Mathematics – Key Technology for the Future
Willi Jäger • Hans-Joachim Krebs (Editors)

Mathematics – Key Technology for the Future

Joint Projects between Universities and Industry

Springer
Preface

Efficient transfer between science and society is crucial for their future development. The rapid progress of information technology and computer systems offers a large potential and new perspectives for solving complex problems. Mathematical modelling and simulation have become important tools not only in scientific investigations but also in analysing, planning and controlling technological and economic processes. Mathematics is playing an important role as a basic science: in setting up models, in analysing their characteristic properties, designing fast and reliable algorithms for the numerical solution of the model equations, estimating the parameters using real data sets, designing and optimising processes, and validating and up-scaling the models with respect to real situations. Mathematics, imbedded in an interdisciplinary concept, has become a key technology. However, the transfer between mathematical research and real applications has to be improved. Connecting mathematics with industry and economics is not a one-way route. Direct, unfiltered access to the results of mathematical research and to the problems and data in real applications are equally important. The toolbox of mathematics contains far more instruments than are being used today. Experts are needed to translate relevant problems in industry and economics into a mathematical setting. “Problem mining” – that means prospecting for areas that can be better and faster explored using mathematical ideas and methods – is an important, but too much neglected task. It is often forgotten that problems can be solved using computers and that mathematical modelling and new mathematical tools could provide better solutions faster and at less expense.

In 1993, the Ministry for Education, Science, Research and Technology of the Federal Republic of Germany (BMBF) started a programme to support joint projects involving mathematics and industry in order to speed up the transfer of knowledge and technology. BMBF entered a new territory since no comparable, concentrated funding of the transfer between mathematical research and industrial and economic applications existed either nationally or internationally. The programme was structured so that existing activities could be enhanced and new initiatives started to break down the prejudice that mathematical research is far from useful in providing practical solutions.

The following conditions were essential criteria for funding within the BMBF programme:

- The relations between mathematical, application-oriented basic research and industry and economics are strengthened.
- New mathematical results are made available for specific applications and new methods are developed using one project as test case.
• The applications selected are relevant for industry or the economy and cannot be solved by standard techniques.
• Partner companies contribute to the projects, at least by providing the special information and data needed.
• Competence in modelling is available; the models and the simulations are validated.
• Teams sharing similar problems co-operate within the network of the programme.
• The projects are planned to achieve results relevant for applications within a period of 3 years.

This initiative proved to be an excellent investment. The impact of the programme within Germany was enormous. It essentially helped mathematics in industry and economics to become a real factor in Germany, both in universities as well as in society. Many faculties of mathematics opened their research and education for industrial and economic applications. Positive experience with this programme is changing the image of mathematics in the non-academic world. This most important effect should not be underestimated.

The results of the projects of the first period were published in the prestigious book “Mathematik – Schlüsseltechnologie für die Zukunft, Verbundprojekte zwischen Universität und Industrie” (K.-H. Hofmann, W. Jäger, T. Lohman, H. Schunck: Springer 1996). The first volume was published in German in order to address the existing and potential partners in Germany. The demand for this publication was so large that it was out of print after a short period of time.

Before starting a third round of projects, a status-meeting was held at the Dresdner Bank, Frankfurt, in December 2000 where a second volume, containing reports on the projects supported by the BMBF up to the end of 2000, was proposed. This publication has now been written in English to make the information programme available to an international readership. The title has been kept since it reflects the message that mathematics provides a key technology.

The spectrum of applications covered by the projects includes the following topics: modelling and simulation of combustion engines; simulation and control of chemical reactors; bio-remediation of polluted soil, modelling and simulation of materials and their technological applications; problems in microelectronics and semiconductor technology; simulation of vehicles, methods for their design; computer tomography, the analysis and evaluation of medical and technological data; methods for quality control of industrial products; optimisation of power plant operation; optimisation of the logistics of factories; optimal planning of traffic and scheduling of vehicles.

New mathematical methods were necessary to solve extremely large differential-algebraic systems, non-linear partial differential equations modelling flow, transport, diffusion and reactions, model equations for crystal growth or
for electromagnetic fields and optics, to solve discrete and continuous optimisation problems, inverse problems, to analyse complex data sets and images and to visualise them. In most projects the algorithms and simulation played a crucial role, however, model design, analysis, calibration and validation were also important topics in the investigations.

A close co-operation with the industrial partners was essential for the projects. The character of the investigations differs in the various projects. Some results can be directly applied in operations; others support the design of new technology. In general, the projects presented in this volume prove that new mathematical ideas and methods can be decisive for the solution of industrial and economic problems. They show that the direct transfer of mathematical tools is advantageous. On the other hand, actual world problems are a challenge for basic mathematical research. We hope that this volume will also encourage students and junior scientists to become more involved with mathematics for industrial and economic applications.

As far as possible, the contributions are arranged according to the areas of applications. This arrangement covers the existing connections and relations in mathematical methods to some extent. It is one of the strengths of mathematical modelling and methods that they depend mainly on the structure of the application. Therefore, similar models and methods can be applied to applications which seem to be rather different at first sight.

Scientists engaged in the programme acknowledge the substantial and effective support provided by BMBF and by all members of the management of this programme. We would particularly like to mention the assistance provided by Dr. Hermann-Friedrich Wagner and Professor Jürgen Richter, BMBF. Their competence and advice were substantial in structuring, planning and pursuing the programme for the transfer of new mathematical results and tools to the solution of industrial and economic problems.

We also appreciate the excellent co-operation with Springer Verlag, Heidelberg, in publishing this volume in its internationally well-known series of scientific publications and helping to spread and promote the messages of the programme of BMBF. Antoniu Bodea and Simina Bodea deserve full credit for the preparation of the electronic files for this volume, a task which demanded far more than just collecting the various contributions.

Heidelberg, January 2003
Willi Jäger                                      Hans-Joachim Krebs
# Table of Contents

## I. Motors, Vehicles

<table>
<thead>
<tr>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthesis of Automotive Cams Using Multiple Shooting–SQP Methods for Constrained Optimization</td>
<td>3</td>
</tr>
<tr>
<td><em>H. G. Bock, R. W. Longman, J. P. Schlöder, M. J. Winckler</em></td>
<td></td>
</tr>
<tr>
<td>Numerical Optimization of Scavenging in Two-Stroke Engines with Transfer Ducts, an Exhaust Port and a Moving Piston</td>
<td>22</td>
</tr>
<tr>
<td><em>D. Kröner, L. Klassen, A. Klimmek, D. Trescher</em></td>
<td></td>
</tr>
<tr>
<td>A Numerical Tool for Flow Simulation in a Wankel Motor</td>
<td>33</td>
</tr>
<tr>
<td><em>R. Rannacher, V. Heuveline</em></td>
<td></td>
</tr>
<tr>
<td>An Efficient Navier–Stokes Solver for Automotive Aerodynamics</td>
<td>43</td>
</tr>
<tr>
<td><em>R. Rannacher, Chr. Becker, S. Turek</em></td>
<td></td>
</tr>
<tr>
<td><em>S. Rjasanow, M. Bebendorf</em></td>
<td></td>
</tr>
<tr>
<td>Combinatorial Optimization Techniques for Three-Dimensional Arrangement Problems</td>
<td>63</td>
</tr>
<tr>
<td><em>T. Lengauer, M. Schäfer</em></td>
<td></td>
</tr>
<tr>
<td>Simulation of Test-drives of Automobiles at Driving Limit</td>
<td>74</td>
</tr>
<tr>
<td><em>H. J. Pesch, M. Gerdts</em></td>
<td></td>
</tr>
<tr>
<td>An Optimal Control Approach To Real-Time Vehicle Guidance</td>
<td>84</td>
</tr>
<tr>
<td><em>R. Bulirsch, M. Vögel, O. von Stryk, C. Chucholowski, Th.-M. Wolter</em></td>
<td></td>
</tr>
<tr>
<td>Theoretical and Experimental Studies of an S-Catamaran</td>
<td>103</td>
</tr>
<tr>
<td><em>K. Kirchgässner, S. D. Sharma, X.-N. Chen, N. Stuntz</em></td>
<td></td>
</tr>
</tbody>
</table>

## II. Environmental Technology

<table>
<thead>
<tr>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust Error Estimators for Interface Problems Occuring in Transport Processes in Porous Media</td>
<td>127</td>
</tr>
<tr>
<td><em>J. Fuhrmann, M. Petzoldt</em></td>
<td></td>
</tr>
<tr>
<td>Modelling and Simulation of a Planned Bio-Chemical in situ Remediation</td>
<td>137</td>
</tr>
<tr>
<td><em>W. Jäger, G. Wittum, W. Schäfer, Ch. Wagner, H. Willershausen</em></td>
<td></td>
</tr>
</tbody>
</table>
Influence of Surfactants on Spreading of Contaminants and Soil Remediation .......................... 152
P. Knabner, S. Bitterlich, R. Iza Teran, A. Prechtel, E. Schneid

Improvement of Environment Observing Remote Sensing Devices by Regularization Techniques .................. 162
P. Maass, Ch. Böckmann, A. Mekler

III. Flow, Transport and Reactions in Technological Processes ........................................ 173

Stability Analysis for Reactors from Chemical Industry .......................... 175
B. Fiedler, M. A. Efendiev, L. Lerman, J. Rademacher, A. Schuppert

Heterogeneous Dynamic Process Flowsheet Simulation of Chemical Plants ................................... 184
F. Grund, K. Ehrhardt, J. Borchart, D. Horn

Numerical Simulation of Annular Chromatography .................. 194
L. Tobiska, A. Thiele

Numerical Methods for Parameter Estimation in Bingham-Fluids ....... 204
G. Wittum, V. Schulz, B. Maar, D. Logashenko

A Viscoelastic Turbulence Model Based on Renormalization Group Theory .......................... 216
M. Niggemann, M. Holzmann, D. Schmidt, K. Soldner

Modelling and Simulation of Capacitor Impulse Welding .................. 233
D. Hömberg, W. Dreyer, F. Duderstadt

Analysis of Transport Processes for Layered Porous Materials Used in Industrial Applications ............. 243
H. Neunzert, A. Zemitis, K. Velten, O. Iliev

Modelling and Numerical Simulation of District Heating Networks with Time-Saving Solution Methods .................. 252
R. D. Grigorieff, R. Köcher

Sensitivity and Robustness Analysis for Construction and Monitoring of Tubine-Generator Shafts .................. 263
D. Prätzel-Wolters, P. Lang, A. Wirsen, S. Kulig

IV. Optics and Sensors ........................................ 277

Adaptive Multigrid Methods for the Vectorial Maxwell Eigenvalue Problem for Optical Waveguide Design .................. 279
P. Deuflhard, F. Schmidt, T. Friese, L. Zschiedrich
Direct and Inverse Problems for Diffractive Structures – Optimization of Binary Gratings ........................................ 293
   J. Elschner, R. Hinder, G. Schmidt

Computation of Electromagnetic Fields for a Humidity Sensor .......... 305
   W. Hackbusch, S. Börm

V. Crystal Growth, Semiconductors .................................... 313

Simulation of Industrial Crystal Growth by the Vertical Bridgman Method .................................................. 315
   G. Dziuk, S. Boschert, A. Schmidt, K. G. Siebert, E. Bänsch,
   K. W. Benz, T. Kaiser

Numerical Simulation and Control of Industrial Crystal Growth Processes ................................................... 331
   K.-H. Hoffmann, A. Voigt, M. Metzger

Optimal Control of Sublimation Growth of SiC Crystals .............. 343
   J. Sprekels, O. Klein, P. Philip, K. Wilmański

Mathematical Modelling and Numerical Simulation of Semiconductor Detectors ......................................... 355
   H. Gajewski, H.-Chr. Kaiser, H. Langmach, R. Nürnberg,
   R. H. Richter

Optimal Design of High Power Electronic Devices by Topology Optimization ............................................... 365
   R. H. W. Hoppe, P. Böhm, G. Mazurkevitch, S. Petrova,
   G. Wachutka, E. Wolfgang

Modelling and Simulation of Strained Quantum Wells in Semiconductor Lasers ........................................... 377

VI. Electronic Circuits .................................................... 391

Efficient Analysis of Oscillatory Circuits ................................ 393
   R. Bulirsch, R. Neubert, A. Schwarz

Modelling and Simulation of Power Devices for High-Voltage Integrated Circuits ....................................... 401
   R. Hünlich, G. Albinus, H. Gajewski, A. Glitzky, W. Röpke,
   J. Knopke

Finding Beneficial DAE Structures in Circuit Simulation ............ 413
   R. März, D. Estévez Schwarz, U. Feldmann, S. Sturtzel,
   C. Tischendorf
CHORAL – a Charge-Oriented Algorithm  
for the Numerical Integration of Electrical Circuits  

P. Rentrop, M. Günther, M. Hoschek, U. Feldmann  

VII. Tomography, Image Analysis and Visualisation  

Reconstructing Crystalline Structures from Few Images  
Under High Resolution Transmission Electron Microscopy  

P. Gritzmann, S. de Vries  

Measurement of Paint Layer Thickness  
with Photothermal Infrared Radiometry  

A. K. Louis, P. Dörr, C. Gruss, H. Petry  

Spatio-Temporal Current Density Reconstruction  
from EEG-/MEG-Data  


Signal Correction in NMR Spectroscopy  

H.-O. Peitgen, T. Boskamp, P. Singer  

On Scattering of Ultrasonic Waves  

P. Mathé, J.H. Zacharias-Langhans  

Smoothing of Tomographic Data  
and Hybrid Volume-Surface Visualisation  

W. Jäger, C. Dărțu  

Video Coding with Adaptive Vector Quantization  
and Rate Distortion Optimization  

D. Saupe, M. Wagner  

VIII. Statistical Methods in Medical Applications  

The Application of Statistical Methods of Meta-Analysis  
for Heterogeneity Modelling in Medicine and Pharmacy, Psychology,  
Quality Control and Assurance  

D. Böhning, U. Malzahn, P. Schlattmann, U.-P. Dammann,  
W. Mehnert, H. Holling, R. Schulze  

An Application for the Analysis of Human Tremor Time-Series  


IX. Optimization in Design and Production  

Free Material Optimization  

J. Zowe, M. Kočvara
Automatic Layout and Labelling of State Diagrams .......................... 584
   P. Mutzel, G. W. Klau

Optimization Problems in a Semi-Automatic Device
for Cutting Leather ...................................................... 609
   A. Pott, H. Glaab

Stochastic Programming for Power Production and Trading
Under Uncertainty ......................................................... 623
   R. Schultz, M. P. Nowak, R. Nürenberg, W. Römisch,
   M. Westphalen

Scheduling Scarce Resources in Chemical Engineering .................. 637
   R. H. Möhring, M. Uetz

X. Optimization in Traffic and Communication ................. 651

Duty Scheduling in Public Transit .................................... 653
   M. Grötschel, R. Borndörfer, A. Löbel

Rotation Planning for the Continental Service of a European Airline . 675
   M. Jünger, M. Elf, V. Kaibel

Computer Aided Scheduling of Switching Engines ..................... 690
   U. T. Zimmermann, M. E. Lübbecke

Train Schedule Optimization in Public Rail Transport .................. 703
   U. T. Zimmermann, T. Lindner

An Integrated Planning Approach for Cellular Radio Networks ....... 717
   R. Mathar, M. Schmeink

Author Index ............................................................. 731
I. Motors, Vehicles

Synthesis of Automotive Cams using Multiple Shooting-SQP Methods for Constrained Optimization


Numerical Optimization of Scavenging in Two-Stroke Engines with Transfer Ducts, an Exhaust Port and a Moving Piston

D. Kröner, L. Klassen, A. Klimmek, D. Trescher

A Numerical Tool for Flow Simulation in a Wankel Motor

R. Rannacher, V. Heuveline

An Efficient Navier–Stokes Solver for Automotive Aerodynamics

R. Rannacher, Chr. Becker, S. Turek

Numerical Simulation of Exhaust Systems in Car Industry – Efficient Calculation of Radiation Heat Transfer

S. Rjasanow, M. Bebendorf

Combinatorial Optimization Techniques for Three-Dimensional Arrangement Problems

Th. Lengauer, M. Schäfer

Simulation of Test-Drives of Automobiles at Driving Limit

H.J. Pesch, M. Gerdts

An Optimal Control Approach To Real-Time Vehicle Guidance

R. Bulirsch, M. Vögel, O. von Stryk, C. Chucholowski, Th-M. Wolter

Theoretical and Experimental Studies of an S-Catamaran

K. Kirchgässner, S. D. Sharma, X.-N. Chen, N. Stuntz
Synthesis of Automotive Cams
Using Multiple Shooting–SQP Methods
for Constrained Optimization

Hans Georg Bock¹, Richard W. Longman², Johannes P. Schlöder¹, and
Michael J. Winckler¹

¹ Interdisciplinary Center for Scientific Computing (IWR), University of
Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
² Dept. of Mechanical Engineering, Columbia University, New York, 10027, USA

Abstract. Cam design is an old field of mechanical engineering. Because of the
complexity of the problem, design procedures have emphasized the use of stan-
dardized approaches and rules-of-thumb, which produce reasonable designs without
attempting to obtain a truly optimized performance. Increased competition among
manufacturers puts pressure on designers to find new ways to deal with the complex-
ity of the problem. In recent years, progress in the mathematics of numerical solution
of optimal control problems has made it possible to obtain numerical solutions
for these problems using realistic models and the needed highly nonlinear state
inequality constraints. The work reported here develops a high level user interface
for cam system designers as well as researchers in the field. It makes available
sophisticated numerical integration that handles the necessary discontinuities, and
numerical optimization with SQP methods to handle the complex optimization cri-
teria and side conditions. This forms a tool for interactively designing cam systems
to create optimal trade-offs between the multiple performance characteristics of
importance.

1 Introduction

Cam follower design is an old field within mechanical engineering, going back
at least as far as internal combustion engines and steam engines. Cams are
used in many types of machines, ranging from sewing machines, to cam con-
trolled subway trains, to the cam controlled valve train in the family sedan,
or in a sports car, or a racing car. Figure 1 shows an example of one such
automotive valve system. As with any complicated design field with a long
history, current design practice in industry makes heavy use of routine design
procedures as well as various rules-of-thumb or heuristic methods. Such rules
may be based on physical modelling of the underlying processes involved, but
they need not be. They can simply be rules that put the design process into
a reasonable region of parameters, one that gives acceptable performance in
most situations. Increased competition among different manufacturers creates
pressure for industry to re-examine these procedures with the aim of finding
ways to improve the designs and produce more competitive products. It is the
purpose of the research effort reported here, to develop numerical methods and software tools that facilitate this process.

The design of cams is a complex process with various performance characteristics of interest, as well as many different complex nonlinear constraints. In automobile cams, one is interested in maximizing the area under the opening curve, in order to facilitate flow through the valve. But one may also be particularly interested in having long life for the cam, or in the case of racing cars, in having the maximum possible operating speed of the cam. On the other hand, it can be of interest to minimize the energy loss of the engine in operating the cam in order to reduce fuel consumption. In some cam applications, the cam nominally operates at a constant speed, whereas in automobile cams, one is interested in good behavior over a wide speed range. A good cam design at one speed is often a poor design at a neighboring speed. Given this situation, what is needed by industry is software that allows one to easily optimize these different possible criteria, or to optimize weighted combinations of these criteria. The designer can then produce an optimized cam design for his single purpose, subject to constraints on other considerations. Or he can adjust the weights among a combination of criteria in order to produce the compromise design most suitable for his product. There are many constraints in the design process, and these constraints are complicated nonlinear functions. Until recently, it was not possible to have software that could handle the complexity of the actual optimization problem. But numerical methods have become available for solving difficult nonlinear dynamical systems with discontinuities, finding optimal control functions for
such systems with complicated cost functionals, and doing so while satisfying complicated nonlinear constraints. In this work, we aim to take an important step, producing software that makes a bridge between these new methods in numerical mathematics and the industrial world.

The evolution of the cam design field took a big step forward in 1955 with the development of the polydyne cam design approach by Stoddart [14]. This approach designs a cam using a polynomial curve, but uses a dynamic model and imposes extra boundary conditions on the polynomial to eliminate residual motion and create a smooth start up. The approach was enhanced in [6], and references [8,9] addressed the issue of residual vibrations in such designs. Reference [16] synthesizes cams as a linear combination of a finite number of sinusoids. The optimization of acceleration performed there is finite dimensional, adjusting the finite number of coefficients. Reference [3] is perhaps the first work to apply optimal control theory to cam design, optimizing in a function space. Reference [15] continues this work, further studying the choice of cost functional, and [7] optimizes energy consumption. Theoretical research has continued in the application of optimal control concepts to cam design, see for example [19]. Reference [1] uses stochastic optimal control, and both [5] and [4] address the issue of optimization over a range of speed. These are developments coming from the academic community, and have not yet had an impact on cam design. Each of these works making use of optimal control concepts, took approaches that were constrained by the lack of effective computational methods for constraining or penalizing all the variables that might be considered important. The research reported here, develops software that effectively addresses this problem. Hence, the software can assist the practicing engineer, and simultaneously be a tool for researchers who examine the issues involved in cam optimization on a more fundamental level. In the process, it can form a bridge between the two communities.

This represents an important opportunity for the cam design field. The field is very challenging because of its complexity, and the new software allows the designers for the first time to address "the real problem". The ability of the numerical mathematics to handle difficult models makes it possible to develop and use realistic models rather than highly simplified models as has been done in most of the literature. The ability of the numerical mathematics to handle difficult optimization problems subject to highly nonlinear state constraints, allows the designer to pose and solve the problems he really wants to solve. The use of rules-of-thumb and heuristics in cam design which represented a way of dealing with the complexity and obtain functional engineering results, can give way to addressing the problems in terms of the actual design objective. The design of a vehicle engine is an engineering task with a lot of different goals. The overall engine layout should be light weighted and small, robust to guarantee a long life and optimized for maximum horsepower while still keeping the fuel consumption low. As the dynamical behaviour of an engine is very complex, numerical optimization should be employed to help to fulfill this difficult project.
2 The Form of the Cam Optimal Control Problem

The formulation of the problem using positions $p$, velocities $v$, accelerations $a$, control function $u$, fixed design parameters $q$ and time $t$ can be written as

$$\max \sum \omega_i F_i(p, v, a, u, t; q) \quad (1)$$

subject to

$$\frac{dp}{dt} = v, \quad \frac{dv}{dt} = a, \quad a = f(p, v, u, t; q), \quad f \text{ piecewise } C^\infty \quad (2)$$

$$g(p, v, u, t; q) \leq 0 \quad (3)$$

$$h(p, v, u, t; q) = 0. \quad (4)$$

The first of these equations defines the objective function (1). Below, there will be considerable discussion of the choice of the objective function. But from the discussion above concerning optimizing opening area, but in addition being concerned about wear, and about energy consumption, one can see the optimization will normally be done as a trade-off cost functional, as a functional consisting of a weighted sum of multiple design objectives. The designer then gives priority to some design objective over others by setting a weight vector. Realistic models for wear, or energy consumption, etc. create cost functionals that are complicated and lead to very nonlinear optimization problems.

The system equations (2) above allow one to numerically simulate the dynamic behavior of any candidate cam profile during the opening and closing process. The dynamics of the valve train allow for several sources of discontinuity, mainly contacts and impacts between the components that can occur for designs during the iterative process of converging to an optimized design. This requires specialized numerical integration software for proper handling.

The inequality constraints (3) are needed to make sure that the dynamical properties of the valve train are within physical bounds imposed by maximum stresses of valve train components. They also handle constraints on curvature of the cam such that the follower can follow the cam curve, and constraints on curvature to meet production restrictions.

For the final design, a set of equality constraints (4) has to be satisfied. These mainly cover initial and endpoints conditions, but also represent geometric continuity conditions for the cam shape.

The solution to this problem is computed using MUSCOD-II, an SQP method for nonlinear programming (see Leineweber [10]). In order to treat the problem with an SQP method, multiple shooting can be used to discretize the state space. Together with this state space discretization, the control function $u(t)$ can be discretized using a piecewise linear and continuous or
even piecewise constant approximation, depending on the choice of the control variable.

Remark: As outlined in section (6), it is often advisable from a technical point of view to use a derivative of the physical control function to be discretized in the numerical solver. In this case additional integral equations to recover the physical control function are added to the set of equality constraints (4).

3 The CAMEO Computer Aided Design Package

The software developed here is called CAMEO (CAM Engineering Optimizer). It contains a high level user interface between several state-of-the-art complex numerical integration and optimization routines. It has the following components:

1a A library of cam models that the user can access as text-based setup files. The file `valve.inp` defines the basic properties of the valve train itself. The current version of the package allows the user to change masses, spring and damper coefficients, radius of the cam base circle and sizes of components. The file `model.inp` selects the modelling components used in dynamical simulation, defining the cam configuration under consideration. A set of 0–1 flags activates the various detailed modelling components. The programming is modular, allowing the user to supply his own model, when he has a specialized cam configuration under investigation that is not included in the library supplied.

1b A state-of-the-art numerical integration package used to integrate the dynamic equations of the cam model, taking care to handle and flag discontinuities as needed both for accurate integration and for the optimization process.

2a The file `opt.inp` is used to set up the optimization criterion and the constraints among a library of options. For all physical variables currently treated, box constraints can be defined. These are treated as optimization constraints during the SQP iterations of the optimization process. By using 0–1 flags and a set of weights the engineer can set up a cost function as a weighted sum of different cost aspects such as the opening area, the lubrication number (Schmierzahl), and the integrated deviation between valve head and valve body movement. The programming is modular to allow for the user to supply his own optimization module when he is concerned with some aspect not currently contained in the library.

2b The optimization is done using the optimal control software MUSCOD-II. When running CAMEO the information inside the above files is used to set up the correct information files for MUSCOD-II. After this initialization process, MUSCOD-II is evoked to solve the optimal control problem. The final result of the MUSCOD-II run is converted into a result file for CAMEO that represents the cam shape and its first and
second derivative, ready to be evaluated in the standard tools used for cam shape design.

3 Finally, graphical representation of the dynamical properties of the resulting cam design is generated for simple visualization. The visualization tool JAFV was enhanced to report loss of contact during an event cycle by color changes of the relevant valve train components. The visualization tool further makes it possible to give a 3D-impression of the actual working area inside the engine during a full rotation cycle (see Fig. 1 and Winckler [18]).

The following sections discuss these aspects in more detail. The range of possible models for cam follower systems is discussed. The types of discontinuities that occur in the dynamic simulation of the model are discussed, and how they are treated in the integration process, and in the optimization process. Then the physical considerations that go into defining the optimal control problem are discussed, including the hard inequality constraints, the choice of control variable, the boundary conditions, and the range of objective functions. During the discussion, examples of the use of CAMEO are given, illustrating its capabilities to deal with complexities of the real world problem.

4 Modelling Valve Trains

In the design process, it is natural to consider a hierarchy of models of increasing complexity and model fidelity. It is best to start with simple models to develop an understanding of the trade-offs or compromises that the designer must make in creating the design. In the next subsection a basic model is described, and in the following subsection various choices of enhancements are treated.

4.1 The Basic Model

Figure 2 shows a basic model that is simple but nevertheless reasonably good. We comment that this model, although simple, is already more complicated and realistic than the models used in much of the literature on the subject. The cam itself is modeled using a vertical displacement function \( r(\phi) \) that is parameterized by the rotation angle \( \phi \) of the camshaft. The cam displacement function is an actuator input for the lumped mass model representing the valve train itself.

As friction forces between valve train components introduce wear into the valve train system, in older valve trains mechanical wear adjustment was necessary. Therefore modern valve train designs introduce hydraulic wear adjustment components to prevent the loss of contact between cam and follower due to wear effects. These hydraulic elements are flexible length links between the upper and lower part of the valve train. They act as a "soft" connector between two rigid bodies and are treated as a spring-damper-element in this
Fig. 2. Valve train: lumped mass model

basic model, linking the valve head mass and the valve body mass. The masses themselves can move along the vertical line of displacement, so the overall model is really a one-dimensional simplification of the planar view. Damping of the head and shaft due to friction is modeled through two additional dampers acting on the masses.

The model represented above assumes that contact is maintained between the cam and the follower, and no provision is made for hard contact while seating, resulting in bounce. The actual simulation must take into consideration these possible changes in the structure of the dynamic equations. Of course in the final design these difficulties should not occur during normal operation, but they can occur for cam designs considered during the optimization process. In addition, these difficulties can define the useful range of operating speeds of the cam. The next section discusses the need for sophistication in handling these matters in the simulation and the optimization codes. So the final basic model not only consists of a set of four ODE (two for the positions, two for the velocities) but also includes a set of at least two switching functions to flag discrete states transitions. These switching functions allow for a mathematically correct and error controlled simulation of the overall dynamics (c.f. Bock [2]).

4.2 Model Refinement

The modelling of a valve train can be quite complex, and is also the subject of considerable research activity. For example, the modelling of wear and of the
relationship between the lubrication and elasticity effect on the follower is not well understood. Some models as in [12] include partial differential equations to model some elements. For practical solution of optimization problems, such models must be discretized, but one should be able to do so without significant compromise in model fidelity. This section discusses a number of model refinements that are treated in the current form of CAMEO, in order to allow the use of more sophisticated models in a routine way. For specialized needs, the user can supply his own model module. The models discussed below can be chosen from a modelling library to replace the standard components of the basic model described above.

**Hydraulic Element**

The hydraulic element can be modeled in detail by analyzing the internal behavior of the mechanics and hydraulics. The principal design (see Fig. 3) is that of a high pressure compartment (A) acting as a “spring” between valve head and valve.

During the opening-closing-cycle of the valve, oil is pressed out of the compartment into a low pressure reservoir (B) through a leakage rim (C). This dynamically reduces the effective length of the element.

After the closing of the valve, the pressure in the refill compartment (D) opens a ball valve (E) to pump up the hydraulic element. In the modelling environment, this refill state is either modeled to be completed at the end of the valve-closed state or treated as a set of additional dynamical equations. In the latter case investigations about the refill behaviour at higher cam speeds can be carried out by the design engineers [11].

![Fig. 3. Hydraulik lash adjustment](image-url)
Valve Spring

At very high cam speeds the valve spring can no longer be modeled as an ideal massless spring. The coils of the spring can surge, having their own internal vibrations. And these can become sufficiently larger that contact between the spring coils in the valve spring occurs. Therefore a refined spring model can be used to simulate the movement of the spring coils during a valve cycle. The coils are treated as a series of lumped masses connected by spring-damper elements. Reduction of the distance between coils below a certain threshold is treated via switching functions to flag a coil contact. This results in a non-smooth change of the force between the coils using a Hertzian contact model.

Valve Head Elasticity

The valve head elasticity depends on the actual position of the contact point of the cam on top of the valve head. Our industrial partner INA provided a non-linear valve head stiffness curve derived from a finite element analysis of the valve train. For additional accuracy this non-linear stiffness curve can be introduce into the contact law for the cam valve-head contact. For the flat-faced valve train the position of the contact point can be analytically determined from the configuration of the valve train and the cam position.

Using these model refinements introduces up to 14 additional differential equations into the model for a maximum total of 18 equations. Selection of the model components can be carried out in CAMEO through a setup file that configures the final dynamical model.

5 Sophisticated Dynamic Equation Solvers Needed for Switching Structures

As mentioned above, the dynamical equations must be able to produce correct results when certain changes in the differential equation structure occur, due to contact or loss of contact. The cam and the valve head interface is modeled as Hertzian contact. Loss of contact will always occur if the cam speed is made high enough. For a poor cam design, one that is not yet optimized through the iterative optimization process, the loss can occur within the normal operating speed range. We implement a switching function to monitor the state of contact (for an introduction to switching functions in dynamical modelling see e.g. v. Schwerin[13]).

Also, proper modelling of the dynamics as the valve approaches its seat is important. A poor design, or one operated at too high a speed will have the valve impact the seat and bounce away, again an undesirable behavior that must be avoided. We use a Hertzian contact model including friction forces to model this contact and set up a switching function to mark the
transition between the two sets of equations (with and without contact at the valve seat). Finally, when a set of lumped masses is used to model the spring, contact between spring coils is monitored in the same manner.

The switching functions described above mark the transition between model components and form a very efficient tool for handling of discontinuities during dynamical simulation. To compute mathematically correct sensitivities for discontinuous dynamical systems as is needed in the optimization, updates of the sensitivity matrices in the presence of discontinuities have to be taken into account. The software ODESIM/ODEOPT by v. Schwerin and Winckler does exactly that, namely generating sensitivities for ODE systems with discontinuities by using switching functions to separate the discreet system component (the switching structure) from continuous parts (the smooth local models) [18].

Optimization of discontinuous dynamical models is a more difficult aspect. Convergence of gradient-based methods for dynamical optimization can usually be guaranteed only in cases where the switching structure (that is: the sequence of switching decisions) is kept fixed over the optimization process. With this restriction SQP methods combined with multiple shooting can be used to solve the optimal control problem. The usual approach for this is to keep the switching points fixed at multiple shooting nodes (e.g. by state-based time transformation) and use standard methods for the generation of sensitivities on the (sufficiently smooth) solution between the nodes.

For the optimization of valve trains this approach is not feasible. The main reason for this is that the switching structure changes during the optimization process (and is indeed supposed to do so). As an example we might start with a cam that looses contact during the event cycle and we are looking for better designs that are smooth enough not to loose contact. Similar effects can be observed for cam designs that reopen the valve after the initial closing due to vibrations inside the valve spring.

To tackle this problem, a first approach is to trace the switching structure during each multiple shooting step. For this purpose the discontinuity monitor ADTMON [17] was redesigned to be able to trace the switching structure of a simulation cycle and compare this to a predefined reference structure. This gives the possibility to allow the switching points to roam freely within a multiple shooting interval. As long as the switch does not cross interval borders the SQP method used for finding the optimal solution works nicely.

The change of the switching structure does not necessarily lead to difficulties in the optimization process. Whereas the change of the switching structure may lead to unsolvable optimization processes in many classed of problems, in the case of the valve train optimization the vanishing of a contact loss during the event cycle leads to a design that is feasible in terms of the constraint conditions formulated during the modelling phase. Further improvement of the optimization criterion can still be possible once loss of contact no longer occurs during the iterations.
In view of this, we implemented ADTMON as a watchdog inside the sensitivity generation module of our SQP code MUSCOD-II. So ADTMON only records changes in the switching structure making it possible to avoid changes that lead to inconsistent optimization problems by a remodelling of the process. As long as convergence of the optimal control problem is achieved within the bounds of the SQP method, the recording of the switching structure tells the user the series of model transitions that were investigated during the optimization process. In case of failure of the optimization, a better set of initial conditions (usually a more sophisticated initial cam shape) has to be provided by the user.

As a side effect, semicontinuous constraints such as a box constraint on the contact force between valve train components can be set under the watchdog function as well. Instead of evaluating such constraints only at the multiple shooting nodes we can trace violations of these constraints at every point of the discretization mesh of the dynamical simulation.

6 Choice of Control Variables and Boundary Conditions

This section discusses issues related to the choice of the control variable, and at the same time addresses some equality conditions for equation (4). These can include conditions imposed on the final solution that produce continuity at the multiple shooting points, and they can also include boundary conditions.

When designing the shape of the cam, the effective variable is the valve position suggested by the point of contact with the follower. This defines a cam excitation curve (see Fig. 4). Since the angle to the point of contact varies as the cam rotates, there is a nonlinear relationship between this curve and the implied actual shape of the cam. At first one might presume that the cam excitation curve should be the choice for the control variable. However, there are additional considerations. One of these is the satisfaction of boundary

Fig. 4. Cam shape and cam excitation curve
conditions, which are contained in the equality constraint represented by equation (4).

Suppose for simplicity that the cam model contained only one mass, and at the end of its motion, we wanted the valve to settle into the seat and remain at rest there. The normal boundary conditions for a second order differential equation allow one to get the mass to the desired position and do so with zero velocity. However, our desire is to have the mass remain at rest in that position. This means that in addition to asking for the usual boundary conditions, we want to impose zero acceleration at the endpoint. To accomplish this, it is necessary to take the derivative of the excitation curve as the control variable, including it in the optimization function. When there are more masses involved, it would require additional derivatives of the excitation curve to get all masses to have zero acceleration at the end. For the work reported here, the third derivative is the chosen control variable. This serves not only to handle the above considerations in a reasonable manner, it also addresses issues of discretization of the control in the numerical optimization process. Multiple shooting is used. With a piecewise constant control as the third derivative, one gets continuity of the second derivative which forms a forcing function of the follower mass. With piecewise linear and continuous parameterization of the control one has continuity of the jerk of the follower mass. Producing such extra smoothness of the cam profile is helpful in limiting the excitation of vibrations.

7 Hard Constraints

Now consider the inequality constraints in equation (3). There are some restrictions on the cam shape itself. Looking at the left part of Fig. 5, it is clear that a flat faced follower would not actually track the reverse curvature portion of the cam. Hence, the actual cam shape is not allowed to have a negative curvature. When considering roller followers as occur in some applications, this constraint is modified to allowing a limited amount of reverse curvature based on the radius of the roller. On the other hand, it is not advisable to make a cam with too strong a curvature as in the right part of Fig.5. Firstly an abrupt change in curvature will produce excessive wear, and secondly, manufacturing such a cam is precluded by the manufacturing machines. Thus there are constraints on both the maximum and minimum curvature of the cam profile.

Note that this is not a constraint on the control variable, which would be handled relatively easily in optimal control theory. Rather it is a complex inequality constraint on the state variables of the system. This is true for two reasons. First the control is the third derivative of the excitation curve, not the curve itself. But in addition, the relationship between the excitation curve and the cam shape is a nonlinear geometric one. Hence, the cam shape must be recovered from the excitation curve in order to formulate these constraints.
on the cam. Fortunately, in the case of flat faced followers, this inversion is possible without resorting to numerical solution.

These curvature restrictions form highly nonlinear inequality constraints on the problem. Not only does the excitation curve itself appear in these constraints, but the first two derivatives are involved in the evaluation of the constraint. This is another reason for using a three times continuously differentiable control parameterization.

There are additional inequality constraints on an acceptable cam, i.e. that contact within spring elements should be avoided, loss of contract between the cam and follower should be avoided (valve float), and bounce of the valve after contact with the seat should be avoided.

8 Optimality Conditions

Section 4 discussed the dynamic model of the cam follower system for equation (1) of the statement of the optimal control problem, equations (1-4). Section 6 discussed equality constraints for equation (4) together with additional equations to be introduced into the state variable system (2) because of the choice of the third derivative for the control variable. Section 7 treated some inequality constraints for equation (3). What remains is to define the optimization criterion for equation (1). Depending how this is done, more inequality constraints (3) can be introduced.

There are many characteristics in the response of a cam follower system that are important, and they must be represented in the problem formulation in some form. For example, a high performance cam will seek to keep the valve as wide open as possible for a long as possible, but one cannot push too far in this direction because it creates large forces, and corresponding wear and short lifetime. Hence, designs are necessarily some form of trade-off between competing desired characteristics. It is the job of the designer to perform this trade-off, for the application in question, and CAMEO is a tool to help him perform this task. The cam for a racing car will pay less attention to issues of wear and lifetime, and more attention to valve opening and high operating speed. There are two options. Equation (1) suggests the use of a weighted sum of all of the variables of concern, such as total integrated valve opening
over the cycle, and some form of wear computations, etc. And then during the design process, the designer adjusts the relative values of the weighting factors. This is typical of modern control theory using the common quadratic cost functional. One does not know in advance what weights one wants, but instead one iteratively adjusts the weights by looking at the performance. The alternative is to optimize some particular set of weighted functions, and apply box constraints to the remaining set of variables, in order to keep them within acceptable limits.

A characteristic of this design problem is its behavior like that of a waterbed— if you push too hard in optimizing one particular property of the cam, in the process of improving this property some other property will go into an unacceptable range. CAMEO then serves as a tool for the designer to perform the trade-offs as needed, in an interactive design process. The following paragraphs discuss possible properties of interest for the cam follower that can form terms in the optimization criterion or be used with box constraints.

8.1 The Area Under the Opening Curve

The main purpose of the valves in an engine is to as freely as possible allow inflow to or outflow from the cylinder at the appropriate times. Hence, one criterion of optimality is the total area under the opening curve over a cycle. One can improve this by weighting the area with some model of the resistance to flow for different opening heights. An ideal valve would open completely, instantaneously, and close again instantaneously.

This defines a rectangular rise-dwell-close behavior as the ideal goal, and one can then formulate the optimization criterion to minimize the difference between the dynamical opening curve reached by the actual cam configuration and this ideal curve. As mentioned above we must not be too ambitious in optimizing for this particular criterion, and therefore impose inequality constraints to keep the solution within an acceptable range with respect to other considerations such as wear.

As an example of this process, consider the simple lumped mass model with a spring-damper element to model the valve spring and the hydraulic element. Take as the initial cam shape a sinusoidal excitation profile. Of course this is a very primitive cam shape that one expects to have poor properties, but we suppose that we do not know anything better to do. Next suppose that the aim is to design a cam to operate at 4500 rpm. The cam speed is one half that of the engine speed, so this corresponds to an engine speed of 9000 rpm, which is far above that of any normal automobile.

The initial sinusoidal profile is unacceptable for a number of reasons. First the cam shape is not convex as required by the inequality constraints (3). This fact is not seen by the simulation model since we use the excitation curve itself for driving the valve train and don’t compute the point of contact between cam and valve head. However, the optimization method
detects the constraint violation, penalizing it during iterations. In addition, the dynamical simulation of the initial design shows a loss of contact for almost 5° cam angle starting at about −40.1° e.g. at the opening flank of the valve cycle. This loss of contact between cam and valve head is typical for a cam shape that is „steep” in the very first part of the valve cycle. Through smoothing of the excitation curve this loss of contact can be avoided. Box constraints on the contact force between cam and valve head, from below and above, are introduced to make sure that the final design stays in contact all the time. This imposes a set of highly nonlinear constraints on the valve train throughout the event phase. Such continuous constraints are treated by discretizing them to a set of evaluation points for which we aim to guarantee their validity. Apart from that we will ensure the constraint throughout the cycle by placing an observation function on the design variables for the whole event phase (see Section 5).

Starting with a discretization of the sinusoidal design as the initial value of the control function and a single shooting forward integration to generate initial values for the state variables, we need 64 SQP iterations with MUSCOD-II to reach a final design that does not violate the constraints and is optimal with respect to the opening area criterion (taking a few minutes on a standard Pentium-III 500MHz PC). A comparison between the cost function values for the initial and the final design gives:

<table>
<thead>
<tr>
<th>Sinusoidal cam</th>
<th>Final cam</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 1.5231018</td>
<td>area = 1.1210788</td>
</tr>
</tbody>
</table>

This example shows the ability of the software to generate designs near the limit of physical performance. This loss of 27% for the opening area is the price one pays to fulfill needed constraints. The above computation can serve as the first stage in the overall design of a very high speed cam. The next stage in the design process is to include some of the considerations below using CAMEO as an interactive design tool.

### 8.2 Optimization for Wear Reduction and Force Minimization

The modelling of wear in cam follower systems is a very complex field, and is still a subject of research effort. Some of the variables that influence wear and contact stress are the pressure angle, the cam offset, the follower mass, the preload, the follower guide dimensions, and follower guide friction, material properties, and cam curvature. One can consider various representations for wear related variables that one can include in the optimization criterion. Reducing the contact force is helpful. One can ask for low Hertzian contact stress between cam and follower [3]. Another issue of concern to cam designers is the pressure angle. As the cam rotates, the point of contact on the follower
moves on the follower face. When it moves far from the axis of the valve stem, this can cause increased wear in the valve guides, as well as binding during the valve motion.

In order to make higher fidelity representations, one needs to study the lubrication effects as they relate to elastodynamic phenomena at the cam-follower interface. This field requires complicated modelling and is a field of current research in the literature.

### 8.3 Lubrication Number

The lubrication number \((Schmierzahl) L\), is a characteristic number to evaluate the amount of wear at the return point when the valve reaches the maximum opening position. It is calculated as

\[
L = r_0 + s + 2 \cdot \frac{a}{\omega^2}
\]  

where \(r_0\) is the radius of the cam circle, \(s\) is the valve opening, \(a\) is the valve acceleration and \(\omega\) is the rotational speed of the cam. Valve designers aim to have as high a Schmierzahl as possible at the return point of the valve thus ensuring a smooth turn-around of the valve. In CAMEO the Schmierzahl is treated as either a direct optimization criterion or else put into the set of constraints with a lower bound value given by the design engineer. The dynamic simulation allows for the Schmierzahl to be evaluated exactly at the return point, while purely kinematically based design tools take the kinematic acceleration at the cam curve maximum as a substitution for the real Schmierzahl. Computation of the Schmierzahl is carried out using an event function on the valve velocity to mark the return point and flag the evaluation.

### 8.4 Energy Optimization

Another consideration that is of importance in economy cars is the energy consumed in running the cam. This is not a trivial amount of energy, taking some horsepower away from the engine. Reference [7] does a preliminary study of the use of this criterion with very simple cam models, and the software developed here allows one to investigate this issue with realistic models and constraints.

### 8.5 Vibrations and Maximum Safe Operating Speed

The excitation of vibrations in the cam follower system is an important difficulty in achieving high speeds of operation. These vibrations are the source of valve float, contact within the coils of the valve spring, and bounce during seating. One approach is to penalize the relative displacement, i.e.
the deviation between the instantaneous actual displacement and the static displacement associated with the follower position. How to handle the preload in the optimization is an issue to be addressed. The smallest possible preload will minimize the forces in the system, but a high preload is necessary to maintain contact at high speeds.

8.6 Optimization of the Ramps

Current design practice uses ramps for a smoother approach to the seat. Normally optimization subject to this design practice requires specifying the boundary conditions at the boundary of the ramp phase. However, with the software capability of CAMEO, one can let the optimization decide what ramp effects would be optimal.

8.7 Optimization Over Speed Ranges

Many cam design problems in industry are for cams that operate at a nominal design speed. The automotive application requires good operation over a wide range of speed. One can ask for perfect matching of the boundary conditions for valve seating at a specific speed. But over a range of speeds one needs to modify the boundary condition specification and include penalties of some kind for deviations from meeting the desired end conditions over the range of speeds. In [5] the cam is designed for a chosen set of operating speeds, and of course it will not be possible to ask for zero acceleration at all these speeds, so a penalty function is included in the cost. In order to influence the performance at speeds between those chosen, sensitivity functions are included into the optimization criterion. In place of this process, reference [4] formulates an integral expression for a desired penalty of final acceleration over the desired range of operating speed. These approaches can now be investigated using more realistic models and constraint specification.

9 Conclusions

In recent years, developments in numerical methods for the solution of optimal control problems has made it possible to produce solutions to complex problems using realistic models and realistic state equality and inequality constraints. This makes it possible for cam system designers to optimize “the real problem” rather than use rules-of-thumb, or heuristic approaches. The field of cam design is complex, having many competing desired properties for good performance. Hence, the design process is one of iteratively making trade-offs between competing characteristics. The software developed here gives a high level interface between the user and state-of-the-art optimization software, in order to facilitate this design process. Modules are supplied to easily prescribe the cam model and to describe the chosen optimization
criterion. If the user needs to specialize these for his purpose, he can supply his own modules. The result makes available to both practitioners in the field and researchers in the field an important new capability.

References


Numerical Optimization of Scavenging in Two-Stroke Engines with Transfer Ducts, an Exhaust Port and a Moving Piston

D. Kröner¹, L. Klassen¹, A. Klimmek², and D. Trescher¹

¹ Institute for Applied Mathematics, University of Freiburg, Hermann-Herder-Straße 10, 79104 Freiburg, Germany, E-mail: dietmar@mathematik.uni-freiburg.de
² Fa. Andreas Stihl AG & Co, 71336 Waiblingen, Germany

Abstract. For the benefit of the environment, the HC-emission of two-stroke engines has to be reduced. This can be done by reducing the losses of scavenging by improving the geometry of the transfer ducts and the exhaust port. Numerical simulations of the flow through the two-stroke engine should be performed for different geometries in order to reveal the geometry with an optimal scavenge process. The simulations can help to accelerate the development of new two-stroke engines. The underlying mathematical model consists of the compressible Navier–Stokes equations in the cylinder with a moving piston. For the discretization we use a stabilized finite volume scheme on a hexahedral mesh. Up to now we have developed a numerical code for computing the flow in the cylinder and the most important integral quantities such as trapping efficiency and the percentage of exhaust gas at the exhaust port. Now we are able to analyze quantitatively the scavenge process and to estimate the quality of different drafts for the geometrical design.

1 Description of the Problem

The main purpose of this project is the improvement of the geometry of two-stroke engines in order to reduce the pollutant emission.

Currently, two-stroke engines are used in great quantities for motorcycles, marines and ship drives, for motorcar and aircraft models, and especially in hand-operated power tools such as chainsaws, brushcutters, etc. Advantages of two-stroke engines over the four-stroke ones consist in favourable ratios of their efficiency and petrol consumption to scavenge volumes. However, the disadvantage is worse exhaust data. To attain a drastic lowering of the exhaust gas, new concepts for improvement of the design have to be developed and employed. One approach is to optimize the geometry of two-stroke engines.

The fresh gas enters the cylinder of the engine (Figure 1a) when the moving piston opens the transfer ducts. The gas is pushed inside the cylinder by the excess pressure in these ducts, achieved as a result of compression of the fresh charge below the piston by its downward motion. The process of inducing the fresh gas into the cylinder is called the scavenge process. With
a bad geometry, the fresh gas streams directly into the exhaust port (see Figure 2, right stream line). It is, therefore, totally lost for the next ignition, it diminishes the power and burdens the environment. In case of a good geometry, the deflagrated gas is ejected into the exhaust port by the fresh charge rising from the transfer duct, and the cylinder is filled with a mixture of the fresh gases. In this case, the fresh charge can also reach the exhaust port, and some part of the deflagrated gas can remain in the cylinder. After the transfer duct and the exhaust port have both been closed by the rising piston (Figure 1b), the compression process in the cylinder begins. As the piston moves downwards, the exhaust port opens, and the deflagrated gas streams in the form of a pressure shock-type wave from the cylinder into the exhaust port [1, p. 52]. The reflection of the pressure wave at the tapered shape of the exhaust pipe can be used to return the fresh gas, which has already arrived at the exhaust port, back into the cylinder (the download effect). The most efficient possibility for reducing the pollutant emission is to minimize the loss of the scavenging. By means of the numerical simulation for the gas motion in the cylinder, one can determine which changes of the geometry of the cylinder, transfer ducts and exhaust port can improve the composition of the gas mixture and minimize the losses of scavenging. In order to analyze the quality of new geometries, the fresh and deflagrated gas fluxes have to be computed in the transfer ducts and in the exhaust pipe and have to be compared for these geometries.

2 Mathematical Model

The mixture of the fresh and deflagrated gases is considered as a compressible fluid characterized by the total mass density $\rho$ and the velocity vector $\mathbf{v}$. If $\sigma, \mathbf{v}_\sigma$ and $\tau, \mathbf{v}_\tau$ denote respectively the densities and the velocities for the
fresh and deflagrated gases, then we have by definition: $\rho = \sigma + \tau$, \( \mathbf{v} = (\sigma \mathbf{v}_\sigma + \tau \mathbf{v}_\tau)/\rho \). Hence, the mixture consists of two components filling one and the same volume. The concentration of the fresh gas in the mixture can be characterized by the ratio $c := \sigma/\rho$. We assume that, in each fixed point of the space, the change of the concentration $c$ occurs only due to convective effects. This means that diffusion of the components in the mixture is neglected. The latter implies $\mathbf{v}_\tau = \mathbf{v}_\sigma = \mathbf{v}$. The combustion effects are also ignored. A further assumption is that, for each gas, the thermodynamic parameters and the equations of state can be introduced independently of the presence of the other component. In accordance with this assumption, the energy of the interaction between the gases is neglected, and the density of the total internal energy is equal to the sum of the internal energy densities for the components. It is assumed that both components of the mixture are perfect gases. In each particle of the mixture, the temperatures of the components are assumed to be equal: $T_\sigma = T_\tau =: T$, while the pressure of the mixture is the sum of the partial pressures: $p = p_\sigma + p_\tau$ (Daltons law, see [21]). Under the above assumptions, the mathematical model for a motion of the mixture is given by the compressible Navier-Stokes equations for the mixture, complemented with the mass conservation equation for the fresh gas and with the perfect-gas equation of state for the mixture. This model can be written as follows.
Numerical Optimization of Scavenging in Two-Stroke Engines

(see [17], [21]):

\[ \partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) - \nabla \cdot \mathbf{h}(\mathbf{u}) = 0 \quad \text{in} \quad (\Omega_{\text{cyl}} \cup \Omega_{\text{duct}} \cup \Omega_{\text{ex}}) \times \mathbb{R}_0^+, \quad (1) \]

with

\[ \mathbf{u} := \begin{pmatrix} \rho \\ \sigma \\ \rho v_1 \\ \rho v_2 \\ \rho v_3 \\ e \end{pmatrix}, \quad \mathbf{f} := \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}, \quad \mathbf{h} := \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \]

and

\[ f_k(\mathbf{u}) := \begin{pmatrix} \rho v_k \\ \sigma v_k \\ \rho v_1 v_k + \delta_{1k} p \\ \rho v_2 v_k + \delta_{2k} p \\ \rho v_3 v_k + \delta_{3k} p \\ (e + p) v_k \end{pmatrix}, \quad h_k(\mathbf{u}) := \begin{pmatrix} 0 \\ 0 \\ \tau_{k1} \\ \tau_{k2} \\ \tau_{k3} \end{pmatrix} \begin{pmatrix} \kappa \partial_k T + v_1 \tau_{k1} + v_2 \tau_{k2} + v_3 \tau_{k3} \end{pmatrix} \]

for \( k \in \{1, 2, 3\} \),

\[ T = \frac{p}{\rho R'}, \quad (2) \]

\[ p = \left( \frac{\gamma_r c_{\text{vr}} (\rho - \sigma) + \gamma_\sigma c_{\text{ur}} \sigma}{c_{\text{vr}} (\rho - \sigma) + c_{\text{ur}} \sigma} - 1 \right) \left( e - \frac{\rho}{2} (v_1^2 + v_2^2 + v_3^2) \right). \quad (3) \]

Here \( \Omega_{\text{cyl}}, \Omega_{\text{duct}} \) and \( \Omega_{\text{ex}} \) denote domains of the cylinder, transfer ducts and exhaust port respectively,

\[ \tau_{ij} := \mu (\partial_i v_j + \partial_j v_i) - \frac{2}{3} \mu (\nabla \cdot \mathbf{v}) \delta_{ij}, \quad i, j \in \{1, 2, 3\} \]

are the components of the viscous part of the stress tensor for the mixture, \( \mu \) is the viscosity coefficient of the mixture, \( v_i \) are the components of the mixture velocity \( \mathbf{v} \), \( \delta_{ij} \) is the Kronecker symbol, \( e \) is the density of the total energy for the mixture (the sum of the internal and kinetic energies), \( \kappa \) is the heat transfer coefficient of the mixture, \( \gamma_\sigma := c_{p\sigma} / c_{u\sigma} \), and \( \gamma_r := c_{p\tau} / c_{u\tau} \) with \( c_{p\sigma}, c_{p\tau}, c_{u\sigma}, c_{u\tau} \) denoting the specific thermal capacities of the gases under constant pressure and constant volume respectively. In the equation of state (2), \( R := (\sigma R_\sigma + (\rho - \sigma) R_\tau) / \rho \) where \( R_\sigma \) and \( R_\tau \) are the gas constants for the fresh and deflagrated gas correspondingly. The relationship (3) is a consequence of Daltons law and of the perfect-gas equation of state assumed for both gases.

Equation (1) contains six unknown functions, they are components of the vector \( \mathbf{u} \). In the expressions for \( f_k, h_k \), the functions \( T \) and \( p \) are determined by the components of \( \mathbf{u} \) by means of equations (2) and (3).

It can be shown that under the assumption

\[ |c_{u\tau} / c_{u\sigma} - 1| << 1, \quad |\gamma_\tau / \gamma_\sigma - 1| << 1, \quad (4) \]
equation (3) is reduced to

\[ p = (\gamma_r - 1) \left( e - \frac{\rho}{2}(v_1^2 + v_2^2 + v_3^2) \right). \]  

(5)

Assuming inequalities (4) are valid, we have used equation (5) instead of (3) for numerical simulations. The favourable property of using equation (5) is that the Navier–Stokes equations and the equation for the mass conservation of the fresh gas are not coupled any more.

For \( t = 0 \) (the piston is at top dead center, i.e. at its highest position) we assume

\[ \rho |_{\Omega_{\text{duct}}} = 1.02 \text{ kg/m}^3, \quad \rho |_{\Omega_{\text{cyl}}} = 5.55 \text{ kg/m}^3, \]
\[ \sigma |_{\Omega_{\text{duct}}} = 1.02 \text{ kg/m}^3, \quad \sigma |_{\Omega_{\text{cyl}}} = 0 \text{ kg/m}^3, \]
\[ (\rho v) |_{\Omega_{\text{duct}}} = 0 \text{ kg/m}^2 \text{ sec}, \quad (\rho v) |_{\Omega_{\text{cyl}}} = 0 \text{ kg/m}^2 \text{ sec}, \]
\[ p |_{\Omega_{\text{duct}}} = 9.199 \times 10^4 \text{ Pa}, \quad p |_{\Omega_{\text{cyl}}} = 4.9 \times 10^6 \text{ Pa}, \]
\[ \rho |_{\Omega_{\text{ex}}} = 0.41 \text{ kg/m}^3, \]
\[ \sigma |_{\Omega_{\text{ex}}} = 0 \text{ kg/m}^3, \]
\[ (\rho v) |_{\Omega_{\text{ex}}} = 0 \text{ kg/m}^2 \text{ sec}, \]
\[ p |_{\Omega_{\text{ex}}} = 1.088 \times 10^5 \text{ Pa}. \]

The values for \( \rho \) and \( p \) are the outcome of calculations performed in a one-dimensional model of the scavenge process, that incorporates also the gas motion effect in the crankcase located below the piston. The temperature \( T \) and the total energy density \( e \) are calculated with the help of the initial values for \( \rho, p \) and \( v \) given above.

On the fixed solid boundaries \( \Sigma_{\text{duct}}, \Sigma_{\text{cyl}} \) and \( \Sigma_{\text{ex}} \), the no-slip and adiabatic-wall assumptions hold

\[ v = 0, \quad \partial T / \partial n = 0, \]

where \( \partial / \partial n \) is the partial derivative in the direction normal to the boundary. On the piston surface \( \Sigma_{\text{pist}} \) moving with the time-dependent velocity \( v_{\text{pist}}(t) \), we have

\[ v = v_{\text{pist}}(t), \quad \partial T / \partial n = 0. \]

In accordance with equations (2) and (5), the internal energy per unit mass of the mixture (that is \( e / \rho - 0.5(v_1^2 + v_2^2 + v_3^2) \)) depends linearly on the temperature \( T \). Therefore the adiabatic boundary conditions imply \( \partial (e / \rho) / \partial n = 0 \). On the inflow boundary \( \Sigma_{\text{in}} \) we use the time-dependent values for \( \rho, p \) and \( v \) determined by the one-dimensional model of the scavenge process, and we put \( \sigma = \rho \) (\( T \) and \( e \) follow from the relationships (2) and (5)). On the outflow boundary \( \Sigma_{\text{out}} \) we use again \( p \) from the one-dimensional scavenge-process simulations (\( \rho, \sigma, v \) are calculated by extrapolation from the inner domain).

Note finally that the Reynolds number calculated by use of the maximum inflow velocity (58.3 m/s) and the maximal diameter of the transfer duct (0.03 m) as characteristic velocity and length scales respectively, has found to be approximately equal to \( 1.2 \times 10^5 \).
3 Numerical Methods

For the discretization of the compressible Navier–Stokes equations we use an explicit time dependent upwind finite volume scheme on an irregular hexahedral 3D mesh. The stabilization of the numerical scheme is necessary since the Reynolds number is of the order $10^5$. The grid is generated on the basis of CAD data of the geometry of the two-stroke engine with the help of [11]. The CAD data are produced by the Fa. Stihl. For the convective terms we use the AUSMDV Riemann solver, which has been proved to be the most effective one compared to several others [6]. The discretization of the diffusive terms is performed by use of the gradients in the direction of the normals to the cell-faces. The complete algorithm has been validated in [6] for examples with known explicit solutions in 2D and 3D.

The moving piston cuts the cells of the grid and the remaining fractions of the cells are considered as ordinary cells of the grid. If these cells become too small and imply a bad CFL condition, then these small parts are added to larger neighbouring cells [23].

This first-order scheme has been implemented on a shared memory parallel computer with 46 processors (SGI ORIGIN 2000), and the second-order scheme is being implemented now. We expect the computations to run relatively fast, so that it will be possible to simulate the flow through different engine geometries in reasonable time and to optimize the geometry iteratively.

Up to now we have performed computations with 400000 hexahedrons and 200000 time steps.

Adaptive local grid refinement [14], [18]–[20] with a dynamical load balancing [5], [22] and higher order discretizations [12], [16] on nonconformal unstructured hexahedral mesh have been implemented and tested for the unsteady Euler equations in 3D [2], [3], [8], [23]–[25] and will be extended to this Navier–Stokes solver.

For the visualization we have used GRAPE [9], [10], [10], [26].

4 Numerical Results

In Figure 2, the cylinder geometry and the particle paths are presented for three particles of the fresh gas entering the cylinder. One particle is transported directly to the exhaust port. Such a kind of motion, called short circuit streaming, has to be avoided. The other two particles flow into the cylinder. Our numerical simulations show that the flow is governed by the pressure difference between the inflow of the ducts and the outflow of the exhaust pipe.

Up to now comparative numerical simulations have been performed for the two different geometries of the engine: “open geometry” and “handle geometry” (see Figure 3). For the qualitative and quantitative study of these geometries, we have calculated the percentage of the fresh gas inside the
Fig. 3. Two-stroke engine: open geometry (left) and handle geometry (right)

cylinder as well as the mass fluxes at the inflow of the transfer ducts and at the outflow of the exhaust pipe. At the outflow of the exhaust pipe, we have calculated also the mass flux for the deflagrated gas. This allowed us to determine the dimensionless parameter

$$\text{percentage of exhaust gas} := \left(1 - \frac{\text{fresh gas mass flux at outlet}}{\text{total gas flux at outlet}}\right) \times 100\%$$

which is a very important measure for the fraction of the fresh gas reaching the exhaust during one scavenge period. If two geometries have equal fluxes of the fresh gas at the inlet, then the better geometry is that one which has the larger percentage of exhaust gas.

The results for the “open geometry” are shown in Figure 4 and those for the “handle geometry” in Figure 5. These numerical calculations have been performed on grids with 250000 elements. Presented are respectively the mass fluxes at the cylinder inflow and outflow, the fresh gas portion in the cylinder, and the percentage of exhaust gas as functions of the crank angle, CA (that defines the piston position). It is seen that for the “open geometry”, the total flux of the fresh gas (that is its inflow for the scavenge period) is larger than that for the “handle geometry” (compare Figure 4a and 5a, solid lines). The fresh gas portions in the cylinder are, however, nearly the same. That is, because, for the “handle geometry”, less fresh gas reaches the exhaust port. The latter is also a cause for a larger percentage of exhaust gas for this geometry.

Typical dimensionless parameters [1, p. 44] for comparison with the physical experiment are

$$\text{scavenge ratio} := \frac{m}{m_{\text{ref}}}$$
Numerical Optimization of Scavenging in Two-Stroke Engines

Freshcharge at Transfer duct

- massflow in kg/s
- total massflow in g

deg CA

0.04
0.03
0.02
0.01
0.00

100 120 140 160 180 200 220 240 260

Fig. 4. Numerical results for the open geometry

Cylinder/Exhaust

- massflow freshcharge in kg/s
- total massflow freshcharge in g
- massflow fresh + exhaust in kg/s
- total massflow fresh + exhaust in g

deg CA

0.04
0.03
0.02
0.01
0.00

100 120 140 160 180 200 220 240 260

in % Percentage freshcharge in Cylinder

60
50
40
30
20
10
0

100 120 140 160 180 200 220 240 260

deg CA

Fig. 4. Numerical results for the open geometry

in % Percentage exhaust gas at Exhaust port

100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0

100 120 140 160 180 200 220 240 260

deg CA

trapping efficiency := \frac{m_{trap}}{m},

where \( m \) is the mass of the fresh gas scavenging through the cylinder inflow, \( m_{ref} \) is the reference mass (the mass of the air filling the cylinder under the normal atmospheric pressure with the largest cylinder volume), and \( m_{trap} \) is the fresh gas mass that is in the cylinder at the end of the scavenging period. The following relationship holds: \( m=m_{trap}+m_1 \) where \( m_1 \) is the mass of the fresh gas, that flows in the exhaust pipe by short circuit streaming.
In Figure 6, the percentage of exhaust gas and the trapping efficiency for the “handle geometry” are presented. They are compared here for calculations performed respectively at grids with 50000, 100000, 250000, and 400000 elements. The results for the percentage of exhaust gas show yet some dependence of the solution from the grid, so that further refinement of the resolution is still needed. The trapping efficiency is presented as function of the scavenge ratio. This dependence shall be needed for comparison with physical experiments (compare [1, p. 44]). Some small dependence on the grid is also seen here. Yet an increase of the number of cells would result in long computing times. An effective possibility to improve the resolution on grids
with a moderate number of elements is a higher-order discretization. In our test simulations [12] we used the reconstruction of Durlovsky, Engquist, Osher [4], [7], [13, p. 213] and showed that in order to reach the same accuracy, the calculation can be done on a much coarser grid in considerably less time.

Fig. 6. The percentage of exhaust gas and the trapping efficiency for the handle geometry

References


A Numerical Tool for Flow Simulation in a Wankel Motor

Rolf Rannacher and Vincent Heuveline

Institut für Angewandte Mathematik, Universität Heidelberg, INF 294, 69120 Heidelberg, Germany
E-mail: vincent.heuveline@iwr.uni-heidelberg.de
E-mail: rolf.rannacher@iwr.uni-heidelberg.de

Abstract. We describe the main steps in the development of a new numerical tool for simulating gas flow and heat transfer in a rotary engine ("Wankel motor"). Our approach comprises a 2D/3D grid generator for the Wankel motor geometry, an implicit finite element discretization for coping with the stiff pressure-velocity coupling and robust multigrid solvers on strongly distorted meshes. These components are implemented within a new FE-software package named Hi-Flow++ which is presently under development.

1 Description of the Project

The aim of this project is to construct a numerical tool for the simulation of gas flow and heat transfer in a rotary engine ("Wankel motor", Fig. 1). Particular numerical difficulties are caused by the extreme geometry changes and the inherent stiffness due to "low-Mach-number" flow conditions. The latter requires the use of an implicit solution approach which is orientated by the incompressible limit case. Many commercial codes are inefficient under such conditions since they are mostly based on explicit methods which may be appropriate at higher Mach number but fail for $Ma \to 0$ and on highly anisotropic meshes.

The first results of this project are the implementation of a 2D/3D grid generator for the Wankel motor geometry and the various components for a Navier-Stokes solver adapted to this special configuration. This particularly comprises a finite element discretization of higher order, an implicit pressure-velocity coupling as well as the treatment of strongly varying anisotropic meshes. To this end, a new finite element software package, Hi-Flow++, has been developed which is particularly designed for flow problems on locally refined and possibly distorted meshes in 2D as well as 3D geometries. At the moment, the 2D version of the code has been completed which solves besides various model equations in particular the stationary compressible Navier-Stokes equations in the low-Mach-number approximation. The current steps in this project are the extension to nonstationary flows and the implementation of the basic ingredients for the 3D solver. For the latter the main components like mesh handling, assembling of system matrices, mesh transfer operations and mesh-point renumbering have already been realized.
1.1 The Wankel Motor

Unlike the conventional combustion engines with reciprocating piston motion, the Wankel Rotary Combustion Engine (RCE) is based on the continuous rotation of the so-called rotor [BEN,AB]. It has four phases in its combustion cycle: intake, compression, power and exhaust.

Fig. 1. Four combustion phases: intake, compression, power and exhaust

The fuel air mixture is swept along, so the four phases take place in different areas of the engine. Intake and exhaust timing is accomplished directly by the motion of the rotor which avoids the moving parts of the classical combustion engines such as valves, cams and timing belts. The most specific feature in the design of the Wankel motor is probably the peripheral envelope, an epitrochoid curve which is generated by rolling a circle around another one. This geometry allows to constantly keep the three rotor tips in contact with the envelope during the eccentric rotation of the rotor and therefore to separate the three chambers. The rotor is designed to minimize the surface to volume ratio in the combustion chamber in order to increase the thermodynamical efficiency. This is best achieved by assuming the biggest possible rotor contour in the peripheral envelope. Unfortunatly such a contour would separate the combustion chamber in two parts during the power phase. A so-called pocket is therefore moulded in the rotor as shown on Fig. 2. In our numerical simulations we assumed the pocket profile to be parabolic. The specifications of the Wankel RCE in our numerical simulations as proposed by the industrial partner (Wankel Rotary GmbH) are given in Table 1.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Generating radius</td>
<td>8.65 cm</td>
</tr>
<tr>
<td>Eccentricity</td>
<td>1.35 cm</td>
</tr>
<tr>
<td>Chamber width</td>
<td>6.5 cm</td>
</tr>
<tr>
<td>Compression ratio</td>
<td>1 : 7.4</td>
</tr>
<tr>
<td>Basic engine speed</td>
<td>3600 rpm</td>
</tr>
</tbody>
</table>

Table 1. Specifications of the Wankel RCE

Fig. 2. Pocket locations
The maximum rotation velocity of this prototype is approximately $v_{\text{max}} \approx 30 \text{ m/s}$ which corresponds to a flow behavior in the low-Mach-number range as $Ma_{\text{max}} < 0.1$ in this case. Accordingly acoustic waves may be neglected which is achieved by the “low-Mach-number” approximation of the compressible Navier-Stokes equations.

1.2 The Mathematical Model

The flow in the rotary engine is described by the compressible Navier-Stokes equations in the “low-Mach-number” approximation. Here, the pressure $p$ is split into a thermodynamic part $P_{\text{th}}$ which is constant in space and a hydrodynamic part $p_{\text{hyd}}$, while only the latter one is used in the equation of state (see e.g. [MAJ]). Due to the temporal variation of the domain, the thermodynamic pressure is time dependent, too. We denote by $D_t := \partial_t + v \cdot \nabla$ the material derivative. The whole system of equations written in primitive variables takes the following form:

$$T^{-1} D_t T - \nabla \cdot v = P_{\text{th}}^{-1} \partial_t P_{\text{th}}$$  \hspace{1cm} (1)

$$\rho D_t v - \nabla \cdot (\rho \mu \sigma) + \nabla p_{\text{hyd}} = \rho f_e$$  \hspace{1cm} (2)

$$\rho c_p D_t T - \nabla \cdot (\lambda \nabla T) - \partial_t p_{\text{hyd}} - \mu \sigma : \nabla v = \partial_t P_{\text{th}}$$  \hspace{1cm} (3)

where $\sigma := (\nabla v + \nabla v^T) - \frac{2}{3} (\nabla \cdot v) 1$, $c_p$, $\lambda$, $f_e$, and $\mu$ represent the shear-stress tensor, the heat capacity, the heat conductivity, the volume forces and the dynamic shear viscosity, respectively. The equation of state is given by

$$\rho = P_{\text{th}} / RT,$$  \hspace{1cm} (4)

The time derivative of $P_{\text{th}}$ is obtained by first averaging the continuity equation (1) in space and then substituting $D_t T$ by using (4), whereas the heating due to $p_{\text{hyd}}$ and $\mu \sigma : \nabla v$ is neglected. This leads to the following scalar ODE for $P_{\text{th}}$:

$$\partial_t P_{\text{th}} = \frac{c_p - R}{|\Omega|} \left( -P_{\text{th}} c_p \int_{\partial \Omega} v \cdot n \, dS + R \int_{\Omega} \nabla \cdot (\lambda \nabla T) \, dx \right),$$  \hspace{1cm} (5)

where $P_{\text{th}}(0) = P_0$ is a given initial value. Our numerical approach to system (1)–(3) is based on its “variational” formulation which will be briefly described below. Let, $(\cdot, \cdot)$ denote the usual $L^2$-scalar product on $\Omega$. Further, $H^1(\Omega)$ is the space of $L^2$-functions with generalized (in the sense of distributions) first-order derivatives in $L^2(\Omega)$. This is the only notation from mathematical theory of function spaces we are going to use in this paper.

The variational formulation of the stationary form of the system (1)–(3) is obtained by multiplying the equations by appropriate test functions $\{ \chi, \psi, \pi \} =: \phi$ and integrating over the domain $\Omega$. This leads us to define
the stationary semi-linear form \( a(\cdot; \cdot) \) by

\[
a(u; \phi) := (T^{-1}v \cdot \nabla T, \chi) - (\nabla \cdot v, \chi) + (\rho(v \cdot \nabla) \psi, \psi) + (\mu \nabla v, \nabla \psi) - (p_{\text{hyd}} - \frac{1}{2} \mu \nabla \cdot v, \nabla \psi) - (\rho f_e, \psi) + (\rho c_p v \cdot \nabla T, \nabla \psi) + (\lambda \nabla T, \nabla \pi) - (\mu \sigma : \nabla v, \pi).
\]

In the diffusive terms, we have used integration by parts. Neumann-type boundary conditions are implicitly represented by the variational formulation, while Dirichlet boundary conditions have to be explicitly imposed on the solution. The variational form of system (1)-(3) then reads in short terms: Find \( u(t) := \{T(t), v(t), p_{\text{hyd}}(t)\} \in V + u_b \), such that \( u(0) = u_0 \) and

\[
(Q \partial_t u, \phi) + a(u; \phi) = F(\phi) \quad \forall \phi \in V.
\]

where \( Q \) is a suitable coefficient matrix multiplying the time derivatives. The right-hand side \( F(\cdot) \) contains the slave variable \( p_{th} \) given by the relation (5), while \( \rho \) is determined through the modified gas law (4). The term \( u_b \) represents prescribed boundary data. The function space \( V \) in which \( u - u_b \) is sought is the tensor product of certain subspaces of \( H^1(\Omega) \) for temperature and velocity while the space for the hydrodynamic pressure is \( L^2(\Omega) \). If Dirichlet conditions for the velocity are imposed along the entire boundary, the hydro-dynamic pressure is only defined modulo a constant and the corresponding pressure-space is \( L^2(\Omega)/\mathbb{R} \).

2 The Solution Approach

2.1 The Galerkin Finite Element Discretization

Our Navier-Stokes solver uses a fully implicit approach for solving the "low-Mach-number" approximation of the compressible Navier-Stokes system (6) (for a detailed description we refer to [BBR, BR]).

The Galerkin finite element method is defined on quadrilateral/hexahedral meshes \( T_h = \{K\} \) covering the domain \( \Omega \). The trial and test spaces \( V_h \subset V \) consist of continuous, piecewise polynomial vector functions (so-called \( Q_p \)-elements) for all unknowns,

\[
V_h = \left\{ \{T, v, p\} \in C(\Omega)^{1+d+1} | T, v|_K \in Q_r^{1+d}, p|_K \in Q_s \right\},
\]

where \( s = 1 \) for \( r = 2 \), and \( s = r - 2 \) for \( r \geq 3 \). Here, \( Q_r \) is the space of (isoparametric) tensor-product polynomials of degree \( r \) (for a detailed description of this standard construction process see [BS]). In order to facilitate local mesh refinement and coarsening, we allow the cells in the refinement zone to have nodes which lie on faces of neighboring cells (Fig. 3). The degrees of freedom corresponding to such "hanging nodes" are eliminated from the system by interpolation enforcing global conformity (i.e., continuity across interelement boundaries) for the finite element functions. For simplicity, we do not allow varying polynomial order across hanging nodes.
Fig. 3. Quadrilateral mesh patch with a “hanging node”

The computation of the strongly anisotropic flow in the rotary engine requires a high resolution in the hole domain which is very time consuming on varying meshes. This difficulty is accounted for by the use of higher-order finite element trial functions with varying orders (“hp-method”). The resulting algebraic systems are lower dimensional with densely filled matrices such that in some stages “direct” algebraic solvers can be used.

We note that by choosing the trial functions for the pressure of (sufficiently) lower degree than of those for the velocity the form \( a(\cdot;\cdot) \) is stable on the discrete spaces \( V_h \) (uniformly in \( h \), i.e., it satisfies the uniform “Babuska-Brezzi inf-sup-stability” condition. This particularly guarantees a stable approximation of the pressure. In the case of equal-order trial functions for \( v \) and \( p \), e.g., the popular \( Q_1/Q_1 \)-ansatz, the scheme requires “pressure stabilization”. In addition, the dominant convection is stabilized by the usual SUPG approach (“streamline upwinding Galerkin”, [HB]). Following this idea, we introduce additional least-squares terms in the continuity equation (least-squares stabilization and streamline diffusion) and in all other equilibrium equations.

In order to formulate this approach in short terms, we write the original system (1)-(3) in the compact form \( L(u)u = f(u) \) with a nonlinear operator \( L(\cdot) \). Then, the stabilization process comprises in the modification of the original semi-linear form \( a(u_h;\phi) \) by a mesh-dependent semi-linear form:

\[
a_\delta(u_h;\phi) := a(u_h;\phi) + \langle Lu_h, S\phi \rangle_\delta,
\]

with a differential operator \( S \) which can be chosen in different ways. Here, we use \( S = -L^* \), and take \( \delta_K \sim h_K \) proportional to the local mesh size,

\[
\delta_K = \alpha \left( \frac{\|Q_K^{-1}\|}{k} + \min \{ \mu, \lambda \} \frac{h^2_K}{h^2_K} + \frac{|v|_{\infty,K}}{h_K} \right)^{-1},
\]

with a constant \( \alpha \sim 0.5 \), and \( k \) denoting a time step when the equations are truly non-stationary or time-stepping is used in solving the stationary equations. The \( \delta \)-dependent inner product is defined as usual by

\[
\langle u, v \rangle_\delta := \sum_{K \in \mathcal{T}_h} \delta_K (u, v)_K.
\]

Accordingly, in the stationary case, we seek a \( u_h \in V_h + v_{b,h} \), satisfying

\[
a_\delta(u_h;\phi_h) = (f(u_h), \phi_h) + \langle f(u_h), S\phi_h \rangle_\delta \quad \forall \phi_h \in V_h.
\]

(7)
Clearly, this ansatz is “consistent” in the sense that the additional terms vanish at the exact solution. This modification serves several purposes: it stabilizes the pressure in the low-Mach-number approximation, it stabilizes the possibly strong convection in the flow, and finally it enhances local mass conservation. This leads to a stable scheme for a wide range of flow conditions.

2.2 Solution of the Algebraic Systems

The nonlinear algebraic system (7) is solved by the damped Newton method. Denoting the derivative of \( a(\cdot;\cdot) \) taken at a discrete function \( u_h \in V_h \) by \( a'(u_h;\cdot,\cdot) \), the arising linear systems have the form

\[
a'(u^i_h; w^i_h, \phi_h) + \langle L'(u^i_h)w_h, S'(u^i_h)\phi_h \rangle_\delta = (r^i_h, \phi_h) \quad \forall \phi_h \in V_h.
\]  

(8)

Here, \( w^i_h \) is the correction and \( r^i_h \) the equation residual of the preceding approximation \( u^i_h \). The updates \( u^{i+1}_h := u^i_h + w^i_h \) are carried until convergence. The linear subproblems (8) are solved by the GMRES method with preconditioning by a multigrid iteration. This multigrid component uses blockwise Gauss-Seidel or ILU smoothing in which the physical unknowns are implicitly coupled on the cell level. For a description of the details of this approach, we refer to [BR]. In computing really nonstationary flows this tide coupling may be lifted by “operator splitting” due to the better conditioning in this case.

2.3 Tests of Solver Components

The discretization described above has first been tested for various model problems of interior flows including the usual “lid-driven cavity” for \( Re = 2000 \) and a new heat-transfer benchmark “temperature-driven cavity” for \( Ra = 10^6 \) comprising large temperature gradients (Fig. 4).

![Fig. 4. Configuration of the two flow benchmarks: “lid-driven cavity” (left), and “temperature-driven cavity” (right)](image-url)
Discretization by higher-order finite elements. Table 2 shows some representative results for the simple test case "lid-driven cavity", while those for the harder test case "heat-driven cavity" are presented in Table 3. It turns out that higher-order finite elements have good potential for accurately computing interior flows even in the presence of strong layers. The solution of the algebraic problems is the bottleneck in using higher-order finite elements within implicit flow solvers. This problem is tackled by using hierarchical multilevel techniques with blockwise smoothers. However, further development is necessary at this point to reach fully satisfactory solution efficiency.

<table>
<thead>
<tr>
<th>FE ansatz</th>
<th>( F_T(u_h) )</th>
<th>#cells</th>
<th>#dofs</th>
<th>#entries</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2/Q1</td>
<td>-0.08657</td>
<td>16,384</td>
<td>14,873</td>
<td>5,994,249</td>
<td>2,100 s</td>
</tr>
<tr>
<td>Q3/Q1</td>
<td>-0.08657</td>
<td>4,096</td>
<td>78,723</td>
<td>4,537,737</td>
<td>1,620 s</td>
</tr>
<tr>
<td>Q4/Q2</td>
<td>-0.08657</td>
<td>1,024</td>
<td>37,507</td>
<td>3,237,897</td>
<td>3,240 s</td>
</tr>
<tr>
<td>Q6/Q4</td>
<td>-0.08657</td>
<td>256</td>
<td>23,043</td>
<td>3,701,769</td>
<td>4,620 s</td>
</tr>
<tr>
<td>Q8/Q6</td>
<td>-0.08658</td>
<td>64</td>
<td>10,851</td>
<td>2,779,785</td>
<td>5,400 s</td>
</tr>
</tbody>
</table>

**Table 2.** Efficiency of higher-order finite elements for solving the "lid-driven cavity" problem with \( \text{Re} = 2000 \) (error \( \sim 1\% \))

<table>
<thead>
<tr>
<th>FE ansatz</th>
<th>( N(u_h) )</th>
<th>#cells</th>
<th>#dofs</th>
<th># entries</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2/Q1</td>
<td>8.843</td>
<td>16,384</td>
<td>214,788</td>
<td>12,069,136</td>
<td>12,324 s</td>
</tr>
<tr>
<td>Q3/Q1</td>
<td>8.859</td>
<td>4,096</td>
<td>115,972</td>
<td>9,558,544</td>
<td>8,765 s</td>
</tr>
<tr>
<td>Q4/Q2</td>
<td>8.871</td>
<td>1,024</td>
<td>54,148</td>
<td>6,587,904</td>
<td>7,484 s</td>
</tr>
<tr>
<td>Q5/Q3</td>
<td>8.893</td>
<td>256</td>
<td>22,084</td>
<td>3,707,536</td>
<td>6,843 s</td>
</tr>
</tbody>
</table>

**Table 3.** Accuracy of higher-order finite elements for solving the "temperature-driven cavity" problem with \( \text{Ra} = 10^6 \) (error \( \sim 1\% \))

---

![Fig. 5. Convection flow in the "heat-driven-cavity": velocity norm (left) and temperature (right)]](image-url)
The time stepping. The simulation of the fully nonstationary flow behavior in the rotary engine, at first, requires a geometric description of the chamber movement. This is obtained in terms of a transformation $\phi_{t'}$ between the time-dependent domains,

$$\Omega_t = \phi_{t'}(\Omega_{t'}).$$

Instead of working on a fixed domain, we consider a variational formulation and discretization on the deforming domain, $\Omega_t$, i.e., the $L^2$-scalar products in (6) are considered on $\Omega_t$ (see [LT]). The discretization is applied in two steps: A spatial semi-discretization of (6) leads to a set of ordinary differential equations which is solved by using a stiffly-stable time stepping scheme like, for example the implicit Euler, the damped Crank-Nicholson, or the second-order Fractional-Step-$\theta$ scheme. Compared to the usual formulation on a fixed domain here an additional term has to be added in the variational formulation (6) which represents the time-variation of the domain $\Omega_t$. After spatial semi-discretization this results in the ODE system

$$(Q\dot{u}_h + Nu_h, \phi_h) + a_\delta(u_h; \phi_h) = (f(u_h), \phi_h) + \langle f(u_h), S\phi_h \rangle_\delta,$$

for all $\phi_h \in V_h$, with the so-called “mesh-velocity operator” $N(t)$. This operator corresponds to the matrix

$$N(t) := \{ (\partial_t \Psi_t D\Psi_t^T \nabla \phi_i(\Psi_t), \phi_j(\Psi_t))_{\Omega_t}\}_{0 \leq i,j \leq l},$$

where $\Psi_t := \Phi_t^{-1}$, with Jacobian $D\Psi_t$, and $\{\phi_i\}_{i=1,\ldots,l}$ is a basis of the finite element space on the mesh for $\Omega_{t'}$. In order to cope with (time dependent) local singularities arising at the intake and exhaust, we use local mesh refinement in these areas (Fig. 7). The mesh refinement is driven by local “error indicators” derived heuristically from properties of the computed solution.

3D-solver components. As a preparatory step for the extension of our Navier-Stokes solver to 3D geometries, several program components for 3D-mesh handling have been implemented:

Fig. 6. Regular 3D meshes for the Wankel motor (half a rotation cycle)
3 Application to the Wankel Motor Configuration

The experimental 2D code has been used for simulating a quasi-stationary model of the Wankel motor, i.e., for a series of different geometries the flow is driven by a prescribed inflow at a small inlet and is left free at an outlet.

Fig. 7. Velocity norm of the flow in the Wankel motor for quasi-stationary flow conditions and locally refined meshes (half of a rotation cycle)
4 Conclusion and Outlook

We have described the main components of a new numerical tool for simulating gas flow and heat transfer in a rotary engine. The method is based on the low-Mach-number approximation of the compressible Navier-Stokes equations and uses a least-squares stabilized finite element discretization of variable order. For the sake of robustness the approach is fully implicit and largely exploits multi-level techniques. At first, a “stationary” flow solver has been developed for interior flows in simple geometries. The natural convection in a box under strong temperature gradients is a prototypical test case. Then, similar computations have been performed for the (still stationary) “Wankel motor” in 2D. The next step is the integration of this quasi-stationary solver component into a time-stepping scheme for computing really nonstationary flows following the variation of the domain. Subsequently, a first 3D version of the code will be compiled.

References


An Efficient Navier–Stokes Solver for Automotive Aerodynamics

R. Rannacher\(^1\), Chr. Becker\(^2\), and S. Turek\(^2\)

\(^1\) Institut für Angewandte Mathematik, Universität Heidelberg, INF 294, 69120 Heidelberg, Germany
E-mail: rannacher@iwr.uni-heidelberg.de, http://numerik.iwr.uni-heidelberg.de

\(^2\) Institut für Angewandte Mathematik, Universität Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany
E-mail: {cbecker,ture}@math.uni-dortmund.de, http://www.featflow.de

Abstract. The aim of this project is the development of efficient and robust numerical methods for the simulation of industrial flows, particularly for incompressible flows in and around vehicles. Since the corresponding flow configurations lead to very huge systems of nonlinear equations with fully nonstationary behavior, the CPU requirements are so high such that special data and matrix structures and hardware-oriented implementation techniques have to be realized. These must be able to exploit a significant percentage of the potentially available computing power of almost 1 GFLOP/s on modern hardware in combination with the powerful FEM discretization and parallel multigrid solution techniques on the mathematical software side. This project is carried out in cooperation with DaimlerChrysler AG at Stuttgart (Dr. M. Wessels, Research and Technology FT1/FB, E222).

1 Description of the Project

The aim of this project is to develop efficient numerical methods and implementation techniques for the simulation of complex flows which are described by the incompressible Navier–Stokes equations. These components have to be integrated into a CFD software package for automotive aerodynamics which is capable of predicting the flow around the car and in the farfield. Preliminary results of the industrial partners based on simulations with different CFD codes and our own experience show that existing research codes and particularly commercial codes have severe problems to provide sufficiently reliable predictions for important flow quantities as the drag and lift coefficient in the fully nonstationary case. Even on today’s supercomputers (see also [5]), the required CPU time is often too large while the results are still too inaccurate.

In Figure 1, the results are shown for a typical steady 3D turbulence calculation with the RNG-\(k-\epsilon\) model which costs 6,5 h CPU time on a SGI Origin2000 with 6 processors for only about 500,000 grid cells. The significant differences between the results from simulation and experiment indicate that much finer computational meshes are needed which in turn will lead to even larger CPU time. Since these codes usually do not contain optimal
<table>
<thead>
<tr>
<th>Quantity</th>
<th>Exp.</th>
<th>Sim.</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drag (('c_d'))</td>
<td>0.165</td>
<td>0.317</td>
<td>92%</td>
</tr>
<tr>
<td>Lift (('c_l'))</td>
<td>-0.083</td>
<td>-0.127</td>
<td>53%</td>
</tr>
</tbody>
</table>

**Fig. 1.** Experiment vs. simulation by DaimlerChrysler obtained with commercial software

multigrid solvers, the ratio between number of grid points and required CPU time increases more than linearly, such that 10 times more unknowns are expected to require at least 20–30 times more CPU time. Since the shown geometrical configuration is still quite simple and nonsteady flow behavior is not considered at all, it can be imagined that the simulation of realistic time-dependent flow behavior is impossible for most of the existing CFD tools, even on the recent supercomputers.

Based on the research code FEATFLOW [11], a new simulation tool for industrial flow configurations is being created which includes modern numerical and algorithmic components for adaptive control of the discretization in space and time, special finite element discretizations and highly efficient multigrid and nonlinear iteration techniques (see [9]). Beside these mathematical tasks we additionally have to apply improved software technologies and implementation strategies which are adapted to modern processor design.

## 2 Computational Bottlenecks and the FEAST Software

One of the main components in iterative solvers are matrix-vector (MV) applications. Hereby, *sparse* concepts are the standard techniques in FEM codes (and others): Depending on the programming language, the matrix entries plus index arrays/lists/pointers are stored as long arrays or heaps, containing the 'nonzero elements' only. For an overview of such techniques, see for instance SPARSKIT [12] and the literature cited therein. While this
sparse approach can be applied for general meshes and arbitrary numberings of the unknowns, no explicit advantage of (possible) highly structured parts can be gained. Consequently, a massive loss of performance compared to the possible peak rates may occur since – at least for large problems with more than 100,000 unknowns – no ‘caching in’ and ‘pipelining’ can be exploited. In this case the higher cost of memory access will dominate the resulting MFLOP/s rates.

To demonstrate this failure, we show some examples for the FEATFLOW code [11] which is one of the most efficient simulation tools for incompressible flow on general domains (see the results in [5]). We apply FEATFLOW to the ASMO2D configuration (see Figure 2) and measure the MFLOP/s rates for the MV multiplication inside of the multigrid solver for the momentum equation (see [9] for algorithmical details). The results in Table 1 show that the computational performance is weak and depends strongly on the problem size and the kind of numbering – Two Level ordering, Cuthill-McKee, STOchastic – while the costs for arithmetic operations, for storage and for the number of memory accesses are identical.

<table>
<thead>
<tr>
<th>computer</th>
<th>#unknowns</th>
<th>TL</th>
<th>CM</th>
<th>STO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN E450</td>
<td>13,688</td>
<td>20</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>54,256</td>
<td>15</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>216,032</td>
<td>14</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>862,144</td>
<td>15</td>
<td>16</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. MFLOP/s rates of sparse MV multiplication in FEATFLOW for different numberings

For a more detailed examination, Table 2 shows typical results for sparse MV multiplication depending on the type of numberings of the unknowns on generalized 3D tensor product meshes: The matrix arises from a discretized scalar Poisson problem with a 27-point stencil. Naturally, rowwise (ROW) numbering is possible (via indexed access), but using the sparse MV techniques, we can apply the different numberings all leading to identical arithmetical costs and numerical results (‘NEQ’ denotes the number of unknowns).

<table>
<thead>
<tr>
<th>computer</th>
<th>NEQ</th>
<th>ROW</th>
<th>TL</th>
<th>STO</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM RS6K/P2SC</td>
<td>17^3</td>
<td>86</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>(160 MHz)</td>
<td>33^3</td>
<td>81</td>
<td>55</td>
<td>16</td>
</tr>
<tr>
<td>‘SP2’</td>
<td>65^3</td>
<td>81</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>‘low cost’</td>
<td>65^3</td>
<td>30</td>
<td>23</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 2. MFLOP/s rates of sparse MV multiplication on tensor product meshes
Based on such studies (see also [7]), we can more precisely characterize the resulting computational run-time behavior of many existing codes in terms of the MFLOP/s rates:

- They are far away from the predicted peak performance.
- They depend significantly on the ‘size’ of the problem and the ‘kind of memory access’.
- ‘Old’ processors (IBM 590) can even be faster (!) than ‘newer’ ones (IBM 597).
- ‘Low cost’ PCs can be significantly faster than processors in ‘supercomputers’!

In addition to Table 2, we further exploit the presence of constant band entries (‘const’) – in contrast to variable matrix entries (‘var’) – as is typical for certain Poisson-type problems. Then, the subsequent results, which are now based on the new Sparse Banded BLAS techniques [8] as part of the FEAST approach [6], show that the same MV applications can actually be performed much faster: We exploit vectorization facilities and data locality [3] which is very important since modern processors are sensible supercomputers with respect to ‘caching in’ and ‘pipelining’.

<table>
<thead>
<tr>
<th>computer</th>
<th>NEQ</th>
<th>‘var’</th>
<th>‘const’</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM RS6K/P2SC</td>
<td>17(^3)</td>
<td>188</td>
<td>480</td>
</tr>
<tr>
<td>(160 MHz)</td>
<td>33(^3)</td>
<td>172</td>
<td>393</td>
</tr>
<tr>
<td>‘SP2’</td>
<td>65(^3)</td>
<td>176</td>
<td>390</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>computer</th>
<th>NEQ</th>
<th>‘var’</th>
<th>‘const’</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTEL PII</td>
<td>17(^3)</td>
<td>51</td>
<td>180</td>
</tr>
<tr>
<td>(400 MHz)</td>
<td>33(^3)</td>
<td>51</td>
<td>137</td>
</tr>
<tr>
<td>‘low cost’</td>
<td>65(^3)</td>
<td>48</td>
<td>124</td>
</tr>
</tbody>
</table>

**Table 3.** MFLOP/s rates of MV multiplication via Sparse Banded BLAS

The FEAST software is conceptualized to combine such highly tuned linear algebra tools with sophisticated FEM simulation strategies (see [10] for a discussion). In contrast to many other approaches which often aim to develop flexible software for research or educational purposes, the FEAST software is designed for high performance applications with industrial background, particularly for Computational Fluid Dynamics (CFD). Consequently, our emphasis lies more on ‘efficiency’ and ‘robustness’ and less on aspects like ‘easy implementable’, ‘flexible’ or ‘most modern programming language’, so that FORTRAN77/90 is used. One of the most important principles in FEAST is the consequent application of (recursive) Divide and Conquer strategies: The solution of a ‘global’ problem is recursively split into smaller independent subproblems on patches as part of the complete set of unknowns. While on certain ‘anisotropic’ parts the usual sparse techniques may be applied (for instance, if local adaptivity is employed), we try to exploit the much higher performance possible on the other – highly structured – patches.
Consequently, the intention is to minimize the number of ‘sparse areas’ and to apply preferably most numerical linear algebra tasks on such ‘structured patches’ via the Sparse Banded BLAS techniques. Hence, the major tasks for realizing such a simulation tool are:

1. Design of the ‘skeleton’ for the recursive splitting into local/global levels.
2. Implementation of the typical FEM facilities on the ‘low level’ patches.

In view of their typically excellent convergence behavior, multigrid methods seem to be most suited for the solution of many PDEs. However, as the previous examples have shown, multigrid on general domains has often poor computational efficiency, at least if the implementation is based on the standard sparse techniques. As a result from our performance measurements (see [7]), the realistic MFLOP/s rates are often in the range of 1 MFLOP/s only, even on modern high performance workstations. Moreover, the linear relationship between problem size and CPU time is hardly realizable, due to the problem-size dependent performance rates of the sparse components. Additionally, the robust treatment of complex mesh structures with locally varying details is often hard to achieve by typical ‘Black Box’ components.

Concerning parallelization some further serious problems occur: The parallel efficiency is often ‘bad’ and far below peak performance, since the solution of the coarse grid problem leads to a large communication overhead. Furthermore, the relation between ‘local’ arithmetic operations and ‘global’ data transfer is poor, in general. Besides these computational aspects, the parallelization of the global smoothers – SOR, ILU – cannot be done efficiently because of their inherent recursive character. So, smoothing can only be performed blockwise, which may lead to a deterioration of the convergence rates. Further, the behavior of such blockwise smoothing is hard to predict for complicated geometries with local and/or global anisotropies. Motivated by these facts, we developed the more general strategy SCARC for solving discretized PDEs aiming for the following goals:

The parallel efficiency shall be high due to a non-overlapping decomposition and a low communication overhead. The convergence rates are required to be independent of the mesh size $h$, the complexity of the domain and the number of subdomains $N$, and they shall be in the range of typical multigrid convergence rates (as $\rho_{MG} \sim 0.1$). Further, the method shall be easily implementable and uses only existing standard methods. The approach shall accomplish the treatment of complicated geometries with local anisotropies (large aspect ratios) without impairment of the overall convergence rates.

The underlying idea is to ‘hide recursively all anisotropies in single subdomains’ combined with ‘block Jacobi/Gauß–Seidel smoothing’ within standard multigrid. This approach is based on the numerical experience that these ‘simple’ block-oriented schemes perform well as soon as all occurring anisotropies are locally hidden, i.e., if the local problems on each block are solved (more or less) exactly. This procedure ensures the global robustness!
On the other hand, this also means that the local solution quality in each block can significantly improve the global convergence behavior. These ideas are combined with hierarchical data and matrix structures, which exploit tensorproduct-like meshes on each *macro* to achieve high performance rates for the linear algebra components in the local (multigrid) solvers. Consequently, all solution processes are recursively organized via sequences of more ‘local’ steps until the lowest level is reached, for instance a single *macro* with the described generalized tensorproduct mesh. Consequently, the complete SCARC approach ([1],[4]) can be characterized as follows:

**Scalable** (with respect to ‘quality and number of local solution steps’ at each stage)

**Recursive** (‘independently’ for each stage in the hierarchy of partitioning)

**Clustering** (for building blocks via ‘fixed or adaptive blocking strategies’)

The main philosophy is to recursively split the global problem into a sequence of local problems, preferably on *macros* which allow the (locally) robust and efficient solution of the auxiliary problems at very high MFLOP/s rates. The number of arithmetic operations tends to be larger than for the standard approach, but these local operations are cheap: *Data moving, not data processing is costly!*

It is obvious that this 'new' parallel approach requires a modified concept of load balancing. In general, there will be more *macros* than processors – realistic coarse meshes in 3D require at least 1,000 up to 10,000 elements – such that each processor will provide the data for several *macros*. First of all, the amount of storage on each *macro* is not identical if for instance the case of variable or constant matrix coefficients can be exploited on the tensorproduct-like meshes, or if auxiliary storage for the *sparse* techniques is needed. Moreover, the application of ILU requires different storage cost than Jacobi- or Gauß–Seidel smoothing. However, the actual CPU time on each *macro* is much more decisive for the solution of the local problems: The elapsed time depends significantly on the corresponding computational efficiency and on the numerical efficiency, i.e. the actual convergence rates which determine the number of multigrid sweeps, and hence the total number of MFLOPs. Additionally, the problem size on each *macro* may vary.

Hence, parallel load balancing is much more complex for this SCARC approach and cannot be determined via *a priori* strategies, for instance, by equilibrating the number of unknowns on each processor. Instead, we have to perform *a posteriori* load balancing techniques, analogously to the adaptive approaches, which are based on the numerical and computational run-time behavior of the last iteration to better equilibrate the total required CPU time on each processor. More information and computational studies for implementations of SCARC are part of the work of Chr. Becker [1] and S. Kilian [4].
3 Numerical Results

We have concentrated on the following aspects which all are part of the FEAST software with special applications to the configurations given by the industrial partner:

1. Further development of FEAST as FEM basic package with direct integration of parallel ‘low level’ components.
2. Realization and optimization of the SPARSE BANDED BLAS library for specified hardware platforms, particularly for parallel computers (SUN ENTERPRISE, CRAY T3E/Jülich, AMD-K7 Linux-Cluster) and vector computers (CRAY T90/Jülich).
3. Integration of the SPARSE BANDED BLAS components into complete multigrid solvers as part of FEAST with special techniques for body-fitted meshes with locally high aspect ratios.

Figure 2 shows a typical macro decomposition (= coarse grid) for the 2D configuration (ASMO2D) and a zoom of some macro element (= 1 quadrilateral) directly at the upper boundary of the car. During the mesh generation process, we have applied locally anisotropic refinement in normal direction towards the boundary such that boundary layers can be better resolved while at the same time the tensorproduct topology is locally preserved.

Due to the local grid distortion, special multigrid components have to be employed which work efficient and robust w.r.t. such mesh anisotropies. In the framework of SCARC as generalized multigrid approach, particularly on parallel platforms, the corresponding local problems (on each refined macro

Fig. 2. Macro decomposition and anisotropically refined macro element (3 refinement levels with exact representation of the curved boundary) for the 2D-ASMO configuration
of the coarse grid) are treated with special linewise Gauß-Seidel schemes as smoothing operations inside of an optimized local multigrid solver. Table 4 shows typical (parallel) multigrid rates for the solution of Pressure–Poisson problems within the framework of an nonstationary Navier–Stokes approach via discrete projection methods [9]. The global as well as the local multigrid convergence rates are more or less independent of the degree of (local) mesh distortion while the resulting CPU times, i.e. the MFLOP/s rates, are not yet completely optimized.

The multigrid convergence rates in Table 4 show the robustness and efficiency of this parallel ScARC approach: ‘I’ corresponds to regular refinements of the macro decomposition in Figure 2, while configuration ‘II’ contains locally anisotropic refinements (see the right example in Figure 2); ‘AR=...’ denotes the minimum and maximum aspect ratio for each configuration, and the averaged multigrid rates for the solution process on the corresponding macro element are given. The resulting CPU times in the right most Table already indicate that high efficiency rates will be achieved, which will be further improved by optimizing the administration of the local multigrid solver and particularly by including the corresponding Sparse Banded BLAS techniques: At the moment, the MFLOP/s rates for the local multigrid solvers on each tensorproduct mesh are of order 25 MFLOP/s and shall be

<table>
<thead>
<tr>
<th>$N_l$</th>
<th>$ScaRC_g$</th>
<th>$T$</th>
<th>$N_l$</th>
<th>$TGS_1$</th>
<th>$T$</th>
<th>$N_l$</th>
<th>$TGS_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.03</td>
<td></td>
<td>32</td>
<td>0.05</td>
<td></td>
<td>32</td>
<td>0.01</td>
</tr>
<tr>
<td>64</td>
<td>0.04</td>
<td></td>
<td>64</td>
<td>0.06</td>
<td></td>
<td>64</td>
<td>0.01</td>
</tr>
<tr>
<td>128</td>
<td>0.04</td>
<td>AR=1</td>
<td>64</td>
<td>0.06</td>
<td></td>
<td>64</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.03</td>
<td></td>
<td>32</td>
<td>0.01</td>
<td></td>
<td>32</td>
<td>0.01</td>
</tr>
<tr>
<td>64</td>
<td>0.04</td>
<td>AR=10</td>
<td>64</td>
<td>0.01</td>
<td></td>
<td>64</td>
<td>0.01</td>
</tr>
<tr>
<td>128</td>
<td>0.04</td>
<td></td>
<td>128</td>
<td>0.01</td>
<td></td>
<td>128</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 4. Multigrid rates (global ($ScaRC_g$), local (with $TGS_1$, see Table 5)), resulting CPU times for 2.5 millions unknowns on the CRAY T3E depending on the number of processors (“#P”) and partitions used on the CRAY T3E/36 proc. and SUN E3500/8 proc.
An Efficient Navier–Stokes Solver for Automotive Aerodynamics 51

improved by the optimized Sparse Banded BLAS up to 50–100 MFLOP/s per processor. The figures in Table 4 show the employed partitions (CRAY T3E/36 proc. + SUN E3500/8 proc.) which are not yet optimal, too, as can be seen from the parallel efficiency. At the moment, the realized MFLOP/s rates for the complete parallel SCARC solver vary between 80 MFLOP/s (4 proc.) and 352 MFLOP/s (36 proc.). By optimization of the implementation of the global SCARC approach (including better loadbalancing strategies) and employing hardware-oriented adaptations of the Sparse Banded BLAS, we expect further improvements in the range of factors 10–50.

While the numerical components seem to be sufficiently efficient, the computational efficiency in terms of the MFLOP/s rates still requires further improvement. Table 5 shows the recent efficiency rates for components from the optimized Sparse Banded BLAS library (matrix-vector multiplication, linewise Gauß–Seidel preconditioner and complete multigrid scheme); here, we can additionally distinguish between the cases of general and of constant matrix entries. It is remarkable that special algorithms and implementation techniques had to be developed for the TGS preconditioner on the CRAY T90. It turns out that multigrid, even with this very robust smoother, can be very efficiently performed. Only on the vector computer CRAY T90, multigrid with (local) Jacobi smoother shows significantly higher MFLOP/s rates, which is to be expected. Additionally, these results show that the actual MFLOP/s rates can significantly depend on the local problem size due to different cache exploitation. This fact can be employed when applying the

<table>
<thead>
<tr>
<th></th>
<th>#NEQ</th>
<th>MVV</th>
<th>MVC</th>
<th>TGSV</th>
<th>TGSC</th>
<th>MGJV</th>
<th>MGJC</th>
<th>MGTV</th>
<th>MGTC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cray T3E</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(600 MHz)</td>
<td>65^2</td>
<td>82</td>
<td>391</td>
<td>59</td>
<td>143</td>
<td>96</td>
<td>391</td>
<td>84</td>
<td>281</td>
</tr>
<tr>
<td>‘MPP’</td>
<td>257^2</td>
<td>72</td>
<td>293</td>
<td>49</td>
<td>118</td>
<td>75</td>
<td>194</td>
<td>70</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>1025^2</td>
<td>47</td>
<td>282</td>
<td>35</td>
<td>86</td>
<td>54</td>
<td>163</td>
<td>51</td>
<td>142</td>
</tr>
<tr>
<td><strong>Cray T90</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(440 MHz)</td>
<td>65^2</td>
<td>855</td>
<td>414</td>
<td>51</td>
<td>52</td>
<td>842</td>
<td>450</td>
<td>183</td>
<td>162</td>
</tr>
<tr>
<td>‘Vector’</td>
<td>257^2</td>
<td>932</td>
<td>887</td>
<td>124</td>
<td>125</td>
<td>912</td>
<td>698</td>
<td>275</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>1025^2</td>
<td>939</td>
<td>1091</td>
<td>223</td>
<td>215</td>
<td>930</td>
<td>950</td>
<td>422</td>
<td>418</td>
</tr>
<tr>
<td><strong>SUN E3500</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(400 MHz)</td>
<td>65^2</td>
<td>298</td>
<td>520</td>
<td>67</td>
<td>70</td>
<td>262</td>
<td>382</td>
<td>158</td>
<td>188</td>
</tr>
<tr>
<td>‘Shared’</td>
<td>257^2</td>
<td>271</td>
<td>588</td>
<td>64</td>
<td>69</td>
<td>252</td>
<td>408</td>
<td>153</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>1025^2</td>
<td>51</td>
<td>224</td>
<td>39</td>
<td>65</td>
<td>62</td>
<td>153</td>
<td>57</td>
<td>117</td>
</tr>
<tr>
<td><strong>AMDK7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(700 MHz)</td>
<td>65^2</td>
<td>182</td>
<td>522</td>
<td>110</td>
<td>159</td>
<td>188</td>
<td>444</td>
<td>167</td>
<td>317</td>
</tr>
<tr>
<td>‘LINUX’</td>
<td>257^2</td>
<td>82</td>
<td>283</td>
<td>55</td>
<td>127</td>
<td>87</td>
<td>189</td>
<td>83</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>1025^2</td>
<td>68</td>
<td>237</td>
<td>51</td>
<td>98</td>
<td>65</td>
<td>131</td>
<td>63</td>
<td>125</td>
</tr>
</tbody>
</table>

Table 5. MFLOP/s rates for different problem sizes (#NEQ = number of grid points) for matrix-vector applications (MV), linewise Gauß–Seidel preconditioner (TGS) and complete multigrid solver (MGJ with Jacobi smoother, MGT with TGS smoother); additionally, we can distinguish between variable (V) and constant (C) matrix entries (Poisson problems on orthogonal meshes)
Fig. 3. Mesh partitioning on the CRAY T3E (64 proc.) and the SUN E3500 (8 proc.) and snapshots of the pressure and the norm of velocity
global SCARC solver since the choice of the local number of macros and the corresponding problem size can be adapted to these effects.

Beside these 2D efforts in the framework of the new FEAST software, we have improved the existing parallel 3D flow solver which is based on FEATFLOW. In particular, for the CRAY T3E (Jülich) and for ‘small’ parallel computers like the SUN E3500 (8 proc., shared memory), the code has been further tuned (see [2] for a documentation). The following pictures show results of nonsteady 3D simulations for a prototypical configuration on different parallel platforms which require different partitions of the computational mesh. This parallel code will be the candidate for all further developments which aim to incorporate the FEAST techniques into this CFD tool.

4 Outlook

The central point of this project has been the development of mathematical components – FEM discretizations, adaptivity and multigrid solvers – and their realization in a software package which directly includes tools for parallelism and hardware-adapted high-performance in low level kernel routines. The code generation uses the new FEAST software in order to achieve highest computational efficiency. These software developments can be viewed as ‘basic research’ in the field of mathematical software for PDEs. It is the special goal in this project to realize and to optimize the SPARSE BANDED BLAS concepts for specific parallel computers (CRAY T3E, CRAY T90, SUN E3500, AMD K7) and to adapt the mathematical components to the ASMO configurations given by the industrial partner. Since the corresponding geometry leads to severe mesh distortions near the boundary, special multigrid components for highly-stretched body fitted meshes have been developed which allow high numerical efficiency and robustness in the multigrid-like SCARC approach.

At the end of this project, we will have completed the 2D Navier–Stokes solver which is based on conforming bilinear FEM and discrete projection methods. This code will be integrated into FEAST such that the parallel/sequential high-performance tools of the SPARSE BANDED BLAS will be directly available. Additionally, the indicated adaptive load-balancing concepts will be realized. Besides that, we will continue our work with the parallel 3D adaptation of the FEATFLOW solver which is presently applied on several parallel computers to prototypical car configurations similar to the 2D-ASMO geometry. This parallel 3D code will be our candidate for all further developments which aim to incorporate the high-performance FEAST techniques into this CFD tool in order to achieve highest computational efficiency on modern computers in combination with the ‘best’ numerical approaches.

References

12. SPARSKIT (by Y. Saad), http://www.cs.umn.edu/Research/arpa/SPARSKIT.
Numerical Simulation of Exhaust Systems in Car Industry – Efficient Calculation of Radiation Heat Transfer

S. Rjasanow and M. Bebendorf

Department of Mathematics, Saarland University, Postfach 15 11 50, 66041 Saarbrücken, Germany
E-mail: rjasanow@num.uni-sb.de

Abstract. The aim of this project is a complete yet simple numerical model for the heat transfer in a system of exhaust pipes of an automobile. The industrial partner Tenneco Automotive, H. Gillet GmbH at Edenkoben, uses this simulation for optimising the construction of the exhaust system in order to improve the efficiency of the catalytic converters. In this model forced convection of the exhaust gas, the heat conduction and the heat transfer due to radiation are taken into account. For the effective numerical solution of the boundary integral equation for the radiation heat transfer a method that is based on matrix compression is developed. Some numerical examples for the matrix compression and calculations using a developed software package are presented.

1 Introduction

Catalytic converters play an important role in the reduction of pollutants in the exhaust of an automobile engine. Modern catalytic converters are very efficient, the reduction of \( CO \), \( NO_x \) and of various hydrocarbons contained in the exhaust being more than 90\%. However, during the warm-up cycle the temperature of the catalytic converter is too low and the efficiency of the converter can be further improved by shortening the warm-up cycle. There are different technical ideas for this purpose. One of the most popular is to use insulated exhaust pipes first introduced in 1991 [1]. The insulating split between two pipes should prevent the energy produced by the engine from getting lost on the way to the catalytic converter. The three-dimensional geometry of an insulated exhaust pipe is presented in Fig. 1. Because of current and future government regulations for emission control there is permanent urgency not only to improve the efficiency of exhaust systems but also to predict output during the development of new automobile models. Thus the mathematical modelling and numerical simulation of exhaust systems become more and more important. In the previous paper [6] we discussed the simplified situation of the straight insulated exhaust pipe. The special structure of the problem leads to special properties of the discrete system, i.e. the matrix is of Toeplitz structure. Using these properties we have designed a fast numerical algorithm for the numerical simulation of the warm-up cycle.

W. Jäger et al. (eds.), Mathematics - Key Technology for the Future
© Springer-Verlag Berlin Heidelberg 2003
In this paper we consider the most general situation of pipes having the geometry which is really used in industry.

The paper is organised as follows. In Section 2 we give a short description of the mathematical model for the heat transfer in an insulated exhaust pipe. In Section 3 we give a short review of the different approximation techniques for full dense matrices. In Section 4 we present the main ideas of our new approximation technique, called Adaptive Cross Approximation, and finally, in Section 5, we present the results of computations and draw some conclusions.

2 Mathematical Model

We consider an insulated exhaust pipe of length $L$ having the cross-section depicted in Fig. 2. We consider the following physical processes in the pipe:

1. Heat transfer due to the forced convection in the inner pipe;
2. Heat transfer due to conduction within and between both walls of the exhaust pipe;
3. Heat transfer due to radiation in the insulating split.

The exact description of the first two equations, the corresponding boundary and initial conditions as well as the standard numerical procedures can be found in [6]. Here we concentrate on the numerical solution of the radiation heat transfer equation which can be written in the following form:

$$((I - (1 - \epsilon)B)R)(y) = (1 - \epsilon)(B^2U)(y)$$

(1)
where \( R(x), x \in \Gamma \) denotes the reflected part of the radiation energy on the inner surfaces of the insulating split \( \Gamma \). The integral operator \( \mathcal{B} \) is defined as

\[
(\mathcal{B}R)(y) = \frac{1}{\pi} \int_{\Gamma} \kappa(x, y) \frac{(n_x, y - x)(n_y, x - y)}{|x - y|^4} R(x) \, ds_x, \quad y \in \Gamma
\]

whereby the function \( U(x) \) is an abbreviation of

\[
U(x) = \epsilon \sigma T^4(x), \quad x \in \Gamma.
\]

The constant \( \epsilon \) is the emissivity of the steel and \( \sigma \) denotes the Boltzmann constant \( \sigma = 5.669 \cdot 10^{-8} \). The numerical solution of equation (1) begins with the discretisation of the surface \( \Gamma \) using a system of plane, triangle panels

\[
\Gamma \approx \Gamma_h = \bigcup_{j=1}^{N} \Gamma_j
\]

The centres of the mass of the panels \( \Gamma_i \) build the system of collocation points

\[
y_i = \frac{1}{3} \left( x_i^{(1)} + x_i^{(2)} + x_i^{(3)} \right), \quad i = 1, \ldots N.
\]

The piecewise constant ansatz \( r \) for the unknown function \( R(x) \) and \( u \) for the function \( U(x) \) leads to the following system of collocation equations

\[
Ar = f, \quad A \in \mathbb{R}^{N \times N}, \quad r, f \in \mathbb{R}^{N}.
\]

The matrix \( A \) and the right hand side of this system are of the following form:

\[
A = I - (1 - \epsilon)B, \quad f = (1 - \epsilon)B^2u.
\]

The elements of the matrix \( B \) are defined as

\[
b_{i,j} = \frac{1}{\pi} \int_{\Gamma_j} \kappa(x, y_i) \frac{(n_x, y_i - x)(n_{y_i}, x - y_i)}{|x - y_i|^4} \, ds_x.
\]

An efficient way of calculating the entries \( b_{i,j} \) is presented in [10].
3 Approximation of the Full Dense Matrices

The system matrix \( A \) of the usual boundary element method has the following properties. The matrix is dense, i.e. the memory requirements are 
\[
\text{Mem}(A) = O(N^2),
\]
where \( N \) denotes the matrix size. The numerical solution of the resulting system of linear equations \( Ay = b \) using classical direct solvers is extremely expensive, i.e. the numerical work is 
\[
O(p(y = A^{-1}b)) = O(N^3).
\]
The only possible increase in efficiency is due to using iterative methods for the resulting system of linear equations. Iterative methods like Krylov subspace methods highly depend on efficient procedures for the fast matrix-vector multiplication. The first fast matrix-vector multiplication tool, the **Fast Multipole Method**, was presented in 1985 by V. Rokhlin [12],[13]. The problem was to compute the force applied to each of the large number of charged particles in each time step. The main ideas here were to divide the whole space into a near and a far field and to cluster particles in the far field. The Fast Multipole Method is widely used in many applications (see also [14]). **The Panel Clustering** algorithm for boundary integral equations, first published by Hackbusch and Novak [4], is based on very similar ideas.

3.1 Matrix Partitioning

Let \( P = \{\Gamma_1, \ldots, \Gamma_N\} \) be the set of panels. We first subdivide the index set \( I \times I \), where \( I = \{1, \ldots, N\} \), into subsets \( t_i^1 \times t_i^2 \), \( i = 1, \ldots, n \) so that with \( S_t = \{y_i, i \in t\} \) either the admissibility condition
\[
\text{diam } S_{t_i^1} \leq \eta \text{dist}(S_{t_i^1}, S_{t_i^2})
\]
is fulfilled or one of the index sets \( t_i^1 \) and \( t_i^2 \) has just one element.

In [4] a set of clusters \( T \) that possesses a tree structure is used to suitably subdivide the set \( P \) with respect to a fixed point. We will use a set of cluster pairs having a tree structure for the partitioning of \( P \times P \) (the Cartesian product of the set of panels with itself). This set \( T' \) is constructed from the set \( T \) by applying the following recursion to \( (\Gamma_h, \Gamma_h) \). Take a cluster pair \( (t_1, t_2), t_1, t_2 \in T \). If \( t_1 \) and \( t_2 \) both have children \( t_{11}, t_{12} \) and \( t_{21}, t_{22} \) in \( T \) respectively, then assign the pairs \( (t_{11}, t_{21}), (t_{11}, t_{22}), (t_{12}, t_{21}) \) and \( (t_{12}, t_{22}) \) as children to the cluster pair \( (t_1, t_2) \) and add them to \( T' \). Now repeat these steps with each child. It is obvious that the covering of \( \Gamma_h \times \Gamma_h \) with the smallest number of cluster pairs can be computed by the following algorithm.

Set \( D = \emptyset \) and call \( \text{Partition}((I, I), D) \), where \( \text{Partition} \) is the following recursive procedure.

\[
\text{procedure } \text{Partition}((t_1, t_2), D) \\
\text{if } (t_1, t_2) \text{ is admissible} \\
\text{then } D := D \cup \{(t_1, t_2)\} \\
\text{else apply the procedure to each of the children of } (t_1, t_2) \text{ in } T'.
\]
Each generated index pair \((t_1, t_2) \in D\) is admissible. Recently Hackbusch and Khoromskij \([2], [3]\) have proven that the storage requirement for such matrices is \(O(N \log N)\). However, they did not investigate the dependency on the parameter \(\eta\) which is essential for our algorithm since the approximation error will be controlled by this parameter.

In \([9]\) it is shown that under the assumption of a quasi-uniform panel set \(P\), i.e. there is a constant \(c_u\) so that

\[ c_u \text{diam } \Gamma_j > h, \text{ for all } \Gamma_j \in P, \]

where \(h = \max_{\Gamma_j \in P} \text{diam } \Gamma_j\), for the number of generated blocks \(N\) it holds that

\[ N = O(N^{1+\alpha} \eta^{-2(d-1)}), \quad \forall \alpha > 0. \]

The partitioning of the matrix can be done so that the number of operations for this purpose is of the same order as \(N\).

### 3.2 Low-Rank Approximation

If the generating function \(\kappa: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}\) is asymptotically smooth (see \([11]\)), i.e. there are constants \(c_1, c_2 > 0\) and \(g \in \mathbb{R}\) so that for any multi-index \(\alpha \in \mathbb{N}_0^d\)

\[ |\partial^\alpha_y \kappa(x, y)| \leq c_1 \, p! \, c_2^p \, |x - y|^{g - p}, \quad p = |\alpha|, \]

it can be shown using a Taylor expansion that an admissible block can be well approximated by low-rank matrices. It is known that the best low-rank approximant is obtained by using singular value decomposition. But for SVD it is necessary to generate the whole block, which finally leads to a complexity of order \(N^2\).

In \([9]\) we proposed an algorithm that computes a low-rank approximant using only a small part of the original matrix.

Assume that we have an admissible block of dimension \(m \times n\). For \(k = 1, 2, \ldots\) calculate

\[
(\bar{u}_k)_i = \kappa(x_i, y_{jK}) - \sum_{l=1}^{k-1} (v_l)_{jK}(u_l)_i, \quad i = 1, \ldots, m
\]

\[
u_k = (\bar{u}_k)_{K}^{-1} \bar{u}_k
\]

\[
(v_k)_j = \kappa(x_{ik}, y_j) - \sum_{l=1}^{k-1} (u_l)_{ik}(v_l)_j, \quad j = 1, \ldots, n,
\]

where in each step \(j_k\) is chosen so that \(\bar{u}_k \neq 0\) and \(i_k\) so that \(|(\bar{u}_k)_{i_k}| \geq |(\bar{u}_k)_i|\). For the approximant \(S_k\) it holds that

\[ S_k = \sum_{l=1}^{k} u_l v_l^T. \]
Thus for the whole approximation we need only the evaluation of \( \kappa(x_i, y_{jk}), \)
\( i = 1, \ldots, m, \) and \( \kappa(x_{ik}, y_j), j = 1, \ldots, n. \) The rest are algebraic transformations, which are easy to implement, whereby it should be remembered that the entries of \( u_k \) at the positions \( i_l \) and the entries of \( v_k \) at the positions \( j_l, l < k \) are zero.

To obtain \( S_k \) we need \((m + n)k\) units of storage and \( O((m + n)k^2) \) operations to generate the approximant \( S_k. \)

In [9] it has been shown that using this method to generate an approximant of accuracy \( \varepsilon, \) i.e. \( \|A - \tilde{A}\|_F \leq \varepsilon, \) we need

\[
N_{st} = O(N^{1+\alpha}\varepsilon^{-\alpha})
\]

storage and the number of operations for a matrix-vector multiplication and for the generation of the approximant does not exceed \( O(N^{1+\alpha}\varepsilon^{-\alpha}). \)

4 Numerical Examples

We begin this subsection with the following trivial situation. Let \( \Gamma \) be the inner surface of the unit sphere

\[
\Gamma = \{ x \in \mathbb{R}^3, |x| = 1 \}.
\]

Then the kernel

\[
\frac{1}{\pi} \frac{(n_x, y - x)(n_y, x - y)}{|x - y|^4}
\]

of the radiation heat transfer equation can be simplified as follows:

\[ n_x = -x, \; n_y = -y \]

and therefore

\[
\frac{1}{\pi} \frac{(n_x, y - x)(n_y, x - y)}{|x - y|^4} = \frac{1}{\pi} \frac{1 - (x, y))(1 - (y, x))}{(2 - 2(x, y))^2} = \frac{1}{4\pi}.
\]

Thus, any appropriate discretisation of this problem will lead to the matrix \( B \) of the rank one.

For the numerical tests of the Adaptive Cross Approximation procedure we will consider a geometry of a simple pipe given by the following parametric representation

\[
\Gamma = \left\{ x \in \mathbb{R}^3, x = \begin{pmatrix} \cos(2\pi t) \\ \sin(2\pi t) \\ 4z \end{pmatrix}, (t, z) \in (0, 1)^2 \right\}.
\]

The panels are plane triangles appearing after the canonical discretisation of the parameter domain \((0, 1)^2\) using \( n_t \) and \( n_z \) discretisation points. The whole number of panels is then \( N = 2n_t n_z. \) In the following table we present the
results of the matrix compression with accuracy $\varepsilon = 10^{-6}$ in the Frobenius norm achieved by the Adaptive Cross Approximation procedure. The first column of this matrix contains the dimension of the matrix $N$, the second column gives the amount of memory in MBytes which would be needed without approximation, the third the corresponding value for the compressed matrix, and the last the percentage of memory used.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>MBytes</th>
<th>MBytes</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>2</td>
<td>2.0</td>
<td>100.0%</td>
</tr>
<tr>
<td>1024</td>
<td>8</td>
<td>5.0</td>
<td>63.0%</td>
</tr>
<tr>
<td>2048</td>
<td>32</td>
<td>12.6</td>
<td>39.5%</td>
</tr>
<tr>
<td>4096</td>
<td>128</td>
<td>27.4</td>
<td>21.4%</td>
</tr>
<tr>
<td>8192</td>
<td>512</td>
<td>63.6</td>
<td>12.4%</td>
</tr>
<tr>
<td>16384</td>
<td>2048</td>
<td>163.8</td>
<td>8.0%</td>
</tr>
<tr>
<td>32768</td>
<td>8192</td>
<td>380.9</td>
<td>4.6%</td>
</tr>
</tbody>
</table>

4.1 Software Package

A main purpose of our project was to develop a software package that allows our partners from industry to simulate an arbitrary exhaust system. To this end, it is necessary to provide a tool for the input of all relevant data. The

Fig. 3. Modelling and simulation tool
developed software allows a complete description of the tube’s geometry, i.e. lengths, bending radii, bending angles, materials, insulations etc. through a graphical user interface. Figure 3 shows this tool. It is possible to make time dependent calculations for arbitrary input profiles given by temperature, mass flow and pressure. As the result of the calculation the temperature of both gas and wall at specified points within the system can be traced and kept for later comparisons.

References

1. T. Nording. Neuartiges Konzept für Abgaskrümmer, Vorrohre und Kata-
8. S. Rjasanow. The Structure of the Boundary Element Matrix for the Three-
11. A. Brandt. Multilevel computations of integral transforms and particle inter-
Combinatorial Optimization Techniques for Three-Dimensional Arrangement Problems

Thomas Lengauer\textsuperscript{1,2} and Mike Schäfer\textsuperscript{1}

\textsuperscript{1} University of Bonn, Department of Computer Science, Römerstraße 164, 53117 Bonn, Germany
\textsuperscript{2} GMD German National Research Center for Information Technology, SCAI Institute for Algorithms and Scientific Computing, Schloß Birlinghoven, 53754 Sankt Augustin, Germany

Abstract. This paper presents two approaches for the automated layout of three-dimensional objects in space. The goal is to achieve high packing densities and fitting of objects in predefined design spaces while satisfying technological side constraints. The focus is on small-sized problem instances (up to 20 objects) with complex, possibly non-convex shapes. Linear programming methods form the common ground of our approaches.

The first approach is a global optimization algorithm based on the branch-and-bound paradigm. We introduce a discretization of the configuration space of all possible arrangements which facilitates a complete enumeration of solutions. The bounding procedure then allows for a drastic reduction of the search space. We use a limited number of discrete object orientations within this method.

The second approach is a local optimization scheme which starts out from a given initial arrangement and is capable to perform continuous object rotations. It is based on a linearization of orthonormal rotation matrices. We also present a perspective for combining global optimization and continuous object rotations.

Examples in this paper are taken from the automobile industry but applications are not limited to this area. Various objective functions may be optimized, including the volume and the location of the center of gravity. We also show how to integrate a wiring area estimation into the global optimization procedure.

1 Introduction

Arrangement problems occur in very different application areas and with a great variety of side constraints. Because of their practical relevance and challenging intellectual properties they have attracted substantial interest in different scientific disciplines. Moreover the design of space-efficient layouts is an increasingly important issue in diverse industrial fields like mechanical engineering and electromechanical design.

We start with a brief overview and a rough classification of arrangement problems to point out related work and to supply some additional links for the reader. A first classification can be done by means of the problem sizes tackled and the geometric complexity of the shapes involved. We distinguish small/large scale problems of identical/different objects with regular/complex geometries in two and three dimensions.
The study of moderately sized problem instances of rectangular and box-shaped objects has a long history in the field of operations research. Examples include the classical bin-packing, knapsack and pallet loading problems. A comprehensive and detailed review of the state of this field can be found in the EJOR Special Issue on Cutting and Packing [BW95]. Two-dimensional knapsack problems with approximately 100 rectangles of 30 different types are now within the scope of optimal solution approaches (see [FS97]). The work presented in this paper differs from these problem types by considering non-orthogonal, non-convex objects and integrating connectivity (e.g. wiring) into the layout problem.

Quite remarkable computational results have also been achieved for large scale problems in VLSI chip layout (see [Len90]). Object shapes are typically rectangular and connectivity plays an important role in the layout process. Partitioning heuristics are often applied to minimize the number of wires crossing the boundary between adjacent subregions. In general, manual handling of these problems is infeasible which stimulated the development of automated methods. In contrast, our work focuses on three-dimensional module assemblies. The problem instances are characterized by a small number of modules and few connections with large cable diameters such that partitioning techniques are not appropriate here.

In the following, we consider the class of small-sized instances with different, complex geometries. Human judgement is very good on these problems, especially for problems with recurring pattern types, and often leads to optimal or near optimal solutions. One case in point here is the computation of cutting images for the textile industry, a two-dimensional layout problem. Our object here is to minimize the amount of wasted fabric. Large sets of manually generated solutions and rigorous mathematically proven bounds for the optimal solutions are available for this application (see [HL98]). Commonly it is quite hard to develop computational methods which are competitive to human experts with respect to solution time and quality. Viewed in this light we cannot promise to improve the quality of the solutions by a substantial amount through the use of computational methods. Rather, one reason to use computational methods is their smooth integration into automated industrial processes. Another advantage is their use as a sort of "advisory"-tool which generates a number of layout variants from which the developer or engineer may choose the best suited one and manually customize it to more specific demands and side constraints. The local optimization procedure then can finally be used to compact the given arrangement which is a time consuming manual task in today’s CAD-systems. This method can also be used to fit a single component into a prespecified design space.

This paper is organized as follows: Section 2 presents the general structure and reasoning behind our two-phase approach to component layout and relates it to the work of other authors considering similar problem types. Section 3 is dedicated to the global optimization procedure and introduces the basic principle of this method. The applicability of this method is illustrated
using the application of designing E-Module-Boxes which includes wiring area estimation. Section 4 is devoted to the local optimization procedure. Results for an industrial test case of module arrangements in the trunk of a car are presented. Finally we present some perspectives to combine global optimization and continuous object rotations in section 5.

2 General Approach and Related Work

The geometric complexity of objects has two important implications for the solution procedure: first of all as object geometries become more complex the combinatorial part of the problem becomes harder because of the growth of the solution space and the lack of efficient computational procedures for NP-hard problems. We use an object hierarchy at different levels of resolution and allow for only a limited number of object rotations to reduce these effects. This leads to a two-phase approach in a natural way. The principal arrangement is fixed in a first phase and the detailed placement and routing of wires is postponed to a second phase. The global and local optimization procedures presented in this paper reflect this two-phase approach.

The second algorithmic aspect influenced by the geometric complexity is enforcing the "solidness" of objects by means of collision detection or object intersection algorithms. There are basically two approaches to tackle this problem: intersection avoidance or intersection detection and correction. The first approach circumvents the intrinsic computations needed to calculate the actual shape of the intersection or other measures like the intersection depth of polyhedra. Rather, our methods avoid intersections by enforcing a linear separation between objects making use of the well known linear separation theorem for convex objects.

We use linear programming methods for the problem formulation and as our basic optimization tool. However, our methods are not limited to convex objects, because non-convex shapes can be modeled by a set of, possibly overlapping, convex parts. Non-polyhedral objects can be approximated with the aid of polyhedral mesh generation tools. While translation and rotation themselves are linear transformations, the optimization of the rotation parameters is not, and we introduce a suitable linearization in the local optimization procedure.

The main functional difference between the global and local optimization procedure is that, within the global optimization algorithm, the object orientations are limited to a discrete, prespecified set. Optimizing the orientation is then subject of the second phase. We use the term arrangement algorithm for global optimization because, in principle, all possible arrangements can be enumerated or generated. The local optimization procedure applied in the second phase is more accurately called a compaction algorithm and starts out from a given initial arrangement of objects. Here in general the relative arrangement of objects, like "P is to the right of Q", is not changed.
There is only a little related work considering packing problems in mechanical engineering. A simulated annealing approach to component packing can be found in [KCR96] and [SC97]. Simulated annealing is a global optimization procedure, which repeatedly applies local adjustments, called "moves". With increasing time it then converges to a solution. The objects are allowed to intersect in the early stages of the algorithm and are separated with time. While allowing and correcting intersections is considered to be a disadvantage of this method, simulated annealing is quite flexible with respect to the integration of new cost criteria and side constraints. On the other hand, it only offers one solution per run, while our approach generates all possible solutions in principle and may present a collection of substantially different layouts to the user, who may judge them by additional criteria which are not formalized in the objective function. In addition, our algorithm offers a quality certificate in the form of the generated lower bounds. Wiring is not considered in the aforementioned papers but has been integrated in [CDY98] and [SCW98]. Two local optimization procedures are presented in [LB94] and [SD88]. In contrast to our linearized problem formulation they use non-linear models solved by gradient based methods.

3 Global Optimization Technique

In this section, we introduce a mixed integer linear program formulation for the arrangement problem of three-dimensional convex polyhedral objects. Non-convex objects can be modeled by a set of convex objects whose relative position is fixed and non-polygonal surface models are approximated by suitable meshes with planar faces. A detailed treatment of our procedure is given in [SL00]. The following well known theorem forms the basis of our problem formulations:

**Theorem 1 (Linear Separation).** Convex objects can be linearly separated if and only if they do not intersect.

The discretization introduced in the next section shows that we can restrict the possible separations to a finite number in the case of polyhedral objects. This technique is also discussed in [ST93] and known as configuration space approach in robotics [AS94]. Another application of this technique can be found in architectural layout optimization [FW95].

3.1 Discretization of the Configuration Space

To introduce the basic ideas, we consider two-dimensional convex polygonal objects with a fixed orientation, for the moment. To characterize the domain of non-intersecting positions of two objects $P_i, P_j$ we use the Minkowski sum
$P_i \oplus P_j = \{p+q \mid p \in P_i, q \in P_j\}$, the pointwise sum of objects (see [Gho90] for a detailed discussion of Minkowski operations). The Minkowski sum of convex polyhedral objects can easily be calculated, for example as the convex hull of the pointwise sum of vertices, and the following theorem holds:

**Theorem 2.** The Minkowski sum of (convex) polygonal objects is a (convex) polygonal object.

Figure 1 shows an example. The dotted line is the boundary of the Minkowski sum. The lower left corner of the triangle $P_j$ is marked with a dot $r_j$. We choose such a unique reference point for each polyhedron. The correspondence between Minkowski sums and intersection analysis is the following: $P_i$ and $P_j$ intersect iff the normalized reference point $(r_j - r_i)$ is inside $P_i \oplus P_j$ mirrored at the origin. The region outside the Minkowski sum (light gray in the figure) is now partitioned into a finite number of convex subregions (marked $s_{ij\ast}$). The number of regions corresponds to the number of faces of the Minkowski sum.

We introduce 0/1 decision variables $s_{ij\ast}$ for each pair $(i < j)$ of objects and each region derived from their Minkowski sum. We interpret $s_{ijk} = 1$ to mean that the reference point $r_j$ of $P_j$ must lie in region $k$. With $M$ a sufficiently large constant we can express this meaning of the $s_{ij\ast}$ by linear inequalities of the form $h_{ijkl}(r_j) \leq 0 + M(1 - s_{ijk})$ ($r_j$ must lie in some halfspaces). Note that the inequality is only “activated” if $s_{ij\ast}$ equals 1. We require that for each fixed pair of objects $i, j$ exactly one of the decision variables must be set to 1, which is formalized as $\sum_k s_{ijk} = 1$. This leads to the basic mixed integer linear program formulation for the arrangement.
problem of convex polyhedral objects:

\[
\begin{align*}
\min f(r_j) & \quad (1) \\
\text{subject to} & \\
\bar{h}_{ijkl}(r_j) & \leq 0 + M(1 - s_{ijk}) \quad (2) \\
\sum_k s_{ijk} & = 1 \quad (3) \\
r_j & \geq 0, \ s_{ijk} \in \{0, 1\} \quad (4)
\end{align*}
\]

This basic formulation can be extended to a fixed number of orientations for each object by introducing additional 0/1 variables \( o_i \). If \( P_i \) and \( P_j \) belong to the cover of the same non-convex object, we can fix their relative position by equalities \( r_i = r_j + \text{offset} \) and drop the \( s_{ij} \) variables for this pair.

The mixed integer program formulation can be solved by a branch-and-bound procedure. The integer variables are handled and set by the branching procedure and the resulting linear program is then solved by standard methods. This enables a complete enumeration of normalized arrangements (in a normalized arrangement objects are shifted down and to the left as far as possible and touch each other). The bounding procedure is used to reject certain regions of the search space. The calculated bounds assure that one dose not cut off the optimal solution. For the details of the branching strategy and bounding procedure we refer the reader to [SL00].

### 3.2 E-Module-Boxes with Wiring Area Estimation

We present some results for a problem class identified together with our industrial cooperation partner. In this application a set of electronic modules (like ABS, electronic engine control etc.) has to be fitted inside a given waterproof box, called the E-Module-Box. The problem instances include cables interconnecting modules and connecting them to external devices. The diameters of cables are not neglectable and we decided to integrate a wiring area estimation into the global arrangement procedure.

The wiring area estimation is based on a rough approximation of the course of cables. For each cable \( w \) with diameter \( d_w \) we calculate a set \( T_w \) of routing variants which represent Steiner trees in a special routing graph. The routing graph is based on routing cells \( C \) whose size can vary. We use 0/1 decision variables \( x_{wi} \) to select a routing variant \( i \) for wire \( w \). Actually, we use only the linear relaxation of the variables requiring \( x_{wi} \geq 0 \) and \( \sum_i x_{wi} = 1 \). This may result in fractional values for the routing variants which are interpreted as a kind of preference value for each variant.

Figure 2 shows the basic idea of adjusting the size of the routing cells. Consider two adjacent cells \( C_a, C_b \) which have a partially overlapping common face with size \( c_x, c_y \). The shorter side \( m_{ab} \) must be at least as long as the maximum diameter of a wire routed across the face \( (5),(7) \). If one of the
variables $c_x, c_y$ is equal to $\min_{ab}$ we require the other side to have a length of at least the sum of wire diameters routed across ((6); wires are stacked on top of each other). This can be expressed by the following linear constraints:

$$
\min_{ab} \geq \sum_i x_{wi}d_w
$$

$$
\min_{ab} + \sum_{w,i} x_{wi}d_w \leq c_x + c_y
$$

$$
c_x \geq \min_{ab}, c_y \geq \min_{ab}
$$

Figure 3 shows results obtained for a problem instance supplied by our cooperation partner. The picture to the left shows the layout actually used in series production. It fills a box of approx. $31 \times 22 \times 19 \text{cm}^3$, amounting to $12,958 \text{cm}^3$. The smallest cube enclosing the space optimal arrangement of modules without wiring has a volume of $7,417 \text{cm}^3$ and of $7,874 \text{cm}^3$ including the wiring area estimation. The solution saves approx. $10 \text{cm}$ in length by arranging four (instead of three) modules parallel to the sidewall of the box. The algorithm uncovered this layout by a rotation of the modules which allows the wiring to pass over the additional one. The layout calculated by the algorithm could be realized with a physical prototype from our cooperation partner.

4 Local Optimization Technique

Our local optimization procedure starts out from an initial arrangement of objects and performs a series of compaction steps which converge to a local optimal solution. A linear program is solved in each step which enforces
the linear separation of objects while optimizing translational and rotational parameters. Figure 4 shows four (out of ten) optimization steps for a simple arrangement. This approach was introduced by Victor Milenkovic for the two-dimensional case [Mil97]. We have extended the method to three-dimensional layouts using a suitable linearization of orthonormal rotation matrices.

4.1 Properties of Three-Dimensional Rotations

Three-dimensional rotations correspond to $3 \times 3$ matrices $R \in SO(3)$ defined by the properties $R^{-1} = R^t$ ($R^t$ transposed matrix with $R_{ij} = (R^t)_{ji}$) and $\det R = +1$. In contrast to 2D rotations they are not commutative. While 2D rotations are defined by exactly one parameter (the rotation angle), three degrees of freedom are required to characterize a 3D rotation. One may choose an axis vector with length equal to the rotation angle or a representation by Euler angles. Quaternion representations with four parameters (but not independent) are also quite common. All these representations have the property, that we cannot state the interdependence of parameters or the resulting transformation in terms of linear (in-)equalities. Consider now matrices $\tilde{R}$ of the form

$$
\begin{pmatrix}
1 & \gamma & -\beta \\
-\gamma & 1 & \alpha \\
\beta & -\alpha & 1
\end{pmatrix}.
$$

These matrices have been studied in physics and are called **infinitesimal rotations** for small values of $\alpha, \beta$ and $\gamma$. The resulting coordinate system $(\tilde{R}e_1, \tilde{R}e_2, \tilde{R}e_3)$ is neither normal nor orthogonal, in general. We can prove the following theorem, which enables us to use these distorted transformations for object separation and optimization of rotational parameters:

**Theorem 3.** Let $\tilde{R}$ be an infinitesimal rotation and $C$ a convex solid object. The image of $C$ under transformation $\tilde{R}$ represents a hull of $C$ under some orthonormal transformation $(T \circ R)$, where $R$ is a rotation and $T$ a translation. $R$ and $T$ can be effectively constructed from $\tilde{R}$. 

**Fig. 4.** Four steps of a local optimization process
Note that the values for $\alpha, \beta, \gamma$ are not restricted in any way and are independent of each other. If $v$ is a point (think of vertices of objects) and $\vec{r} = (\alpha, \beta, \gamma)$, then $\vec{R}v = v + \vec{r} \times v$ with $\times$ being the standard 3D vector product.

We can now derive the separation inequalities. If we have a fixed normal vector $n_{ij}$ for each separating plane between convex objects $P_i, P_j$ with vertices $v_{i*}, v_{j*}$ (be reminded that we always have an initial arrangement without intersections) then the following linear inequalities in rotation variables $\vec{r}_*,$ translation variables $t_*$ and right hand side scalar variables $b_{ij}$ form the linear separation constraints for three-dimensional rotational compaction:

$$n_{ij}(v_{i*} + \vec{r}_i \times v_{i*} + t_i) \leq b_{ij} \leq n_{ij}(v_{j*} + \vec{r}_j \times v_{j*} + t_j) \quad (8)$$

After each compaction step we reconstruct an orthonormal transformation $(T_i \circ R_i)$ from $\vec{r}_i$ and $t_i$ for each object which leads to the new arrangement. Actually we do not use fixed normal vectors $n_{ij}$ but we use a small perturbation term $\Delta n_{ij} \perp n_{ij}$. In this way the separating plane may slightly tilt, as well.

### 4.2 Arrangement of Devices in the Trunk

A second test application supplied by our cooperation partner was the arrangement of devices in the trunk of a passenger car. In this case some restrictions due to accessibility requirements had to be obeyed. Figure 5 shows the arrangement realized in series production. We used a combination of our global and local optimization procedures to obtain substantially different and more compact layouts. Figure 6 visualizes one of our solutions compared to the original layout shown with outlined shapes.

![Fig. 5. Arrangement of devices in the trunk of a passenger car](image1)

![Fig. 6. Alternative compacted layout compared to original arrangement](image2)
5 Perspectives

The ultimate goal from a perspective of optimization would be an exact global optimization procedure for the non-linear mixed-integer formulation of polyhedral arrangement problems. Remember that the reason behind our two-phase approach is the computational infeasibility of such an approach. We want to sketch one step towards this goal, by combining the Minkowski sum approach of the global optimization procedure and our linearized local optimization model. Suppose we have two objects $P_i, P_j$ and consider their Minkowski sum while changing their relative orientation. The topology of the Minkowski sum, which is a polyhedral object, only changes at certain rotation angles. This property can be used to define an equivalence relation with a finite number of equivalence classes on the domain of possible orientations. Decision variables $s_{ij}$ are now introduced to select one equivalence class and a linearized rotation approach is used to describe the separations within this class. The convolution operation introduced in [BGRR97] can be used to calculate the equivalence classes. We are currently investigating how to implement this approach to be useful in practice.

Acknowledgments. The authors would like to thank our cooperation partners at DaimlerChrysler AG, Stuttgart, for helpful discussions and the supplied test cases. This work was funded by the Federal Ministry of Education, Science, Research and Technology (BMBF) under grant no. 03LE7B014.

References

Three-Dimensional Arrangement Problems


Simulation of Test-drives of Automobiles at Driving Limit

H. J. Pesch and M. Gerdts

Lehrstuhl für Ingenieurmathematik (FAN), Universität Bayreuth, 95440 Bayreuth, Germany, URL: http://www.uni-bayreuth.de/departments/ingenieurmathematik

Abstract. This project has been carried out in cooperation with the Volkswagen AG at Wolfsburg/Germany. The scope of the project is the computer-aided simulation of test-drives of an automobile at driving limit. Today the development of a new automotive model is a very expensive process. A large amount of the development costs is consumed by the building of prototypes. Therefore comparatively cheap computer based simulation methods are used to win insights to the future dynamical behavior of the automobile as soon as possible during the development process. The use of powerful software tools allows the complete modelling of automobiles by differential-algebraic equation systems of possibly higher index, e.g., as mechanical multi-body systems. An optimal test-driver is modeled by formulating an optimal control problem subject to the equations of motion of the automobile and additional constraints given by the test-course. The mathematical challenge is to solve large scale optimal control problems with differential-algebraic equation systems of higher index. The benefit for the industrial partner may be described by the slogan “test-drives without prototypes”.

1 Problem Formulation

The task of the project is the simulation of test-drives of an automobile at driving limit. In connection with this task it is interesting to study the behavior of a complete car model to detect potential malfunctions in the handling of the automobile. Because such malfunctions can have severe effects on the security of the automobile it is necessary to study the dynamical behavior in extreme situations, e.g. at high speed in conjunction with demanding manoeuvres. In this project the following items are considered:

- Modelling
  - automobile: mechanical multi-body system;
  - test-courses and manoeuvres;
  - driver: formulation of an optimal control problem;
- Development and implementation of suitable numerical methods for the solution of the resulting large scale optimal control problems with differential-algebraic equation systems of higher index;
- Computation of selected test-drives.
2 Mathematical Modelling

2.1 Automobile

The automobile is modeled as a mechanical multi-body system (MBS). Figure 1 shows some elementary components of a realistic car, which have to be considered in more or less detail in a realistic car model. In a MBS each component is considered as a rigid body. Because of the large number of components in a car it is nearly impossible to describe the model mathematically "by hand". Today powerful software packages, e.g. SIMPACK [17], for the design and simulation of MBS exist. These programs allow an automatic generation of the equations of motion and exportation into programming languages as FORTRAN 77.

The equations of motion are given by the differential-algebraic equation system (DAE system) in descriptor form

\[ \dot{p}(t) = v(t) \]
\[ M(p(t)) \dot{v}(t) = f(p(t), v(t)) - (c_p(p(t)))^T \lambda(t) \]
\[ 0 = c(p(t)) \]  

with the generalized position coordinates \( p \), the generalized velocities \( v \), the constrained forces \( \lambda \) and the positive definite mass matrix \( M \), see [11]. DAE systems can be viewed as singular ordinary differential equation systems because in this formulation the derivative of \( \lambda \) is not given explicitly but implicitly by the algebraic constraint \( c(p(t)) = 0 \). Twofold differentiation of the algebraic constraint w.r.t. time \( t \) leads to hidden constraints \( 0 = c_p(p)v \) and \( 0 = \frac{\partial (c_p(p)v)}{\partial p} v + c_p(p)(M(p))^{-1}(f(p, v) - (c_p(p))^T \lambda) \) that have to be fulfilled at each time \( t \). If the matrix \( c_p(p)(M(p))^{-1}(c_p(p))^T \) is regular, the second hidden constraint can be solved for \( \lambda \) and one more differentiation w.r.t. \( t \) yields a differential equation for \( \lambda \). So, three differentiations are necessary to receive a differential equation for \( \lambda \). Consequently the DAE system is said to be index three. One idea to solve the DAE system numerically is to solve the so called underlying ODE system given by the differential equations for \( \dot{p}, \dot{v}, \dot{\lambda} \) instead of (1). However, due to round-off errors and error propagation the algebraic constraint as well as the hidden constraints
may not be satisfied during the integration. This is the so called "drift-off effect". Projection methods overcome this effect by projecting the numerical solution on the constraints, v. [4]. Other methods make use of (1) directly or modifications of it, v. [16], [8] and [7]. In praxis often the index-reduced form of (1) with \(0 = c(p)\) replaced by \(0 = c_p(p)v\) is used. For short integration intervals the drift-off effect is rather moderate. Apart from problems concerning the integration of DAE systems another difficulty occurs. In contrast to ODE systems not any initial values for \(p, v\) and \(\lambda\) are admissible because the algebraic constraint and the hidden constraints have to be fulfilled, i.e. initial values have to be consistent with the DAE system.

2.2 Track

The mid-line of the roadway is described by a unit speed curve which is piecewise defined by clothoids, i.e. curves with linear curvature, see [13]. The boundaries of the test-course are given by the width of the track depending on the arc length. Alternatively spline interpolation or piecewise defined polynomials can be used to formulate the boundaries of the test-course, compare [12]. In addition local properties of the roadway such as bumps or slick sections can be considered, v. [18].

2.3 Driver

According to [12] an optimal test-driver is modeled by the optimal control problem

\[
\begin{align*}
\text{Minimize} & \quad J[x, u, p] = \Phi(x(t_f), t_f, p), \quad t_0 < t_f, \\
\text{subject to} & \quad F(x(t), \dot{x}(t), u(t), p) = 0, \quad t \in [t_0, t_f], \\
& \quad g(x(t), u(t), p) \leq 0, \quad t \in [t_0, t_f], \\
& \quad \psi(x(t_0), x(t_f), p) = 0, \quad u(t) \in U, \\
& \quad p \in P.
\end{align*}
\]

(\textit{OCP})

In this formulation \(x\) denotes the state of the dynamical system, \(u\) denotes the control variables, e.g. the motion of the steering wheel and the acceleration of the automobile, and \(p\) are parameters, e.g. spring and damper characteristics of the wheel suspension, to be chosen in a way that the objective \(\Phi\) is minimized. The equations of motion of the automobile are written as an implicit DAE system \(F(x, \dot{x}, u, p) = 0\). The optimal solution has to satisfy path constraints \(g(x, u, p) \leq 0\), e.g. boundaries of the test-course, and boundary conditions \(\psi(x(t_0), x(t_f), p) = 0\), e.g. initial or final position on the test-course. The final time \(t_f\) may be fixed or free.
3 Mathematical Methods

3.1 Discretization of the Optimal Control Problem

The optimal control problem (OCP) is discretized by introducing a suitable chosen grid \( t_0 = t_1 < t_2 < \cdots < t_{N+1} = t_f \) on which the control \( u \) is approximated by \( u_{app} \). For example the continuous and piecewise linear function

\[
u_{app}(t) = u_i + \frac{t - t_i}{t_{i+1} - t_i}(u_{i+1} - u_i), \quad t_i \leq t \leq t_{i+1}, \quad i = 1, \ldots, N,
\]

with \( u_i \approx u(t_i) \) is very popular. The state \( x \) is approximated by the solution \( x_{app} \) of the DAE system

\[
F(x_{app}(t), \dot{x}_{app}(t), u_{app}(t), p) = 0, \quad t_0 \leq t \leq t_f,
\]

by modifications of standard integrators like DASSL, [14], or ODASSL, [7]. For simplicity we restrict ourselves to this single shooting approach, though a generalization by a multiple shooting technique is possible and has been implemented. Introducing these approximations into the infinite dimensional optimal control problem (OCP) and evaluating the constraints on the grid lead to a finite dimensional nonlinear optimization problem

\[
\begin{align*}
\text{Minimize} & \quad \Phi(x_{app}(t_{N+1}), t_{N+1}, p), \\
\text{subject to} & \quad g(x_{app}(t_i), u_{app}(t_i), p) \leq 0, \quad i = 1, \ldots, N + 1, \\
& \quad \psi(x_{app}(t_1), x_{app}(t_{N+1}), p) = 0, \\
& \quad u_{app}(t_i) \in U, \quad i = 1, \ldots, N + 1, \\
& \quad p \in P,
\end{align*}
\]

which is solved numerically employing a SQP method, [10]. Gradients and Jacobians needed by the SQP method are computed via the sensitivity DAE corresponding to (2) with respect to the variables of (NLP), compare [5]. For this purpose DASSL and ODASSL have been augmented by appropriate routines. Note that consistent initial values have to be provided to solve (2) in each iteration of the SQP method. Especially in the presence of free initial values to be chosen in an optimal manner the situation becomes more complicated. In this case a projection method in combination with a sensitivity analysis of the projection function in the sense of [6] is employed, see [2], [3] and [9]. The depicted method augmented by routines for the calculation of consistent initial values for the DAE system has been implemented in the software package SODAS which allows a reliable numerical treatment of (OCP). For very long and complex test-courses the solution of (OCP) may be very difficult due to stability and accuracy problems and the large-scale dimensioning of (NLP). In this case the alternative approach of driving with reduced range of vision (RRV) is applied. Instead of formulating an optimal control problem for the test-drive along the complete test-course only
a short section with local information about the roadway is considered, for example a range of 100 [m] or 10 [s]. The resulting optimal control problem for this fraction is less dimensional and easier to solve. The local solutions are coupled by appropriate transient conditions such that the whole test-course is completed.

4 Simulation Results

In this section some selected results of simulated test-drives are presented.

4.1 Example 1: Mechanical Multi-Body System of a VW Iltis

Figures 2 and 3 show two typical manoeuvres of standard type: The double-lane-exchange manoeuvre and the slalom manoeuvre. The double-lane-exchange manoeuvre depicted in Figure 2 serves to investigate the behavior of the automobile when an obstacle necessitates a jink. Its measurements are specified in [20]. On the basis of the slalom manoeuvre shown in Figure 3 predictions for the stability of the chassis tuning are possible. The considered model is a 25 degree of freedom model of the VW Iltis explained in [15] in detail. Its equations of motion are of type (1) and are generated by SIMPACK [17] using the automatic exportation facility into FORTRAN 77. With a suitable approximation of the test-course by piecewise defined polynomials the corresponding optimal control problems (OCP) for the two manoeuvres are formulated and solved using SODAS.

Fig. 2. Double-lane-change manoeuvre ("elchtest") of the ILTIS model
The driver’s control potentiality is restricted to the steering mechanism of the automobile because the considered manoeuvres are driven with constant velocity typically. The results are obtained using equidistant grids with 41 respectively 101 grid points. The final position of the automobile on the roadway is restricted in such a way that it can be ensured that the automobile will stay on the roadway for some time after the simulation ends. If no restriction to the final position is given, the optimal control problem often ends in a feasible solution, where its extrapolation beyond the final time violates the boundaries of the roadway for any admissible control. The objective was the minimization of the steering effort in the case of the double-lane-change manoeuvre and the minimization of a linear combination of the steering effort and the final time for the simulation of the slalom manoeuvre. The final time was free in each case. The first result was obtained for the index reduced model, which has index 2. The second simulation was computed by employing a stabilization technique introduced in [7].

![Fig. 3. Slalom manoeuvre of the ILTIS model with a velocity of 18 [m/s] (solid) and 15 [m/s] (dashed)]
4.2 Example 2: Driving with Reduced Range of Vision

The computations are based on a slightly modified model of a car described in [19] primarily. The resulting model is given as an index-1 DAE system of dimension 43 with 4 algebraic components. The test-course of length 1200 [m] is modeled by a unit speed curve with a piecewise defined curvature, which is not continuous in contrast to the recommendation of the German road construction authority. Therefore the handling of the car at these junction points becomes more difficult. In addition the width of the roadway depends on the arc length and has a stricture at the bottom right part of the roadway, compare Figure 4. The arrow in Figure 4 marks the initial orientation of the automobile’s motion. The figure shows the resulting path of the center of mass of the automobile for a simulation with the RRV approach. The objective in each local optimal control problem was the minimization of the steering effort on the one hand and the maximization of the velocity on the other hand, thus a linear combination of these objectives. The driver modeled by (OCP) is able to control the steering mechanism and the acceleration of the automobile. The final position of the car on the roadway again is restricted by the postulation that the car should be able to proceed its drive without leaving the roadway. Please note that in general boundary conditions of the type given in (OCP) can not be guaranteed by the RRV approach because of the local character.

Fig. 4. Driving with reduced range of vision: Resulting path for index 1 model
In Figure 5 time plots for the steering velocity, which is one of the control variables of the problems, as well as for the resulting lateral acceleration and the roll angle of the simulation are depicted. The computation was carried out with a range of vision of 7 seconds and a shifting range of 1.05 seconds marked by the vertical lines in Figure 5. This means that 1.05 seconds of each trajectory of the local optimal control problems are accepted. The state at 1.05 seconds of the local optimal control problems serves as an initial state of the successive optimal control problem. The control is allowed to jump at the junction points, compare Figure 5. The local optimal control problems are formulated with a fixed final time of 7 seconds. Except for the final local optimal control problem, no boundary conditions at final time are requested. Though a potentially large number of optimal control problems have to be solved, another advantage of the method is that the solution of the previous fraction may serve as a very good initial estimate for the subsequent problem and therefore the computational effort is not immoderate.

**5 Outlook**

In the remaining time of the project we focus attention on the visualization of the obtained results and the validation and extension of the developed software as well as the computation of further simulations of virtual test-
drives. For the future it is intended to perform real-time optimization for optimal control problems with DAE systems of higher index. Therefore the ideas in [1] should be adapted respectively extended to this problem class. In this context the simulation of test-drives of automobiles is an ideal area of application.

References


An Optimal Control Approach
To Real-Time Vehicle Guidance

Roland Bulirsch¹, Martin Vögel¹, Oskar von Stryk¹, Cornelius Chucholowski², and Thieß-Magnus Wolter²

¹ Technische Universität München, Zentrum Mathematik/M2, 80290 München, Germany, URL: http://www-m2.ma.tum.de
² Tesis DYNAware GmbH, Implerstraße 26, 81371 München, Germany, URL: http://www.tesis.de

Abstract. A newly developed two-level driver model is presented. On the anticipation level, optimal control problems for a reduced vehicle dynamics model are solved repeatedly on a moving prediction horizon to yield near optimal setpoint trajectories for the full model. On the stabilization level, a nonlinear position controller is developed to accurately track the setpoint trajectories with a full motor vehicle dynamics model in real-time. The formulation of the optimal control problems on the anticipation level is based on a nonlinear single track model which is extended by a complex tire model and further nonlinear model details such as to match the main properties of the full vehicle dynamics model. The optimal control problems are solved efficiently by a recently developed sparse direct collocation method. Numerical results for various vehicle maneuvers are presented, including a time-optimal double lane change at high speed.

1 Introduction

Driving comfort and safety of modern passenger cars can highly be improved by electronic vehicle control units (ECUs) actively interfering in the vehicle dynamics. Popular examples include anti-lock braking systems (ABS) and electronic stability programs (ESP). However, through the use of ECUs the vehicle design and the dynamical driving properties become even more complex. Therefore, major car manufacturers as well as automotive suppliers make use of suitable software for virtual prototyping in order to cut the product development time and cost, and to improve the design quality.

Specifically, Hardware-in-the-Loop (HIL) experiments provide efficient, cost effective, reproducible, and save tests for ECUs. Here, a test bench is used to link the ECU to the numerical real-time simulation of the full motor vehicle dynamics (Fig. 1). Furthermore, to rate the handling and driveability properties of a virtual prototype, the performance of the closed loop of driver, vehicle, and environment must be investigated in real-time [27]. Thus, handling characteristics of the physical prototype such as body roll, ride quality including vibration and bumps, vehicle safety, and performance parameters can be predicted.
Fig. 1. Hardware-in-the-Loop setup for a braking system with ABS control unit linked to the real-time simulation of full motor vehicle dynamics

To realistically simulate the full motor vehicle dynamics in real-time tailored models of the vehicle dynamics and the road properties are needed (Sect. 2). A new driver model consisting of two levels is outlined in Sect. 3. On the anticipation level, optimal control problems on a moving prediction horizon (Sect. 3.3) are solved for a newly developed extended single-track vehicle model (Sect. 3.2). Thus, optimal open-loop controls and trajectories are obtained, e.g., specifying the location and the velocity of the vehicle's center of gravity along the road. The latter are used as setpoint trajectories for the nonlinear real-time position control which serves to compensate any disturbances on the stabilization level of the driver model (Sect. 3.5). For the efficient numerical solution of the optimal control problems on the anticipation level, a recently developed version of a sparse direct collocation method has been employed (Sect. 3.4).

The numerical results for various virtual test drives demonstrate the efficiency of the approach (Sect. 4).

2 Modelling and Simulation of Full Vehicle Dynamics

To realistically “test drive” entire vehicles in the computer, computer models of the whole car, including the suspensions, the powertrain, the engine, the steering mechanism, and ECUs are required along with models for the road geometries and conditions.

2.1 Numerical Simulation of Full Motor Vehicle Dynamics in Real-Time with veDYNA

A detailed and comprehensive vehicle model is needed to allow for the nonlinear kinematics of wheel and axle, and to describe the drive train, the steering
Fig. 2. Submodels of the full motor vehicle dynamics model of *veDyna*

mechanism and the tire dynamics. Our vehicle model consists of a suitable multibody system with kinematical connections and force elements which is supplemented by a sophisticated tire model. General purpose methods for modelling multibody systems use the descriptor form of the equations of motion leading to a large-scale system of differential-algebraic equations (DAEs) of index 3 [20]. For DAE systems particular numerical techniques must be applied to prevent a “drift-off” from the algebraic constraints [21]. Here, we make use of an optimally tailored model description which yields a system of ordinary differential equations (ODEs) which is suitable for real-time simulation.

The vehicle model of *veDyna* consists of a system of nine rigid bodies comprising the vehicle body, the axle suspensions and the wheels. Further submodels are employed to depict the characteristics of the drive train, the steering mechanism, and the tires (Fig. 2). Suitable minimum coordinates and generalized velocities are used to describe the spatial state of the vehicle and its components [19]. The equations of motion are derived from Jourdain’s Principle yielding

\[ M_{BV}(y_{BV}) \ddot{y}_{BV} = Q_{BV}(y_{BV}, z_{BV}, y_{ST}, z_{ST}, y_{DT}, z_{DT}) \]  

(1)
An Optimal Control Approach To Real-Time Vehicle Guidance

\[ \dot{y}_{BV} = K_{BV}^{-1}(y_{BV}) \ z_{BV} \] (2)

\[ M_{DT} \ \dot{z}_{DT} = Q_{DT}(y_{DT}, z_{DT}) \] (3)

\[ \dot{y}_{DT} = V_{DT} \ z_{DT} \] (4)

\[ M_{ST}(y_{ST}, y_{BV}) \ \dot{z}_{ST} = Q_{ST}(y_{ST}, z_{ST}) \] (5)

\[ \dot{y}_{ST} = V_{ST} \ z_{ST} \] (6)

\[ D \ \dot{y}_{T} = F_{stat} - C \ y_{T}. \] (7)

Thus, the vehicle dynamics is fully characterized by the system of 24 first-order ODEs comprising the vehicle body and the axles, (1) and (2). Eight ODEs (7) describe the lateral and longitudinal deviations of the tires by means of spring and damper elements. The vertical deformations of the tires are covered by (1). The dynamic model of the drive train consists of 19 ODEs, (3) and (4), including four equations governing the angular wheel speeds. Five additional ODEs account for the dynamics of the steering system (5) and (6). Couplings between the separate systems occur via the generalized forces and torques \( Q_{BV} \). Wind forces and moments result in additional forces applied to the multibody system of the vehicle [17,19].

The tire forces have a significant impact on the dynamical behavior of a vehicle. The semi-empirical tire model that is used here describes the behavior of a real tire accurately [11,19]. About 80 parameters which can be measured or estimated enter the model for each tire in \textit{veDYNA}. The model covers different driving situations, including effects at the driving limits such as sliding and spinning. The actual tire model is selected online depending on the respective road and weather conditions [1].

Due to the stiffness of the system (1)–(7) its numerical integration is carried out recursively with a semi-implicit one-step Euler scheme using a constant step size [19]. In particular, the integration method makes efficient use of the special block structure of the ODEs which yields

\[ z_{ST}^{k+1} = z_{ST}^k + h \left( M_{ST} - h^2 \frac{\partial Q_{ST}}{\partial z_{ST}} - \frac{\partial Q_{ST}}{\partial y_{ST}} V_{ST} \right)^{-1} Q_{ST}^k \] (8)

\[ y_{ST}^{k+1} = y_{ST}^k + h V_{ST} z_{ST}^{k+1} \] (9)

\[ z_{DT}^{k+1} = z_{DT}^k + h \left( M_{DT} - h^2 \frac{\partial Q_{DT}}{\partial z_{DT}} - \frac{\partial Q_{DT}}{\partial y_{DT}} V_{DT} \right)^{-1} Q_{DT}^k \] (10)

\[ y_{DT}^{k+1} = y_{DT}^k + h V_{DT} z_{DT}^{k+1} \] (11)

\[ z_{BV}^{k+1} = z_{BV}^k + h \left( M_{BV} - h^2 \frac{\partial Q_{BV}}{\partial z_{BV}} - \frac{\partial Q_{BV}}{\partial y_{BV}} K_{BV}^{-1} \right)^{-1} Q_{BV}^k \] (12)

\[ y_{BV}^{k+1} = y_{BV}^k + h K_{BV}^{-1} z_{BV}^{k+1}. \] (13)

It turns out that a fast and stable solution is possible in real-time on modern PC hardware.
The development of intricate vehicle control devices requires the performance of the virtual car to match the actual vehicle behavior accurately. For calibrating the vehicle model of veDYNA, a parameter estimation tool has been developed which relies on observations obtained from physical test drives. The associated nonlinear least-squares problems are solved efficiently by means of mathematical optimization algorithms. Significant speed-ups in the computational time are achieved when employing a low-cost parallel computing platform, such as a heterogeneous cluster of PCs, which is well suited for the needs of the automotive industries and suppliers applying vehicle dynamics simulations [4].

2.2 Parameterized Road Model

To exercise the dynamic vehicle models under various road conditions (test tracks, public streets) in the computer, three-dimensional road geometries such as banked curves, hills, and bumpy roads, as well as variations in the road parameters due to weather and road surface conditions must be considered. These parameters vary widely on real courses, but can be controlled easily for virtual test drives if suitable models are employed.

For this purpose, we developed the tailored and parameterized road model veDYNA Advanced Road (Fig. 3) for the application in real-time vehicle dynamics simulations [22]. The course of the road is defined by means of the center line which is a curve in \( \mathbb{R}^3 \). Its vertical projection onto the horizontal plane is defined by a curve in \( \mathbb{R}^2 \) whose curvature is a piecewise linear function of the arc length as it is usual in road construction. The vertical coordinate of the center line can be selected arbitrarily. A variable transverse profile can

(a) The curvature of the road's center line is a piecewise linear function of the arc length.

(b) The variable transverse profile enables the modelling of bumps, dips, treads etc.

(c) Road sections with different surface properties (traction values, roughness)

Fig. 3. Components of the parameterized road model veDYNA Advanced Road
be added at each point of the center stripe for generating three-dimensional road properties like bumps and dips. Road surface properties such as traction coefficients and the grade of roughness or smoothness are further parameters of the transverse profile, thus, permitting a description of different road and weather conditions.

The road model is flexible and easy-to-use. A realistic road model can be obtained with quite a few parameters, though it is possible to increase the level of detail arbitrarily. Moreover, the road model is well suited for real-time simulation. We refer to [22] for details. The road model has been incorporated into a graphical user interface of the MATLAB/Simulink emulation of veDYNA and into the 3D-visualization socket for computed test drives. Thus, the comfortable, modular modelling and visualization of any test course of interest is made possible.

3 Optimal Control Approach to Driver’s Anticipation and Response

To enable virtual test drives the control actions of the vehicle’s driver, such as steering, braking, accelerating, gear shifting, and operating the clutch, must be simulated by a virtual driver. In this context, we do not want to investigate the specific biomechanical, neuromuscular or psychological behavior of a human driver. Rather, a mathematical driver is needed for investigating the objective handling properties of the virtual prototype vehicle. Therefore, a virtual driver can use information which a human driver usually does not have (and cannot use directly), e.g., the traction values at the tires or the exact side slip angle. However, we require that the driver model is able to guide the virtual car along a test track at the dynamical driving limits in a way which is close to the performance of experienced human test drivers. An optimal control approach is motivated by “the basic assumption ... that the well-motivated, well-trained human operator behaves in a near optimal manner subject to his inherent limitations and constraints, and his control task” [12].

The two-level model of Donges [6] has been a fundamental contribution to modelling the driver’s response. The control tasks in vehicle guidance are separated into tasks on the guidance and the stabilization level. Nowadays, ECUs are typically considered to be control systems on the stabilization level to compensate for disturbances and enabling the driver to manage the vehicle’s dynamic behavior even in critical situations [7]. By this definition, control systems on the guidance level are systems that assist the driver in his/her steering task by looking ahead of the vehicle (anticipation).

3.1 Two-Level Driver Model

We have developed a two-level driver model for the use within the vehicle dynamics simulation program veDYNA (Sect. 2.1). On the stabilization level
of our driver model, a position control algorithm precisely guides a virtual vehicle, described by the full motor vehicle dynamics simulation of Sect. 2, along setpoint trajectories for the location and the velocity of the vehicle’s center of gravity (Sect. 3.5). On the *anticipation level*, the near optimal setpoint trajectories are computed by repeatedly solving optimal control problems for an extended single-track vehicle model by means of numerical optimization (Sects. 3.2, 3.3). The result is a virtual driver who is able to guide the virtual car on a virtual road at high speeds as well as during extreme maneuvers where skidding and sliding effects take place (cf. Sect. 4 and [5,23]).

3.2 Extended, Nonlinear Single-Track Vehicle Dynamics Model

For the numerical computation of optimal setpoint trajectories for the position and the velocity of the vehicle’s center of gravity along the road, a reduced vehicle dynamics model is required. The model must be reduced sufficiently such as to enable the online numerical solution of optimal control problems, but it must also be detailed enough to represent the main properties of the full vehicle.

For this purpose, we have developed a significantly extended, nonlinear single-track model (Fig.4). The single-track model due to [18] is based on the assumption that the center of gravity of the vehicle has zero height above a planar road. Thus, the left and right tire loads are equal even in case of large lateral accelerations; the two tires at the front and the rear axle are treated as one single “wide” tire. Moreover, under the given assumptions the tire loads are also constant during acceleration and braking maneuvers, and changes in the roll or pitch angle cannot occur. Therefore, it is not necessary

![Fig. 4. Single-track vehicle model](image-url)
An Optimal Control Approach To Real-Time Vehicle Guidance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>sideslip angle [rad]</td>
</tr>
<tr>
<td>$\psi$</td>
<td>yaw angle [rad]</td>
</tr>
<tr>
<td>$\omega_z$</td>
<td>yaw angular velocity [rad/s]</td>
</tr>
<tr>
<td>$v$</td>
<td>velocity of the center of gravity [m/s]</td>
</tr>
<tr>
<td>$X,Y$</td>
<td>position of center of gravity [m]</td>
</tr>
<tr>
<td>$\delta_v$</td>
<td>steering angle (at tire) [rad]</td>
</tr>
<tr>
<td>$\alpha_v, \alpha_h$</td>
<td>sideslip angle at front and rear wheel [rad]</td>
</tr>
<tr>
<td>$S_v, S_h$</td>
<td>front and rear lateral forces [N]</td>
</tr>
<tr>
<td>$H$</td>
<td>front and rear longitudinal forces [N]</td>
</tr>
<tr>
<td>$T$</td>
<td>drag force (wind) [N]</td>
</tr>
<tr>
<td>$l_v, l_h$</td>
<td>horizontal distance of front/rear axle to center of gravity [m]</td>
</tr>
<tr>
<td>$m$</td>
<td>mass of the vehicle [kg]</td>
</tr>
<tr>
<td>$\theta$</td>
<td>inertia due to z-axis [kg m$^2$]</td>
</tr>
</tbody>
</table>

Table 1. Variables of the single-track vehicle model

to model the axle kinematics, e.g. by spring and damper elements, and the single-track model behaves as if the tires were rigidly linked to the vehicle body.

The single-track vehicle dynamics depend on the acceleration, the braking and the lateral forces acting on the fictitious front and rear wheels. The acceleration and braking forces $V$ and $H$ at the front and rear wheels vitally determine the vehicle’s velocity $v$; the steering angle $\delta_v$ is used to guide the vehicle along any curve. Only six vehicle state variables (cf. Table 1) are contained in the equations of motion (14)–(19).

$$\dot{\beta}(t) = \omega_z(t) - \frac{1}{m v(t)} \left[ \left( H(t) - T(v(t)) \right) \sin(\beta(t)) 
+ S_v(\alpha_v(t)) \cos(\delta_v(t) + \beta(t)) + S_h(\alpha_h(t)) \cos(\beta(t)) \right]$$  \hspace{1cm} (14)

$$\dot{\psi}(t) = \omega_z(t)$$  \hspace{1cm} (15)

$$\dot{\omega}_z(t) = \frac{1}{\theta} \left[ S_v(\alpha_v(t)) l_v \cos(\delta_v(t)) - S_h(\alpha_h(t)) l_h \right]$$  \hspace{1cm} (16)

$$\dot{v}(t) = \frac{1}{m} \left[ H(t) \cos(\beta(t)) - T(v(t)) 
- S_v(\alpha_v(t)) \sin(\delta_v(t) + \beta(t)) + S_h(\alpha_h(t)) \sin(\beta(t)) \right]$$  \hspace{1cm} (17)

$$\dot{X}(t) = v(t) \cos(\psi(t) - \beta(t))$$  \hspace{1cm} (18)

$$\dot{Y}(t) = v(t) \sin(\psi(t) - \beta(t))$$  \hspace{1cm} (19)
Approximations for the side slip angles at the front and the rear wheel can be computed from

\[ \alpha_v = \beta - \omega_z \frac{l_v}{v} + \delta_v \]  

\[ \alpha_h = \beta + \omega_z \frac{l_v}{v}, \]  

provided that the yaw angular velocity \( \omega_z \) is "small". Although the lateral forces \( S_v \) and \( S_h \) mainly depend on the respective side slip angle, also the longitudinal force \( H \) has to be taken into account. In particular, the vector \((S_v, S_h, H)^T = F(\alpha_v, \alpha_h, H^{des})\) is a highly nonlinear function of the side slip angles and a desired longitudinal force \( H^{des} \). The latter can be interpreted as accelerating or braking pedal position. The function \( F \) serves to account for the interdependencies between lateral and longitudinal forces. Under standard driving conditions (i.e., \( \alpha_v, \alpha_h, H^{des} \) are small enough) \( S_v \) and \( S_h \) are almost independent of \( H^{des} \), and \( H \) is essentially given by \( H^{des} \). However, if the overall slip – which is calculated internally for each tire when evaluating \( F \) – becomes larger, the direction and the magnitude of the overall tire force determines the magnitude of the lateral and the longitudinal forces. The function \( F \) also switches between accelerating and braking, and considers different drive train configurations, such as front-, rear- and all-wheel drive.

When combining the single-track model with a suitable tire model, the main dynamical vehicle properties, including skidding effects, can be described properly [14,15,27]. As the most important enhancement we linked the nonlinear single track model with the sophisticated tire model used within the realistic full vehicle dynamics simulation package veDYNA (Sect. 2.1). This tire model, which is denoted by \( F \) here, covers the nonlinear interdependencies between the lateral and the longitudinal forces. Both veDYNA and our extended single-track model use the same subprograms and input data to evaluate the current road contact and to compute the respective tire forces. Thus, the forces \( H, S_v, \) and \( S_h \) in the single-track vehicle dynamics are computed by means of the tire model of the full motor vehicle dynamics program. Each tire model depends on about 40 parameters, i.e., half of the parameters of the full veDYNA tire model. Moreover, the large extension of the nonlinear lateral and longitudinal dynamics and the consideration of wind forces make the extended single-track model suitable for computing optimal setpoint trajectories which are consistent with the full vehicle dynamics simulation. The computational results for the extended single-track model are in good agreement with the full vehicle dynamics simulations, although only six ODEs are needed instead of 56 (cf. Sect. 4).

3.3 Optimal Control Problem on a Moving Horizon

The minimum time control problem for the extended nonlinear single-track model is stated as to

\[
\text{minimize } J[u] = t_f + \rho_0 x_b(t_f)
\]
where the $n_x = 8$ state variables and the $n_u = 2$ control variables are given by

$$x = (\beta, \psi, \omega_z, v, X, Y, \delta_v, x_8)^T, \quad u = (H^{des}, u_2)^T.$$  

(23)

The optimal control problem is subject to the equations of motion given by the Eqs. (14) – (19) and the additional equations

$$\dot{x}_7 = \delta_v = u_2, \quad \dot{x}_8 = \rho_1 u_1^2 + \rho_2 u_2^2.$$  

(24)

Here, $\rho_0, \rho_1, \rho_2 \geq 0$ denote suitable non-negative weights. Even if only a small weight $\rho_0$ for $x_8(t_f) = \int_{t_0}^{t_f} (\rho_1 u_1^2 + \rho_2 u_2^2) \, dt$ is considered in the objective (22), the Hamiltonian becomes regular and the solution is more smooth.

The maximum vehicle speed is constrained by

$$\dot{x}_4 = v \leq v_{\text{max}}.$$  

(25)

The non-negativity of two nonlinear state constraints

$$0 \leq g_i(x(t)), \quad i = 1, 2,$$  

(26)

ensures that the left and right road limits are obeyed during the maneuver.

By the Principle of Optimality, the optimal control problem over a long course may be decoupled into several problems over smaller sections without loss of optimality, provided that the optimal trajectory $x^*$ is known for some intermediate values. From the basic dynamical vehicle properties and the geometry of the course, intermediate values $x_{0,i}^*, x_{f,j}^*$ of the optimal trajectory can often be estimated quite well, e.g., by considering the maximum vehicle speed and the maximum lateral acceleration in a curve. Then, the optimal control problem on a prediction horizon $[t_0, t_f]$ with the unknown duration $t_f - t_0$ consists in optimally steering the vehicle from a given initial state to a given final state, i.e.,

$$x_i(t_0) = x_{0,i}, \quad x_j(t_f) = x_{f,j}, \quad i \in I_i, \quad j \in I_j, \quad I_i, I_j \subset \{1, \ldots, 8\}. \quad (27)$$

The optimal trajectory $x^*(t)$, $u^*(t)$, $t_0 \leq t \leq t_f$, must satisfy various necessary conditions derived from the Euler-Lagrange differential equations (EL-DEQs) and the Maximum Principle, e.g., [16,24]. Some of them can be used for an a-posteriori verification of the consistency of the solution computed by the method described in Sect. 3.4. Let $\lambda : [t_0, t_f] \rightarrow \mathbb{R}^{nx}$ denote the adjoint or costate variable, $\eta$ the multiplier function of the state constraints, and $\mathcal{H}(x, u, \lambda, \eta) = \sum_{i=1}^{nx} \lambda_i f_i(x, u) + \sum_j \eta_j g_j(x, u)$ the Hamiltonian. Then:

- The Hamiltonian $\mathcal{H}$ of an autonomous problem must be a (piecewise) constant function of time for the optimal trajectory.
- The final value of $x_8(t_f)$ is free. Thus, the corresponding adjoint variable $\lambda_8$ must satisfy

$$\lambda_8^*(t_f) = \frac{\partial J}{\partial x_8(t_f)} = \rho_0$$  

(28)

for the Mayer type objective (22).
The Hamiltonian is a nonlinear function of the control $u_2 = \delta_v$. Thus,

$$\frac{\partial H}{\partial u_2} = \lambda_7 + \lambda_8 2\rho_2 u_2 \equiv 0 \iff u_2 = -\frac{\lambda_7}{2\rho_2 \lambda_8} \quad (29)$$

must hold for all parts of the optimal trajectory lacking active constraints that affect $u_2$.

### 3.4 Sparse Direct Collocation for Numerical Optimal Control

Consider the general optimal control problem on the horizon $[t_0, t_f]$

$$J[u] = \varphi(x(t_f), t_f) + \int_{t_0}^{t_f} L(x(t), u(t), t) \, dt \rightarrow \text{min}! \quad (30)$$

subject to

$$\dot{x} = f(x(t), u(t), t), \quad (31)$$

$$0 \leq g_i(x(t), u(t), t), \quad i = 1, \ldots, n_g, \quad (32)$$

$$0 = r_j(x(t_0), t_0, x(t_f), t_f), \quad j = 1, \ldots, n_r, \quad (33)$$

where $x : [t_0, t_f] \rightarrow \mathbb{R}^{n_x}$ and $u : [t_0, t_f] \rightarrow \mathbb{R}^{n_u}$ denote the state and the control variables respectively.

Direct shooting and direct collocation methods both promise high flexibility and robustness when solving optimal control problems numerically with low or moderate accuracies [3,26]. However, in many practical applications the problem functions only have low, local differentiability properties, i.e., discontinuities in the first or second derivatives. Thus, obtaining a useful gradient approximation for shooting-type discretizations takes much more effort, since a numerical sensitivity analysis for initial value problems with switching points must be carried out. On the other hand, for a collocation-type discretization a careful, but much cheaper finite difference approximation may be sufficient, where no special treatment of discontinuities in first or second derivatives by switching functions is required. A further advantage of the direct collocation approach is the potentially faster computation as compared to direct shooting, because the ODE simulation (31) and the control optimization problems (30), (32) are solved simultaneously for collocation, but not iteratively as by shooting methods. To achieve an optimal speed-up for collocation, the NLP sparsity must fully be utilized. Otherwise the large size of the NLP will severely limit the efficiency.

In the sequel, we make use of a discretization for $x$ by piecewise cubic Hermite polynomials $\tilde{x}(t) = \sum_k \alpha_k \tilde{x}_k(t)$ and a discretization for $u$ by piecewise linear functions $\tilde{u}(t) = \sum_k \beta_k \tilde{u}_k(t)$ on a discretization grid $t_0 = t_1 < t_2 < \ldots < t_{n_t} = t_f$. Adherence to the equations of motion is prescribed at the grid points and their respective midpoints (collocation at Lobatto points), thus yielding a set of nonlinear NLP equality constraints $a(y) = 0$. Moreover, any inequality constraints on the control or state variables are to be satisfied at the grid points which results in set of nonlinear NLP
inequality constraints \( b(y) \geq 0 \). In both cases, \( y = (\alpha_1, \alpha_2, ..., \beta_1, \beta_2, ..., t_f)^T \) denotes the \( n_y \) parameters of the parameterization. The resulting nonlinearly constrained optimization problem basically reads as

\[
\text{NLP: } \min_y \Phi(y) \quad \text{subject to } a(y) = 0, \ b(y) \geq 0, \quad (34)
\]

where \( \Phi : \mathbb{R}^{n_y} \rightarrow \mathbb{R} \) denotes the parameterized cost index (30) with \( y \in \mathbb{R}^{n_y} \), and \( a : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_a} \), \( b : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_b} \) are the nonlinear NLP constraints.

The selected discretizations \( \tilde{u}, \tilde{x} \) must satisfy certain convergence properties. One requirement is that the discretized solution must approximate a solution of the EL-DEQs and the Maximum Principle if the grid becomes arbitrarily fine, i.e., for \( n_t \rightarrow \infty \) and \( \max\{t_{i+1} - t_i : i = 1, \ldots, n_t - 1\} \rightarrow 0 \) [24]. A great advantage of the direct collocation approach is that it provides reliable estimates \( \lambda \) of the adjoint variable trajectory along the discretization grid. These estimates are derived from the Lagrange multipliers of the NLP [24]. They enable a verification of the optimality conditions for the discretized solution although the EL-DEQs have not been solved explicitly.

Moreover, local optimality error estimates can be derived which enable efficient strategies for successively refining a first solution on a coarse grid [24,25]. Thus, a sequence of related NLPs must be solved whose dimensions increase with the number of grid points. The numerical solution of the NLPs can be done efficiently by sequential quadratic programming (SQP) methods. In each SQP iteration the current guess of the solution \( y^* \) is improved by solving a quadratic subproblem, derived from a quadratic approximation of the Lagrangian of the NLP, subject to the linearized constraints [2,9]. The NLPs which result from a direct collocation discretization have several specific properties:

- The NLPs are large-scale with very many variables and very many constraints.
- Most of the NLP constraints are active at the solution, e.g., the equality constraints from collocation. Thus, the number \( n_s \) of free NLP variables is much smaller than the total number of variables \( n_y \).
- The constraints’ Jacobians \( \nabla a(y), \nabla b(y) \) are sparse and structured. Only a small number of the elements will be nonzero, and the percentage decreases as the number of grid points increases.
- The NLP objective \( \Phi(y) \) only depends on a few, fixed number of variables regardless of the actual grid size, provided that the objective (30) is of Mayer type, i.e., \( L \equiv 0 \).

These features are fully utilized by means of the recently developed large-scale SQP method SNOPT [9] which partitions the NLP variables into basic, superbasic and nonbasic variables. The Hessian of the NLP Lagrangian function is approximated by limited-memory quasi-Newton updates, and a reduced Hessian algorithm is used for solving the QP subproblems. The nullspace matrix of the working set in each iteration is obtained from a sparse LU factorization.
When compared to standard "dense" SQP methods, the computational speed-up which can be achieved by fully utilizing the NLP structure is more than a factor of one hundred for typical discretized optimal control problems (Sect. 4). The described sparse direct collocation method is implemented in the software DIRCOL [24,25], which is freely available from the authors and has already been distributed upon request to about 40 institutions from research and industry.

3.5 Nonlinear Position Control of Optimal Setpoint Trajectories

On the stabilization level of our driver model, a nonlinear position control of the vehicle's center of gravity along the road is employed to guide the full vehicle dynamics model [14,22]. We use the computed optimal trajectory \((X^*(t), Y^*(t)) = (w_x(t), w_y(t))\) of the center of gravity as a setpoint ("target") for the actual position \((X(t), Y(t))\) of the vehicle. The setpoint trajectories \(((w_x(t), w_y(t))\) are repeatedly computed by solving optimal control problems for the reduced vehicle model (cf. Sections 3.2, 3.3 and 3.4). The computed optimal controls \(u^*\) are not directly used here.

The position control algorithm is based on the theory of nonlinear system decoupling and control originally developed for robot control [8]. Several state variables of the full vehicle dynamics model in veDyna enter into the control law, among them the direction of the velocity of the vehicle’s center of gravity \(v = \psi - \beta\), the side slip angle \(\beta\), the actual position \((X(t), Y(t))\) and velocity \((\dot{X}(t), \dot{Y}(t))\), as well as the position of the target \(((w_x(t), w_y(t))\), its velocity \((\dot{w}_x(t), \dot{w}_y(t))\) and its acceleration \((\ddot{w}_x(t), \ddot{w}_y(t))\). We compute a desired front lateral force \(S_v^{des}\) and a desired longitudinal force \(H^{des}\) by

\[
S_v^{des}(t) = -S_h(x(t)) - m \left[ \left( \cos(v(t)) \beta(t) + \sin(v(t)) \right) a_x^{des}(t) + \right.
\]
\[
\left. \left( \sin(v(t)) \beta(t) - \cos(v(t)) \right) a_y^{des}(t) \right] \tag{35}
\]
\[
H^{des}(t) = T(x(t)) + m \left[ \cos(v(t)) a_x^{des}(t) + \sin(v(t)) a_y^{des}(t) \right],
\]

where the desired second derivatives \(a^{des} = (a_x^{des}, a_y^{des})^T\) of the coordinates \(X, Y\) are given in inertial coordinates and calculated as follows

\[
a_x^{des}(t) := A \left( \dot{w}_x^{cor}(t) - X(t) \right) + 2\sqrt{A} \left( \ddot{w}_x - \ddot{X}(t) \right) + \ddot{w}_x \]
\[
a_y^{des}(t) := A \left( \dot{w}_y^{cor}(t) - Y(t) \right) + 2\sqrt{A} \left( \ddot{w}_y - \ddot{Y}(t) \right) + \ddot{w}_x. \tag{36}
\]
4 Numerical Results for Virtual Test Drives

4.1 Minimum Time Double Lane Change of a Passenger Car

As an example, we consider the ISO double lane change which shall be performed in minimum time with a standard passenger car, i.e., a BMW E30 with rear wheel drive.

First, on the anticipation level of our driver model, the optimal control problem for the corresponding single-track model (Sects. 3.2, 3.3) is solved by the method of Sect. 3.4. The control constraint during the lane change consists of a maximum speed of $v_{\text{max}} = 33 \text{ [m/s]} \approx 119 \text{ [km/h]}$, and the weights are given by $\rho_0 = 5 \cdot 10^{-6}$, $\rho_1 = 10^{-7}$, $\rho_2 = 10^3$. The problem formulation also includes an optimal acceleration maneuver on a straight road before the actual double lane change is performed. There, the vehicle is accelerated from the initial velocity $v(t_0) = 8 \text{ [m/s]} \approx 29 \text{ [km/h]}$ to its maximum value $v_{\text{max}}$ within 5.1 s which is kept until $t^* = 11.2958 \text{ s}$.

The numerical results for different improvements V1 – V4 of the sparse direct collocation method DIRCOL are reported in Table 2 (standard SQP: NPSOL-5.0 [10], sparse SQP: SNOPT-5.3-5 [9]). For reasons of demonstration, only equidistant grid points have been used here. Though, usually it is much more efficient to apply grid refinement based on suitable local error monitoring functions [24]. The speed-up factor for the improved collocation methods depends on the grid size; for the most advanced version V4 of DIRCOL, it is more than 500 when using 161 grid points (Table 2). The computed solution for the finest grid is depicted in Fig. 5. It should be noted, however, that for the online computation it is not necessary to solve the problem on the finest grid. A moderate accuracy of the solution is sufficient for which a computational time smaller than the time needed for the maneuver or its simulation can be achieved.

The optimality condition (29) for the computed trajectory is satisfied to reasonable accuracy (Fig. 5, right), although the adjoint differential equations

![Fig. 5. Numerical solution of the minimum time double lane change maneuver with the extended nonlinear single track model for 161 grid points. Left: trajectory in the $(X,Y)$-plane. Right: comparison of the computed control $u_2 = \delta_v$ (solid, red line) and the function $-\lambda_7/(2\rho_2\lambda_8)$ (dashed, green line) over time (cf. Eq. (29))](image-url)
Table 2. Numerical results for the minimum time double lane change with the extended single-track model for the varying utilization of the NLP structure and sparsity

| \( n_t \) | \( n_y \) | \( n_a \) | \( n_b \) | \( n_s \) | Jacbn | itn. | CPU [s] | PIII/500 Mhz CPU speed-up | memory | \( J[u] \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 200 | 160 | 42 | — | 9.3% | V0 | 32 | 0:26 | 1.0 | 1.4 | 11.28235 |
| | | | | | | V1 | 58 | 0:36 | 0.7 | 1.4 | 11.28233 |
| | | | | | | V2 | 36 | 0:05 | 5.2 | 1.7 | 11.28235 |
| | | | | | | V3 | 30 | 0:03 | 8.7 | 1.3 | 11.28235 |
| | | | | | | V4 | 24 | 0:01 | 26.0 | 0.5 | 11.28235 |
| 41 | 400 | 320 | 82 | — | 4.7% | V0 | 55 | 7:17 | 1.0 | 5.3 | 11.29100 |
| | | | | | | V1 | 86 | 9:57 | 0.7 | 5.3 | 11.29102 |
| | | | | | | V2 | 151 | 2:27 | 3.0 | 6.6 | 11.29142 |
| | | | | | | V3 | 44 | 0:20 | 21.9 | 4.8 | 11.29107 |
| | | | | | | V4 | 23 | 0:03 | 145.7 | 1.2 | 11.29147 |
| 81 | 800 | 640 | 162 | — | 2.4% | V0 | 75 | 73:30 | 1.0 | 21 | 11.29657 |
| | | | | | | V1 | 110 | 108:40 | 0.7 | 21 | 11.29658 |
| | | | | | | V2 | 74 | 9:27 | 7.8 | 25 | 11.29660 |
| | | | | | | V3 | 46 | 2:00 | 36.8 | 17 | 11.29670 |
| | | | | | | V4 | 65 | 0:29 | 152.1 | 2 | 11.29660 |
| 161 | 1600 | 1280 | 322 | — | 1.2% | V0 | 56 | 486:40 | 1.0 | 83 | 11.29852 |
| | | | | | | V1 | 39 | 329:10 | 1.5 | 83 | 11.29853 |
| | | | | | | V2 | 26 | 31:40 | 15.4 | 94 | 11.29852 |
| | | | | | | V3 | 18 | 7:58 | 61.1 | 64 | 11.29856 |
| | | | | | | V4 | 24 | 0:55 | 530.9 | 3 | 11.29852 |

V0: standard finite differences + standard SQP
V1: analyt. deriv. & fin. diffcs. + standard SQP
V2: analyt. deriv. & fin. diffcs. + sparse SQP – no structure
V3: analyt. deriv. & fin. diffcs. + sparse SQP + NLP-structure
V4: analyt. deriv. & fin. diffcs. + sparse SQP + NLP-structure + sparse Jacobian

have not been solved explicitly. Furthermore, condition (28) is also satisfied with high accuracy. The computed estimate \( \hat{\lambda}_0(t_f) = 4.999325 \cdot 10^{-6} \) agrees with the optimal value \( \lambda_0(t_f) = \rho_0 = 5 \cdot 10^{-6} \) to four digits.

In the second step, the computed optimal trajectories for the single-track model serve as near optimal setpoint trajectories for the nonlinear real-time position control of the full motor vehicle dynamics model in veDyna (Sect. 2.1). The difference between the state space trajectories \((X^*, Y^*)\) and \((X, Y)\) of the reduced and the full vehicle dynamics model is hardly visible. A zoom into the end of the double lane change (cf. Fig. 6) shows a maximum difference of about 7 cm. The small tracking errors for the double lane change maneuver at high speed is quite remarkable for the only rear wheel driven car, and demonstrates both, the capabilities of the position controller of the driver
model, as well as the good prediction of the full vehicle dynamics behavior by the significantly smaller, extended single-track model.

### 4.2 Further Vehicle Maneuvers

Recently, Lim and Hedrick suggested a longitudinal and lateral vehicle controller for automated vehicle operation of passenger cars [13]. The position

![Fig. 6. Zoom into the state space trajectories for the last part of the double lane change maneuver with a BMW E30 at 119 km/h. Left: center of the front axle (---) and its near optimal setpoint trajectory (- - -). Right: deviation from the setpoint trajectory for the recent position controller (---) as compared to an earlier version (- - -).](image)

![Fig. 7. Results for the driving cycle of [13]. Left: numerical simulation of the longitudinal and lateral acceleration of the vehicle's center of gravity with veDYN A and the position controller from Sect. 3.5. Right: longitudinal and lateral deviations of the setpoint trajectory](image)
controller was applied to a challenging test maneuver with quick changes in road curvature and braking in turns (Fig. 7, left). High lateral accelerations of about 5 m/s occur. We have investigated the same maneuver with the full vehicle dynamics model of a BMW E30 and the position controller from Sect. 3.5 which was used to track the given setpoint trajectory. For our approach (Fig. 7, right), we observed only one tenth of the deviation from the setpoint trajectory that have been reported in [13].

Finally, a fast virtual test drive of the full-scale vehicle dynamics model of a BMW E30 has been performed along the BMW handling course at Aschheim which is about 2 km long (Fig. 8). Animated results for further vehicle maneuvers are also available from http://www-m2.ma.tum.de/Projekte/kfz.

---

**Fig. 8.** Virtual test drive along the BMW test track at Aschheim including a short icy section of the road surface. The red line denotes the near optimal setpoint trajectory for the vehicle’s center of gravity on a dry road.
Acknowledgements. The authors thank Dr.-Ing. Karl Naab of BMW AG for helpful discussions and support and Dipl.-Math. Torsten Butz for carefully proofreading the manuscript. This research has been supported by the Federal BMBF under grant 03-BU7TM2-8.

References

Theoretical and Experimental Studies of an S-Catamaran

Klaus Kirchgässner¹, Som Deo Sharma², Xue-Nong Chen³, and Norbert Stuntz¹

¹ Mathematisches Institut A, Universität Stuttgart, Germany
² Institut für Schiffstechnik, Gerhard-Mercator-Universität Duisburg, Germany
³ VBD–Europäisches Entwicklungszentrum für Binnen- und Küstenschifffahrt, Duisburg, Germany

Abstract. Using a nonlinear shallow-water solitary-wave theory it was demonstrated that for a ship moving at supercritical speed along the centerline of a rectangular channel, if the hull sectional-area curve is of a special form determined by the solution of an oblique double-soliton interaction and the channel width is chosen to ensure complete wave cancelation through sidewall reflection, the ship waves can be made to form a purely localized pattern around the ship so that its wave resistance, which results only from far-field free waves, theoretically vanishes. To get rid of the crucial dependence on impractical sidewall reflection, this mechanism was developed further to obtain a novel catamaran comprising twin hulls with curved centerlines, yaw and skegs; it has theoretically zero wave-resistance at a chosen supercritical design speed in laterally unrestricted shallow water. Despite certain deviations from the ideal form for practical reasons, the wave-resistance of the new curved-yawed-hull catamaran with and without skeg was numerically found to be less than that of an equivalent straight-unyawed-hull catamaran by 50 and 30%, respectively. Now, the new design, albeit without skeg, has been validated by model experiment and comparison with a state-of-the-art reference catamaran of equal main dimensions that was developed and tested earlier in the VBD. Up to 28% wave-resistance reduction was achieved in the experiment, although not in the originally designed configuration but at a reduced yaw angle found by trial and error.

1 Introduction

The S-catamaran has twin mirror-symmetric component hulls, each of which has a slightly curved centerline (S-form), a global yaw angle and a suitable skeg, such that the overall configuration has strict transverse symmetry. The hull sections including skegs, curved centerline and transverse separation are theoretically determined to eliminate or at least substantially reduce wave resistance at a supercritical design speed in shallow water. In previous papers the S-catamaran was also called super-catamaran. The idea originated from a fortuitous observation followed by a theoretical analysis.

It was found by numerical computation and verified by physical measurement (Chen & Sharma 1994) that the wave resistance of a ship model
moving at supercritical speed parallel to the tank axis is reduced significantly by shifting its track from the tank centerplane to a certain speed-dependent location near one of the tank sidewalls. It was surprising since one might generally expect a symmetric configuration to produce the least resistance. As can be seen from the numerical calculation, the reason for it was that the bow wave after reflexion from the near sidewall eventually hits the afterbody and counteracts the stern wave so that the resultant wave in the wake is weakened, see Fig. 1. At supercritical speeds in shallow water the ship wave pattern is almost like the shock wave of an aerofoil in supersonic flight. The bow creates a free-surface elevation; the stern, a depression. Here in Fig. 1 the starboard bow-wave (on the near-wall side) is reflected by the tank sidewall and in extending across the stern (under the keel) it almost completely cancels the port stern-wave (on the far-wall side) as evidenced by comparison with the strong starboard stern-wave (on the near-wall side).

Fig. 1. (a) Calculated wave pattern of a standard ship model of Series 60 hull form in off-center motion (from left to right) parallel to channel axis at 20% channel width from the near wall at supercritical depth Froude number $U = 1.3$. Its wave resistance reduction is about 30% with respect to the motion along the channel centerline. (b) Schematic of the mechanism of wave resistance reduction by sidewall reflexion

Naturally, one is inspired by Fig. 1 to move also the far wall closer to the ship so that the starboard stern-wave would be canceled as well by the reflected port bow-wave. By numerical experimentation the optimal channel width for a symmetric ship-channel configuration at depth Froude number 1.3 was obtained. Coincidentally, it turned out to be 4.9 m, just half of the
original value. The wave resistance is now reduced by 75%. The associated wave patterns in the optimal narrow channel and in the original channel are shown in Fig. 2. It is obvious that the resultant waves in the ship wake are much weaker in the optimal-width channel than in the original. The lower energy density in the optimal channel multiplied with its reduced width explains the dramatic drop in wave resistance.

![Fig. 2. Calculated wave patterns of a standard ship model of Series 60 hull form moving along the channel centerline at depth Froude number $U = 1.3$ in (a) original channel of width $w^* = 9.81m$ with specific wave resistance $C_{w_o} = 0.051$ and (b) optimised channel of width $w^* = 4.9m$ with specific wave resistance $C_{w_o} = 0.013$; achieved wave resistance reduction $(0.051 - 0.013)/0.051 = 75\%$.](image)

This success stimulated a complete theoretical investigation of this intriguing phenomenon and eventually led to the discovery that the wave resistance of a ship in a channel can be made to vanish within the framework of a linear shallow-water wave approximation and, furthermore, even in a more accurate nonlinear theory. By analogy to electrical conductors, which are known to become superconductive (zero electrical resistance) under certain conditions, the name “shallow channel superconductivity” was proposed for this phenomenon, see Chen & Sharma (1997) and Pöppe (1998).

This phenomenon of zero wave resistance is not just an intellectual curiosity but seems to have a potential of practical application. It is possible to achieve this superconductivity also without the help of admittedly unrealistic sidewall reflection and, therefore, independently of channel width, namely, by means of catamarans with appropriately shaped sections and curved-yawed hulls.

In order to calculate accurately the wavemaking of a catamaran, either straight or curved, yawed or unyawed, it is necessary to consider the crossflow under the keel. The basic method to treat this effect is the concept of blockage of the crossflow under a 2-D body floating in shallow water as quantified by
an asymptotic potential difference between the two sides (Newman 1969). This was exploited in a nonlinear matched asymptotic theory by Chen & Sharma (1994) to treat the asymmetric flow about a slender ship moving in shallow water with the asymmetry arising from yaw angle and/or off-center motion in a channel. The theory was numerically implemented in the computer program SHALLOWTANK and verified by several model tests in the VBD. Its generalization to locally varying yaw angle allows the treatment of a curved-yawed hull. Hence, in particular, the hull can be so curved and yawed that bow waves are created on one side only. The theoretical centerline resembles the letter S.

A curved-yawed hull and its mirror image can now be configured so as to produce a waveless catamaran. This is achieved by applying a two-soliton solution of the KP equation to fit the afterbody to the forebody with respect to both cross-sectional area and local yaw angle. The aim is that at the supercritical design speed the bow wave generated only on the inner side by each of the two forebodies should hit the other afterbody and cancel its stern waves on both sides. Consequently, the catamaran has theoretically zero wave-resistance. A more detailed description and further mathematical development are given in Chen's dissertation (1999).

The curved-yawed-hull catamaran was designed within the framework of the Kadomtsev-Petviashvili (KP) theory, see Sharma & Chen (1997). It had the same principal dimensions as a reference catamaran designed and tested much earlier in the VBD, see Heuser (1973). The new design was numerically evaluated over a large speed range at the design water-depth using the computer program SHALLOWTANK, both in the curved-yawed-hull and in a conventional straight-unyawed-hull configuration. The present paper reports, for the first time, results of a physical model-test of the novel S-catamaran recently carried out in the VBD. The measured waves and wave resistance of the S-catamaran are compared with those of its component hull and of the reference catamaran. Up to 28% reduction in wave-resistance as compared to the reference catamaran is found. Although wave cancelation to the theoretically expected extent could not be observed in this first experiment with an S-catamaran, the considerable reduction of wave resistance achieved shows that further theoretical and experimental effort would be worthwhile.

2 Theory of the S-Catamaran

The basic idea in the construction of the S-catamaran is that the two component hulls are so shaped in terms of cross-section and centerline that at the supercritical design-speed the forebodies generate waves only on the inner sides. At the same time, the transverse separation of the two hulls is so chosen that the afterbody waves are canceled completely by the incident bow wave from the opposite hull. In other words, the two bow waves are to be absorbed perfectly by the opposite afterbodies. This may be called the no-outward-wave condition. Each bow wave can be an inward oblique soliton.
The two identical solitons stemming from the two forebodies undergo an oblique interaction in the area between the two hulls. The ship geometry is determined inversely by the nonlinear solution of interacting solitons.

2.1 Simplified Mathematic Model

Consider a catamaran with two hulls of length $l^*$, beam $b^*$ and draft $d^*$ moving along the centerline of a shallow channel of water depth $h^*$ and width $w^*$ at speed $U^*$. The hulls need not be straight, i.e., each hull can be curved and yawed as defined by a local yaw angle $\psi(x)$, see Fig. 3. This is indeed the basic recipe for reducing wave resistance here. The two component hulls are mirror-symmetric to each other yielding overall transverse symmetry. The same symbols are used to denote dimensional variables and their nondimensional forms, the former being marked with asterisks. Unless otherwise stated, all variables throughout this paper are nondimensionalised by reference to water depth $h^*$, the acceleration due to gravity $g^*$, and water density $\rho^*$. For example, $l = l^*/h^*$ is nondimensional ship length, $S_0 = S_0^*/h^{*2}$ is nondimensional midship section area, $s = s^*/h^*$ is nondimensional separation between the mid-points of the midship sections of the two hulls, etc. Note that nondimensional speed $U = U^*/\sqrt{g^*h^*}$ is usually called the depth Froude number $F_{nh}$.

![Fig. 3. Schematic of the problem: (a) local yaw angle, (b) near-field cross-flow](image)

Due to symmetry, only the port half-channel will be considered in the following. A righthanded Cartesian coordinate system $Oxyz$ moving at the same speed as the ship is used with origin $O$ located in the center-point of the port hull, plane $Oxy$ on the quiet free surface, $z$ positive upward, and $x$ positive forward. The problem is formulated and solved by the method of matched asymptotic expansions in Chen & Sharma (1994). Here the mathematical model is simplified for both far and near fields in order to obtain analytic solutions and to design a catamaran explicitly. Therefore, the far-field equation is just repeated with a few explanations, and the crucial matching condition is rederived briefly by simple reasoning.
In order to obtain an explicit two-soliton solution, the KP equation is taken for the far-field, although it is not the best model in the supercritical range. Its stationary form reads

\[(1 - U^2)\varphi_{xx} + \varphi_{yy} + 3U\varphi_x\varphi_{xx} + \frac{U^2}{3} \varphi_{xxxx} = 0,\]  

where \(\varphi\) is depth-averaged velocity potential.

In the near field the shallow-water slender-body theory is applied. The flow in the near field is sketched in Fig. 3 (b). Two key results of the slender-body theory are employed: (i) the asymptotic normal velocity is equal to the cross displacement-velocity due to lengthwise change of cross-sectional area of the hull, superimposed on the mean cross-velocity due to asymmetry; (ii) the mean cross-velocity yields an asymptotic potential difference between the two sides of a body floating in shallow water (Newman 1969). In mathematical terms,

\[
V_n|_{y=\pm 0} = \mp \frac{1}{2} V_\tau \frac{dS}{dx} + V_n, \tag{2}
\]

\[
\Delta \varphi = 2\bar{V}_n C(x), \tag{3}
\]

where

\[
\Delta \varphi = \varphi|_{y=+0} - \varphi|_{y=-0} \tag{4}
\]

is the potential jump, \(y = \pm 0\) or \(0^\pm\) denote the port and starboard sides of the port component-hull, \(V_n\) is local normal velocity, \(\bar{V}_n\) is its mean value, \(V_\tau\) is local tangential velocity, \(S(x)\) is lengthwise cross-sectional area distribution, and \(C(x)\) is blockage coefficient at each cross-section \(x\). In terms of far-field variables, the normal and tangential velocities can be expressed and approximated by neglecting disturbance velocities \(\varphi_x, \varphi_y\), which are small compared to the ship speed \(U\), as

\[
V_n = \frac{\partial \varphi}{\partial y} \cos \psi - \left(\frac{\partial \varphi}{\partial x} - U\right) \sin \psi \approx \frac{\partial \varphi}{\partial y} \cos \psi + U \sin \psi, \tag{5}
\]

\[
V_\tau = -\left(\frac{\partial \varphi}{\partial x} - U\right) \cos \psi - \frac{\partial \varphi}{\partial y} \sin \psi \approx U \cos \psi. \tag{6}
\]

Substituting (5), (6) and (3) into (2), one obtains an approximate matching condition,

\[
\frac{\partial \varphi}{\partial y}|_{y=\pm 0} = \mp \frac{1}{2} U \frac{dS}{dx} + \frac{\Delta \varphi}{2C(x) \cos \psi(x)} - U \tan \psi(x). \tag{7}
\]

If local yaw angle \(\psi(x)\) is small, then \(\cos \psi \approx 1\) and (7) reduces to

\[
\frac{\partial \varphi}{\partial y}|_{y=\pm 0} = \mp \frac{1}{2} U \frac{dS}{dx} + \frac{\Delta \varphi}{2C(x)} - U \tan \psi(x). \tag{8}
\]
This serves as a simplified boundary condition at the ship location for the KP equation (1). A more accurate boundary condition involving the disturbance velocity and local free-surface elevation has been laboriously derived and implemented in the computer program SHALLOWTANK. The mathematical model still has to be closed by imposing the boundary condition \( \varphi_y = 0 \) on the channel sidewall and centerplane, as well as the Kutta condition at the ship stern \( x = -l/2 \),

\[
\varphi_x \big|_{y=0^+} = \varphi_x \big|_{y=0^-} .
\]

2.2 Super-Catamaran Solution

The prefix "super" stands here for both superconductive and supercritical. A super-catamaran, by definition, would radiate no waves outward or into the wake and, hence, have zero wave-resistance at its supercritical design speed. This is achieved in the following by a logical combination of two steps.

First, the elimination of outward waves requires that the transverse velocity must vanish on the outer side,

\[
\frac{\partial \varphi}{\partial y} \big|_{y=0^+} = 0 .
\]

This means that the potential along the outer side (the port side) of the port hull is constant in the far-field. As there is no discontinuity at the bow, this constant is equal to the value of the potential on the inner side (the starboard side) of the port hull at the bow,

\[
\varphi(x,0^+) = \text{const.} = \varphi(l/2,0^-) .
\]

By using (8) and (11), the no-outward-wave condition (10) yields a prescription for the local yaw angle,

\[
\tan \psi(x) = -\frac{1}{2} \frac{dS}{dx} + \varphi(l/2,0^-) - \varphi(x,0^-) \left( \frac{2UC(x)}{2UC(x)} \right) .
\]

Here, \( \varphi(x,0^-) \) must be determined by the inner-side condition. Substituting (12) into (8) for the inner side, \( y = 0^- \), easily provides the boundary condition on the inner side,

\[
\frac{\partial \varphi}{\partial y} \big|_{y=-0} = U \frac{dS}{dx} .
\]

So far it has been ensured that a hull curved and yawed according to (12) will generate no waves on its outer side, on which the transverse velocity is zero. Therefore, it pushes the entire displacement flow toward the inner side, on which the transverse velocity thus becomes twice as much as that for a straight-unyawed hull of same cross-sectional area. Consequently, all the
waves are generated on the inner side. Normally, they would extend into the wake behind the catamaran, with or without reflection from the inner side of the other hull. The next step, therefore, must ensure that the waves are confined to the area between the two hulls. Since the problem defined by (1) together with (13) is exactly the same as that of a straight-unyawed hull with twice the displacement of the curved-yawed hull, all that is needed is merely an application of the original idea of shallow-water superconductivity in a rectangular channel, see Chen & Sharma (1997).

Second, the elimination of wake waves requires that the stern waves be canceled by the bow wave of the opposite hull. The afterbody cross-sectional area distribution should be so selected that the bow wave generated by the inner side of each component hull and incident upon the inner side of the other hull will be perfectly absorbed by its afterbody. The bow wave can be an oblique soliton. Thus the two oblique solitons interact with each other in the area between the two hulls. The two afterbodies are then determined by the incident wave. This two-soliton solution of the KP Equation (1) was presented in Chen & Sharma (1997) and can be used again here to derive the hull shape. The major expressions needed to determine the catamaran are: the potential on the inner side, which is used in (12),

$$\varphi(x, 0^-) = \frac{8Uk}{3} \cosh(2kx_1) + \sqrt{A_2} \exp(2kx); \quad (14)$$

the cross-sectional area distribution, which is obtained inversely from (13) as \(\varphi|_{y=0^-}\) is known,

$$S_\alpha(x) = \frac{8}{3} \frac{k\sqrt{U^2 - 1 - AU}}{\cosh(2kx_1) + \sqrt{A_2} \cosh(2kx)} \sinh(2kx_1); \quad (15)$$

and the mean separation between the two hulls,

$$s = 2x_1/\sqrt{U^2 - 1 - AU}; \quad (16)$$

with

$$k = \frac{\sqrt{3AU}}{2U}, \quad A_2 = \frac{U^2 - 1 - AU}{U^2 - 1 - 4AU}. \quad (17)$$

Here, \(A\) and \(x_1\) are two free parameters, allowing the design of various hulls. Theoretically, \(x_1\) is arbitrary and \(0 < A < (U^2 - 1)/(4U)\). The theoretical ship length \(l\) is infinite but \(S_\alpha(x)\) vanishes exponentially at infinity so that it can be safely truncated at a reasonable length. Since the disturbance velocity at the stern on the inner side also vanishes at \(x = -\infty\), the Kutta condition (9) is satisfied.

Actually, to construct the catamaran, the two above steps are reversed, i.e., one starts with the two-soliton interaction solution between the two hulls and then derives the cross-sectional area distribution as well as the local yaw angle.
3 Design of an S-Catamaran

The design depth Froude number was chosen as $U = 1.7$, corresponding to an absolute speed $U^* = 8.42 \text{ m/s}$ in water of depth $h^* = 2.5 \text{ m}$. Ship length $l^* = 34.13 \text{ m}$ and ship draft $d^*_0 = 1.1 \text{ m}$ were the same as those underlying the VBD model at scale $1:10$, see Heuser (1973). Values of parameters $x_1$ and $A$ were sought to match the other principal dimensions (displacement volume $V^*$ and midship-sectional area $S^*_o$). By trial and error, the desired solution was found to be $x_1 = 2.85$ and $A = 0.17754$.

<table>
<thead>
<tr>
<th>Item</th>
<th>Symbol (Unit)</th>
<th>VBD M601</th>
<th>S-Catamaran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length at waterline</td>
<td>$l^*$ (m)</td>
<td>34.13</td>
<td>34.13</td>
</tr>
<tr>
<td>Beam at midship</td>
<td>$b^*_0$ (m)</td>
<td>3.33</td>
<td>3.102</td>
</tr>
<tr>
<td>Draft</td>
<td>$d^*_0$ (m)</td>
<td>1.10</td>
<td>1.100</td>
</tr>
<tr>
<td>Area of midship section</td>
<td>$S^*_o$ (m$^2$)</td>
<td>3.02</td>
<td>3.024</td>
</tr>
<tr>
<td>Displacement</td>
<td>$V^*$ (m$^3$)</td>
<td>50.65</td>
<td>50.65</td>
</tr>
<tr>
<td>Wetted surface area</td>
<td>$S^*_w$ (m$^2$)</td>
<td>106.40</td>
<td>109.9</td>
</tr>
<tr>
<td>Block coefficient</td>
<td>$C_B = V^* / l^* b^<em>_0 d^</em>_0$</td>
<td>0.405</td>
<td>0.4349</td>
</tr>
<tr>
<td>Midship coefficient</td>
<td>$C_M = S^<em>_o / b^</em>_0 d^*_0$</td>
<td>0.8245</td>
<td>0.8863</td>
</tr>
<tr>
<td>Wetted surface coefficient</td>
<td>$C_{WS} = S^<em>_w / \sqrt{V^</em> l^*}$</td>
<td>2.5591</td>
<td>2.6433</td>
</tr>
<tr>
<td>Hull separation</td>
<td>$s^*$ (m)</td>
<td>13.32</td>
<td>12.72</td>
</tr>
<tr>
<td>Relative separation</td>
<td>$K = s^* / b^*_0$</td>
<td>4.00</td>
<td>4.10</td>
</tr>
<tr>
<td>Mean yaw angle</td>
<td>$\psi$ (deg)</td>
<td>0</td>
<td>4.86</td>
</tr>
<tr>
<td>Water depth</td>
<td>$h^*$ (m)</td>
<td>2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 1. Principal dimensions and parameters of each component hull of the two catamarans (Note: Model tests were done to scale $1:10$)

The principal dimensions of the new design and the reference VBD catamaran are compiled in Table 1. The design procedure is outlined in following.

Since $S_a(x)$ is an even function, $S_a(-l/2) = S_a(l/2)$, the following operation,

$$S(x) = S_a(x) - S_a(l/2),$$

yields a truncated ship of finite length $l$ and sharp ends $S(-l/2) = S(l/2) = 0$. Since $l/2 - x_1$ is large enough, $S_a(l/2) << 1$, and the practical cross-sectional area distribution $S(x)$ is close to the theoretical one $S_a(x)$. The displacement of each component hull is

$$V^* = h^* \int_{-l/2}^{l/2} S(x) \, dx.$$  

The theory does not specify the detailed geometry of the cross sections. So they can be designed by practical considerations. Here, the hull lines are
described by exponential functions:

\[
y(x, z) = \pm \frac{b(x)/2}{1 - \exp(-f(x)d(x))} \left(1 - \exp\{-f(x)\left[z + d(x)\right]\}\right),
\]

where

\[
y = \pm \frac{1}{2} b(x)
\]

is the waterline;

\[
z = -d(x) = -d_0 \{1 - 0.9 \exp[-1.2(x + l/2)]\} \{1 - 0.99 \exp[6(x - l/2)]\}
\]

is the keel-line, and \(f(x)\) is a parameter function chosen to be

\[
f(x) = 20 \sech(0.3x).
\]

Integrating (20) over \(-d(x) < z < 0\) yields

\[
S(x) = 2 \int_{-d(x)}^{0} y(x, z) dz = \frac{b(x)}{1 - \exp(-f(x)d(x))} \left\{d(x) + \frac{1}{f(x)}[\exp\{-f(x)d(x)\} - 1]\right\}.
\]

Since \(S(x)\) is known from (18), \(b(x)\) can be determined from the above equation as

\[
b(x) = S(x) \frac{1 - \exp(-f(x)d(x))}{d(x) + [\exp\{-f(x)d(x)\} - 1]/f(x)}.
\]

Next, the hull centerline is determined by theory as moderated by practical requirements of real flow. It is found that, for the given clearance \(c = 1 - d_0 = 0.5\), the theoretical yaw angle according to (12) together with (14) is too large, the maximum value being about 18 degrees. In order to obtain a more suitable curvature, the following considerations are introduced: (i) Any yaw angle will cause more vortex shedding, thereby increasing viscous resistance and counteracting the favorable reduction of wave resistance. So, from the point of view of total resistance, the best local yaw angle should be smaller than the theoretically predicted value. (ii) The theoretical model based on potential theory with the Kutta condition at the stern does not hold very well for strong vortex shedding beginning well ahead of the stern. Generally speaking, this viscous effect will weaken the potential jump across the hull. Since the potential jump is determined by the outer field solution, it cannot account for this effect. To keep the term \(\Delta \varphi(x)/C(x)\) in (8) realistic, especially in the afterbody, either the blockage coefficient should be increased or the potential jump decreased artificially. The former artifice, tried as a numerical experiment as reported in Chen & Sharma (1994), was found to be useful for improving the prediction of side force and yaw moment on a ship at drift angle.
Theoretical and Experimental Studies of an S-Catamaran

Fig. 4. (a) Blockage coefficients according to (25) (solid line) and (26) (dotted line); (b) Local yaw angle of the curved-yawed starboard component hull

Fig. 5. (a) Cross-sectional area curve; (b) Beam at waterline

For the above reasons the theoretical potential jump in (12) is reduced by dividing it by a factor function as follows,

\[ \tan \psi(x) = -\frac{1}{2} \frac{dS}{dx} + \frac{\varphi(l/2, 0^-) - \varphi(x, 0^-)}{f_d(x)} \frac{1}{2UC(x)}, \] \hspace{1cm} (23)

where

\[ f_d(x) = 2.5 + \frac{(6 - x)}{12}. \] \hspace{1cm} (24)
Since \( \frac{dy_m}{dx} = \tan \psi(x) \), integrating (23) once yields the centerline \( y_m(x) \). Additionally, it is proposed to fit a suitably sized small skeg under the keel in the afterbody to purposely increase the blockage coefficient until it approximately reaches a value given by the following formula,

\[
C(x) = 0.95 \exp(-0.18x). \tag{25}
\]

The actual design may have to be done by experiment rather than by theory. In order to show the difference made by the skeg, the above curve (solid line) is drawn in Fig. 4 (a) along with the theoretical blockage coefficient (dotted line) for a rectangular cross section of beam \( b(x) \) and draft \( d_o \) as given by the formula of Taylor (1973),

\[
C(x) = \frac{1}{2} b(x)(1/c - 1) + \frac{2}{\pi} [1 - \ln(4c)] + \frac{2}{3\pi} c^2 + \frac{281}{90\pi} c^4, \tag{26}
\]

where \( c = 1 - d_o \) is the clearance beneath the keel.
The cross-sectional area and beam at waterline of the S-catamaran are presented in Fig. 5; blockage coefficients and local yaw angle, in Fig. 4; the body plan (ignoring curvature and yaw of the centerline), the side profile, and the waterline, in Fig. 6. Note that each hull-centerline has not only an S-shaped curvature but also a significant mean outward yaw angle.

4 Numerical Calculation

The hull separation used in the present computation is $s^* = 12.7209$ m ($s^*/b_o^* = 4.1011$) since numerical experiments at the design speed $U = 1.7$ revealed that its optimum value for the truncated hull $S(x)$ is slightly different from the theoretical design value $s^* = 11.3075$ m for the infinite hull $S_o(x)$. A straight-unyawed monohull, identical to each component hull of the proposed design except for curvature and yaw, is used as a basis for comparing wave resistance and ascertaining the favorable effect of wave interference. The catamaran calculations hold for a channel width $w^* = 98.94$ m, which corresponds approximately equal to the VBD tank-width (scale 1:10) and is large enough to avoid sidewall effects; taking advantage of symmetry, only half of it is simulated in the computer. The monohull calculations using the same grid for simplicity, without taking advantage of symmetry, hold for a symmetric channel of half the width and, annoyingly, show a strong sidewall effect, localized around $U = 1.2$, which is irrelevant for the comparison intended here. The computational domain extends longitudinally about 10 ship lengths upstream and about 20 ship lengths downstream. More details were reported in Sharma & Chen (1997).

Fig. 7 shows the calculated results of specific wave resistance $R_w/W = R_{w*}/W^*$, where displacement weight $W^* = \rho^* g^* V^*$, and specific lateral force $F_y/W = F_{y*}/W^*$ experienced by (i) the port hull of the curved-yawed-hull catamaran (super-catamaran), (ii) the port hull of the same catamaran

![Figure 7](image_url)

Fig. 7. Calculated specific wave resistance $R_w/W$ and lateral force $F_y/W$ vs. the depth Froude number $U$, where dots denote the straight monohull; “c”, the conventional straight-unyawed-hull catamaran; “s”, the curve-yawed-hull super-catamaran with skeg.
Fig. 8. 3-D plots of the wave patterns of the two catamarans. The vertical axis denotes $\zeta^*/(\epsilon h^*)$ with $\epsilon = 0.10789$. The vertical scale is exaggerated 10 times.
without curvature and yaw (conventional catamaran), and (iii) the straight-unyawed component hull (monohull), which can also be interpreted as the limiting case of a conventional catamaran of infinite hull-separation. The two catamarans have almost the same behavior in the subcritical and near-critical speed ranges, where their wave resistance is clearly higher than that of the monohull due to unfavorable wave interference. The catamaran advantage shows up in the supercritical speed range. In the interval $1.5 \leq U \leq 2.5$, the wave resistance of the conventional catamaran is already significantly lower than that of the monohull, and the curved-yawed-hull catamaran, of course, is even better. At the design speed $U = 1.7$ the specific wave-resistance values for monohull, conventional catamaran and super-catamaran are 0.02738, 0.01598, 0.007852, respectively. The superiority of the super-catamaran is by no means restricted to a narrow range around the design speed. In fact, the maximum relative wave-resistance reduction (about 80%) occurs at $U = 2$. The super-catamaran also develops much less squat than the monohull and the conventional catamaran in the speed range of wave-resistance reduction. But its lateral force is much larger than that of the conventional catamaran, obviously owing mainly to yaw and partly to curvature. The straight-unyawed monohull, of course, develops no lateral force or yaw moment.

To estimate the importance of the theoretical curve of blockage coefficient (solid line in Fig. 5), which can only be achieved by means of a skeg, a fourth case was calculated: curved-yawed-hull catamaran without skeg (dotted line in Fig. 5). Its specific wave-resistance at the design speed was found to be 0.01137, showing that substantial savings can be achieved even without a skeg.

Fig. 8 shows 3-D plots of the wave patterns of the two catamarans at $U = 1.7$ and is fully consistent with the finding of Fig. 7. It is seen that the super-catamaran (curved-yawed hulls with skeg) generates substantially lower waves on the outer side and in the wake than the conventional catamaran (straight-unyawed hulls), while its inner-side waves, trapped between the two hulls, are about twice as high as those of the conventional catamaran.

5 Model Experiment

To verify the foregoing theory a physical model experiment with an S-catamaran was recently carried out in the VBD towing tank. The S-catamaran was designed to have the same principal dimensions as two reference catamarans M601 (straight, unyawed hulls) and M635 (slightly cambered but unyawed hulls) developed and model-tested by Heuser (1973) in the VBD. So his systematic model tests can serve as a basis for comparison.

Besides the depth Froude number as basic variable, mean yaw angle and transverse hull-separation were chosen as easily variable parameters, since they have an obvious influence on resistance. In the original design, each hull-centerline has a maximum local yaw angle of 8.6 deg outward and a
mean value of 4.86 deg outward with respect to the direction of motion. The catamaran model was tested mainly without the skegs under the keel required by the theory. A few tests were also run with improvised skegs but they are omitted here.

Disappointingly, the measured total resistance of the S-catamaran in the original configuration turned out to be much higher than that of the reference catamarans at supercritical speeds. In retrospect, the reason seems to be clear. Each component hull of the S-catamaran acts as a 3-D foil of very low aspect ratio. The dynamic lift (here as side force) generated by yaw and curvature is inevitably associated with considerable vortex shedding and induced drag, not properly accounted for in the foregoing theory. Moreover, curvature and yaw may promote boundary-layer separation and hence enhance viscous pressure resistance, also not considered in the present potential-flow model.

Fig. 9 shows the measured specific total resistance of the S-catamaran and the S-monohull, both at the original designed yaw angle, in comparison with Heuser’s reference catamaran M635 and its monohull. As already mentioned, the S-catamaran disappointingly has a higher resistance than the reference catamaran at supercritical speeds. Not unexpectedly, the S-monohull has an even higher resistance. But it is encouraging that the beneficial catamaran effect, namely, the reduction of wave resistance by mutual wave cancelation of twin hulls, is much stronger in the S-catamaran than in the reference catamaran.

Presuming that the poor performance of the S-catamaran was due to an exaggerated mean yaw-angle of the component hulls, the model experiment was repeated with four stepwise decreased mean yaw angles, i.e., by rotating both component hulls inward by 2, 3, 4 and 5 deg, also because this was relatively easy to conduct without having to build new models with modified curvature. For each configuration, the viscous form-factor $1 + k$ based on the

![Fig. 9. Measured specific total resistance of four models](image)
ITTC 1957 Correlation Line was individually determined by the standard Hughes-Prohaska method and the thus estimated viscous resistance was deducted from the measured total resistance to obtain the "measured" wave resistance. The results are displayed in Fig. 10. It is remarkable that the best results are obtained by reducing the mean yaw angle from originally 4.86 deg to about zero. Then the total resistance is reduced by 11% and the wave resistance by 28% compared to the reference catamarans at a near-design speed $F_{nh} = 1.8$, as seen in Fig. 11. Unfortunately, specific viscous form-factors $1 + k$ could not be determined for the reference catamarans since no
Fig. 12. Measured wave cuts of the S-catamaran at $F_{nh} = 1.7$ with the mean yaw angle at original value (solid line) and reduced by 4 deg (dashed line). From top to bottom, the wave cuts shown are at transverse distance 1, 1.5, 2.0, 2.5, 3.0 and 3.5 m from the tank centerplane to port.

Resistance measurements at sufficiently low Froude numbers were available from old records. Therefore, the best value 1.15 obtained for an S-catamaran was applied. This may have still led to a slight overestimation of the reference
catamarans' viscous resistance and, hence, to an underestimation of its wave resistance. So the S-catamaran's advantage shown is on the safe side.

Furthermore, longitudinal cuts through the wave patterns of the towed models were recorded by means of an array of six stationary wave gauges installed in the tank on the port side at different transverse distances from the tank centerplane. Only two comparisons of major interest are reproduced here.

First, Fig. 12 compares the measured wave cuts of the S-catamaran in the original configuration to those at the mean yaw angle reduced by 4 deg; the speed is $F_{nh} = 1.7$. Obviously, the wave pattern is not sensitive to 4-5 deg reduction of yaw angle. Only the first crest becomes a little higher and the second a little lower when the mean yaw angle is reduced. This means that the wave resistance related to the wave pattern is also hardly affected by the yaw angle. It seems that the yaw angle is responsible mainly for a transverse force and the associated induced resistance. It is not clear how far the form factor, evaluated by the Hughes-Prohaska method, can account for the induced resistance. The "wave resistance" shown in Fig. 10 may, in fact, include a part of the induced resistance.

Second, Fig. 13 compares the measured wave cuts of the S-catamaran to those of the S-monohull (the starboard component hull), both towed at the original mean yaw angle at $F_{nh} = 1.7$. Since the S-monohull was towed along the tank centerline, whereas the twin hulls of the S-catamaran were both offset on opposite sides, certain adjustments are necessary to ensure that the wave cuts are compared at equal transverse distances from the relevant

![Fig. 13. Measured wave cuts of the S-catamaran (thick line) and the S-monohull (thin line) at $F_{nh} = 1.7$. (a) 2 m wave-gauge offset for S-cat; 1.5 m, for S-mono. (b) 2 m offset for S-cat; 2.5 m, for S-mono](image-url)
Fig. 14. Photographs of wave patterns of the starboard S-monohull (top) and of the S-catamaran (bottom) at $F_{nh} = 1.6$. Caution: Motion is here from right to left.
hull segments. The bow wave of the S-catamaran at 2 m offset is, therefore, compared to that of the S-monohull at 1.5 m offset in Fig. 13 (a). But the stern wave of the S-catamaran at 2 m offset is compared to that of the S-monohull at 2.5 m offset in Fig. 13 (b). Because of the outward yaw angle, the starboard component hull naturally makes a stronger bow wave on the port side than the S-catamaran. But it is remarkable that also in the wake the waves of the S-monohull are stronger than those of the S-catamaran whose displacement is twice as large. This is evidence of effective wave cancelation within the catamaran. The strong bow wave on the port side of the starboard hull is partly absorbed by the afterbody of the port hull. This phenomenon is clearly visible in the photographs reproduced in Fig 14.

6 Conclusion

A mathematical theory was developed for the practical design of a low-wave S-catamaran and verified by numerical calculation and physical model test in a towing tank. The calculated wave-cancelation could not be fully achieved in the experiment at first try, presumably because of the failure of theory to account for vortex shedding and induced resistance of the curved-yawed component hulls of the S-Catamaran. However, around the supercritical design speed a large reduction of wave resistance to an extent not attainable by a conventional catamaran was obtained. The experimental results are in good agreement with numerical predictions at reduced mean yaw angle.

Quantitatively, the measured wave resistance of the S-catamaran was 28% lower than that of a reference catamaran at $F_{nh} = 1.8$; the measured total resistance, 11%. The S-catamaran has 52% and 58% lower measured specific wave-resistance than its S-monohull at the original and reduced mean yaw angles, respectively, at $F_{nh} = 1.7$; Heuser’s reference catamaran, only 30%. Numerically, the S-catamaran with curved-yawed hulls but without skeg has 29% and 58.5% lower specific wave-resistance at $F_{nh} = 1.7$ than a straight-unyawed catamaran and a straight-unyawed monohull, respectively, of the same cross-sectional area-curve.

References

II. Environmental Technology

Robust Error Estimators for Interface Problems Occuring in Transport Processes in Porous Media
   J. Fuhrmann, M. Petzoldt

Modelling and Simulation of a Planned Bio-Chemical in situ Remediation
   W. Jäger, G. Wittum, W. Schäfer, Ch. Wagner, H. Willershausen

Influence of Surfactants on Spreading of Contaminants and Soil Remediation
   P. Knabner, S. Bitterlich, R. Iza Teran, A. Prechtel, E. Schneid

Improvement of Environment Observing Remote Sensing Devices by Regularization Techniques
   P. Maass, C. Böckmann, A. Mekler
Robust Error Estimators for Interface Problems Occuring in Transport Processes in Porous Media

Jürgen Fuhrmann and Martin Petzoldt

Weierstraß-Institut für Angewandte Analysis und Stochastik im Forschungsverbund Berlin e.V., Mohrenstraße 39, 10117 Berlin, Germany

1 Introduction

The aim of the project is the adaptive numerical simulation of transport processes in porous media within the groundwater simulation code FEFLOW [4]. These processes can be very complex and take place in highly heterogeneous media. The nature of the describing equations is such that error estimators for the full nonlinear model can be defined only heuristically based on considerations on known model problems and numerical experiments. Here one could only proceed with heuristical error indicators, which give a criterion where to refine the grid to become probably more accurate. As adaptivity and error control suggest to the engineer that we really give him a measure of the error – i.e. an error estimate, in consensus with the industry partner, we disregarded the heuristical approach as much as we could and rather tried to verify both theoretically and experimentally estimators for a relevant problem class where we still are able to proceed with mathematical theory.

Such a problem is the flow of groundwater in a confined, heterogeneous aquifer. The strong form of this linear problem in the stationary case looks like follows:

\[-\nabla \cdot k(x)\nabla u(x) = f(x)\]

where \(k\) is the permeability of the medium and \(f\) contains sources. Additionally, we impose boundary conditions of first, second or third kind. It is important to note that \(k\) may have large variations and may be discontinuous. One may think of several layers of soils, where the permeability takes different values within each layer. During this paper, we will assume that \(k\) is piecewise constant causing a partition of the computational domain into several subdomains with constant \(k\), and we will call this type of problem interface or transmission problem.

A natural wish is to be able to control the error made by the approximation of this problem with finite elements. This rises the In this paper, we will mainly discuss the issue of error estimation for an interface problem like that described above.

W. Jäger et al. (eds.), Mathematics - Key Technology for the Future © Springer-Verlag Berlin Heidelberg 2003
2 Approximation Properties of Finite Elements in the Presence of Singularities

If we look more carefully on the problem, we see that the error is large at so called singularities. Singularities are parts of the solution with lower smoothness, where the gradient changes strongly. It is known that in our problem, singularities can occur at re-entrant corners of the outer boundary of the problem, at places where boundary conditions change and at interfaces between domains with different permeabilities.

![Fig. 1. functions with singularities a) λ ≈ 0.7  b) λ ≈ 0.99  c) λ ≈ 0.1](image)

In Fig. 1 we plot examples of typical functions with singularities. Note that the functions in Figs. 1a,b) are rather smooth in the neighborhood of the origin, whereas the function from Fig. 1c) is very steep near the origin. The smoothness of the singularities can be measured by its so called degree $\lambda$ (the solution is contained in the Sobolev space $H^{1+\lambda-\epsilon}, 0 < \epsilon$). This is a positive number between 0 and 1. The singularities with small degree $\lambda$ are worse then those with larger degree $\lambda$ (compare with Fig. 1).

It is known that the degree of the singularity influences the degree of approximation with finite elements [9]. The smaller the degree of the singularity the worse is the approximation with finite elements. Usually, singularities cause the major part of the approximation error in interface problems. For this reason it is useful to study the dependence of the degree of the singularity on the geometry and the permeability.

We want to point out that the degree of the singularity depends also on $k$ but we are interested in results which are independent of the jumps of $k$.

In Fig. 2 we plot three different geometrical situations. In case a) the permeability in the neigbourhood of a point takes two different values and $1/2 < \lambda$ [10]. In Fig. 2b) the permeability takes three different values and we can show that $1/4 < \lambda$ [12]. When the permeability takes four or more different values (see Fig. 2c)) the degree of the singularity can be arbitrary close to 0 [10]. But if we assume that in a neigborhood of the singularity
Fig. 2. degree of singularity a) $1/2 < \lambda$  b) $1/4 < \lambda$  c) $0 < \lambda < 1$

![Fig. 2. degree of singularity](image)

Fig. 3. a) quasi-montone case b) non quasi-montone case

![Fig. 3. a) quasi-montone case b) non quasi-montone case](image)

we can find continuous paths not containing the singular point itself along which we can move from any subdomain to that with the highest value of the permeability $k$ with an increasing value of $k$, we can show $1/4 < \lambda$ [12]. Such a situation is shown in Fig. 3a). This case is called quasi-monotone. Thus, we see that in the quasi-monotone case the singularity is not very "bad". The situation depicted in Fig. 3b) is not quasi-monotone, since we cannot move from the subdomain with permeability 10 to that with permeability 20 in a monotone way. One has to move through subdomains with permeability 1. In such a situation the singularity can be arbitrarily bad.

If one knows the singularities and their degree $\lambda$ one can adapt the discretization to these singularities and reduce the error with optimal order [9]. But this approach has several drawbacks. First, it is not feasible to obtain the degree of a singularity in a generic situation. Second, there may occur hundreds of singularities, but only few of them have larger contributions to the solution. Only those with larger contributions need to be treated but one is not able to identify them.

3 A Posteriori Mesh Adaptation
Based on a Robust Residual Based Error Estimator

An alternative approach is to look for a method which works without any a priori knowledge about the singularities. The idea is to construct a better discretization with more degrees of freedom based on a given one. Clearly, if
one has the choice to insert new degrees of freedom, one does so in parts of
the domain where the local contribution to the approximation error is large.
For an overview in the case of the Laplace equation see [13].

But how to decide where the error is large and where not? For this we
one can use the theory of residual based error estimators. On each element
$T$ of the discretization we calculate a positive number $\eta_T$ on the basis of the
numerical solution and other known data. This number should be large if the
approximation error is large on an element and small if not. Now we insert
new degrees of freedom in elements which large error estimators, for instance
by subdividing these elements into smaller ones.

![Numerical solution on refined meshes with approx. 220 DOFs, refinement
near singularity with different degree depending on $\lambda$, compare with Fig. 1](image)

**Fig. 4.** Numerical solution on refined meshes with approx. 220 DOFs, refinement
near singularity with different degree depending on $\lambda$, compare with Fig. 1

In Fig. 4 we plot examples of refined meshes together with the according
numerically obtained solution. These solutions correspond to those depicted
in Fig. 1. Note that the smaller $\lambda$ (that is the higher the degree) the more the
mesh is refined near the singularity.

Until 1999 there were no estimators with satisfying theoretical background
for the problem under consideration. We want to point out that the straight-
forward use of according estimators for the Laplace equation, that is for the
case $k = 1$ will fail immediately. First, the estimator for the Laplace equations
does not yield robust estimates and second, used in a refinement strategy, it
would lead to overrefinement at the interface.

During the work on this project, we developed and investigated residual
based a posteriori error estimators [11] which are robust in the case of quasi-
monotone distributions of the permeability. See also the recent independent
works [2], [6] where stronger restrictions on the permeability distribution have
been imposed. The results hold for the two-dimensional case as well as for
the three-dimensional case.

Below, we define the residual error estimator $\eta_R(T)$. Denote with $F \subset \partial T$
all faces of elements $T$ which do not belong to the Dirichlet boundary (in 2D
a face would be simply an edge). Let us denote with $d$ the space dimension.
We define:

$$\eta_R(T)^2 := \frac{h_{T}^{2+d}}{k_T} f_h^2 + \sum_{F \subset \partial T} \frac{h_{F}^{d}}{k_{F_1} + k_{F_2}} \left( k_{F_1} \frac{\partial u_h}{\partial n_{F_1}} + k_{F_2} \frac{\partial u_h}{\partial n_{F_2}} \right)^2.$$ 

Here $k_{F_1}, k_{F_2}$ are the values of the permeability at both sides of the face and $\frac{\partial u_h}{\partial n_{F_1}}, \frac{\partial u_h}{\partial n_{F_2}}$ are the normal derivatives of the numerical solution $u_h$ with respect to both sides of $F$. We denote with $f_h$ a piecewise constant approximation of $f$. In the case that $F$ belongs to the Neumann boundary we insert the according boundary condition.

The estimator consists of two parts. The first part is made of the strong form of the problem, whereas the second part measures the jump of the normal flux through the element boundaries. There is a natural interpretation of the second term. As the solution $u$ of the interface problem by definition has continuous normal derivatives, the nonzero jump term of the discrete solution $u_h$ can be seen as a measure of the error.

We can show [11] that in the quasi-monotone case the error measured in a weighted $H^1$-seminorm is estimated by the error estimator from above

$$\int_{\Omega} k \nabla (u - u_h)^2 \leq c \sum_T \eta_R(T) + \sum_T \frac{h_{T}^2}{k_T} \int_T (f - f_h)^2,$$

where the sum is taken over all simplices. The estimate from below yields

$$\eta_R(T) \leq C \int_{T^+} k \nabla (u - u_h)^2 + \sum_{T \subset T^+} \frac{h_{T}^2}{k_T} \int_T (f - f_h)^2,$$

where the set $T^+$ consists of neighbors of $T$. It is important to note that the constants $c, C$ do not depend on the values of the permeability $k$ or on other data than $u, u_h, f$. In this sense the above bounds are robust. The weighting factors in the definition of $\eta_R(T)$ are necessary in order to assure robustness of the error estimator.

The proof is technical and relies upon robust interpolation results in appropriate Sobolev spaces. It is known that quasi-monotonicity is necessary and sufficient for these interpolation results [7] [14].

The robustness has been confirmed by numerical experiments [11]. We observed that the error is usually overestimated by a factor between [3, 6]. Rescaling by a factor $1/3$ gives an overestimation with a factor between [1, 2].

In the non-quasimonotone case we can show only nonrobust estimates. Numerical results with very strong singularities indicate non-robustness of the error estimator.

### 3.1 Numerical Experiments

The numerical experiments have been carried out using pdelib [8] with its adaptive online grid interface together with the adaptive grid management
kernel of KASKADE 3.x [1]. Geometry definition, problem description and coarse grid meshing have been carried out using FEFLOW [4] and an offline interface to pdelib. Via the online grid interface, the code currently is being incorporated into FEFLOW.

To judge the quality of the error estimators we calculated different numerical examples with known analytical solutions [11]. These examples confirmed robustness and optimal error reduction of the derived estimators. Analytical solutions are known only for simple problems. Now let us come to more real problems. Since here we do not dispose of an analytical solution we calculate a so called reference solution on a very fine grid which we believe is sufficiently close to the error.

Fig. 5. adaptive grid for example 1 with a) 47 b) 145 c) 843 nodes

Fig. 6. Example 1: a) numerical solution and grid refined using $\eta_R(T)$, b) refined grid using Zienkiewicz-Zhu estimator
**Example 1** (Fig. 5): The problem under consideration is known as “hydraulic window”. The domain consists of three layers with a broken middle layer with lower permeability. We set constant inflow Neumann boundary conditions on the upper side of the box and Dirichlet boundary conditions on the right side. On the remaining part we impose zero Neumann boundary conditions. There are at least 10 singular points. We observe that the singular points are not refined equally. For instance the singular points on the left side of the boundary are almost not refined. The major refinement takes place where the middle layer is broken. In Fig. 6 b) we see that the Zienkiewicz-Zhu estimator is not adapted to the case of non constant $k$ since refinement takes place near the interface of the different permeabilities and not only around the singular points. In Fig. 7, we observe an optimal reduction of the error with order $O(N^{-1/2})$ where $N$ is the number of DOFs. This optimal decrease demonstrates that refinement takes place only where needed. Further, the ratio estimated error versus error is moderately constant in the range $[3, 4.5]$.

**Example 2** Here we use “real world” data from the WASY GmbH. The coarse mesh has about 11000 degrees of freedom and the permeability is varying by a factor of 10 is fairly heterogeneous, see Fig. 9.

The solution of the continuous problem has several hundreds of singularities. In Fig. 10 we see that refinement takes place in the vicinity of interior boundaries and in parts of the domain where the gradient is large. The reduction of the estimated error $\eta_R$ takes place with order $O(N^{-1/2})$ (Fig. 8). We know that the reference error $\eta_{ref}$ underestimates the error slightly. The reduction of $\eta_{ref}$ proceeds also with optimal order $O(N^{-1/2})$. Note that in the beginning of the refinement the reduction of the error estimator $\eta_R$ takes place with higher order than implied by the asymptotical order. Further, we see that the efficiency index $\eta_R/\eta_{ref}$ takes values between 2.8 and 4 showing that the global error estimate is reliable and efficient.
4 Conclusions

As seen in the second example, we can make a difference between the pre-asymptotic and asymptotic stages of refinement. The asymptotic stage begins when each simplex is refined at least once, indicating that the error is equally distributed over the whole domain.

One could believe that the error can be made arbitrarily small. But simple calculations show that we can not expect to reduce the error in realistic problems by a factor greater than 1000 in two space dimensions and a factor of 100 in three space dimensions in the asymptotic phase.

To see this note that the optimal (asymptotic) relation between error $e$ and the number of unknowns $N$ is $e \approx N^{-1/2}$. This is seen in the $H^2$ regular case [3]. In real life discretizations the coarse mesh may consist of 1000 DOFs. With the above error reduction rate on a mesh with $100 \cdot 1000^2 = 10^9$ DOFs the error would be about 1000 times less. Such a mesh could be treated by a supercomputer but not by a today's workstation, which is capable of solving problems with $10^6$ DOFs. So it would be realistic to expect that when starting with a coarse mesh with 100 DOFs the error could be reduced by a factor of 100 resulting in a fine mesh with about $10^6$ DOFs. In 3D application the optimal relation is $e \approx N^{-1/3}$ and thus is worse. With a coarse mesh of 1000 unknowns and a reduction of the error by a factor of 100 we stop with mesh with $10^9$ DOFs.

Thus the stopping criterion for real applications will be rather the maximal resources allocated to the solution.

We want to emphasize that the "bad" rate $e \approx N^{-1/2}$ is due to the discretization method and not to the adaptive algorithm. It is not likely that one could improve this rate by finite element methods of higher order, since in realistic applications always singularities occur, which reduce the approximation order of any polynomials [5].
What we can expect from the method in this situation is the optimal usage of the limited resources. Further, we can conclude that the feasible error reduction enables the engineer to obtain a solution with accuracy comparable to that of the measured data of the problem. The estimate based on the asymptotic rate however appears to be too pessimistic, as we observe a faster error improvement during the preasymptotic phase where the insertion of a relatively small number of elements near the (a priori not treated) singularities has a relatively large impact on the error.
References

Modelling and Simulation of a Planned Bio-Chemical in situ Remediation

Willi Jäger¹, Gabriel Wittum¹, Wolfgang Schäfer¹, Christian Wagner¹, and Harald Willershausen²

¹ IWR, INF 368, Universität Heidelberg, 69120 Heidelberg, Germany
² IBL Umwelt- und Biotechnik GmbH, INF 517, 69120 Heidelberg, Germany

Abstract. In this paper, the simulation of a bio-chemical in situ remediation approach is discussed. We consider a field case of a contaminant plume consisting of dissolved chloroethenes. A regional flow model shows, that the contaminants may reach a water works located 2 km downstream of the spill area. Thus, an in situ remediation is anticipated, where the dechlorination and mineralization of the chloroethenes is stimulated by a sequence of aerobic and anaerobic treatments. The equations of the reactive transport model are discretized using implicit time schemes and locally refined meshes. The resulting discrete equations are solved simultaneously with a Newton method and a point-block multigrid method, which turned out to be an efficient solver for the considered systems. Simulation results for several remediation configurations show a fast bio-chemical degradation of the chloroethenes which is however limited to the vicinity of the injection wells. After two years about 60% of the initial contaminant mass is removed, while a pure pump-and-treat scheme removes only 40% of the initial mass in the same time. The simulations suggest that the remediation scheme could be optimized with respect to the in situ biodegradation efficiency.

1 Introduction

The field case under consideration is a subsurface contaminant plume caused by a hydrocarbon spill at a former chemical plant located in the Upper Rhine Valley about 10 km south of Heidelberg. Until 1984 used solvents mainly composed of chlorinated hydrocarbons were reprocessed there. The improper handling of the solvents ended in a severe contamination of the subsurface which was detected first in 1980. After a succesful bioremediation measure [20], today, the spill area is cleaned up, but elevated chloroethene concentrations can be found in downstream sampling points. The current extension of the chloroethene plume is shown in Fig. 1.

1.1 Hydrogeological Situation

The subsurface of the contaminated area is composed of thick quaternary sand and gravel deposits which are subdivided by silt and clay layers. The aquifers under consideration are the so called “oberes Kieslager OKL” (upper gravel beds) and the “mittleres Kieslager MKL” (intermediate gravel beds).
Fig. 1. Extension of the chloroethene plume in 1997. The dots denote water works which might be affected by contaminant spreading.

OKL and MKL are separated by the generally low permeable “obere Zwischenhorizont OZH”. However, in the central part of the model domain the OZH consists of sandy material and allows for a better connection between the two aquifers. In the north-eastern parts of the model domain the OKL is subdivided into two layers by a confining layer called ZH1. The southern boundary for the extension of the ZH1 is only approximately known. It is not clear whether it still exists in the subsurface of the spill area. All data concerning the hydrogeology have been provided by the Geological Survey of Baden Württemberg [11]. A vertical cross section for the aquifer system reaching from west-northwest to east-southeast about 1 km south of the spill area is displayed in Fig. 2.

The average thicknesses and hydraulic conductivities of the different layers are shown in Table 1.

<table>
<thead>
<tr>
<th>hydrogeological unit</th>
<th>average thickness [m]</th>
<th>average hydraulic conductivity [m/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>OKL</td>
<td>40</td>
<td>1 – 2 · 10^{-3}</td>
</tr>
<tr>
<td>MKL</td>
<td>120</td>
<td>5 · 10^{-4}</td>
</tr>
<tr>
<td>OZH</td>
<td>5 – 10</td>
<td>1 · 10^{-7}</td>
</tr>
<tr>
<td>ZH1</td>
<td>10</td>
<td>1 · 10^{-5}</td>
</tr>
<tr>
<td>ZH3</td>
<td>10</td>
<td>1 · 10^{-7}</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of the hydrogeological units in the model domain.
In general the groundwater flows from the eastern flank of the Upper Rhine Graben sediments to the west and finally discharges into the Rhine river (outside the model domain). However, the aquifer system under consideration is extensively used for drinking water supply and the groundwater extractions at the water works have an appreciable effect on the regional flow field.

1.2 Concept of the in situ Bioremediation

The IBL Umwelt- und Biotechnik GmbH is preparing the clean-up of the central part of the contaminant plume which shows concentrations of up to 3500 µg/l of chloroethenes. In principal it is intended to apply the same scheme that was used for the remediation of the source area [20]. Contaminated water will be extracted with two pumping wells fully screened in the OKL which shows saturated thicknesses between 35 m and 42 m there. The extraction rates will be 2.8 l/s and 1.4 l/s respectively. The extracted water will be completely re-infiltrated. 1.4 l/s will be charged to an infiltrating well outside the contaminated area while the remaining 2.8 l/s will be equally distributed over 5 infiltration wells located inside the plume area. This partitioning is performed to guarantee that the water infiltrated in the contaminated area will be completely captured by the two pumping wells. The two infiltration wells which are closest to the pumping wells will be screened over the whole aquifer thickness, while the three more distal wells are screened over the top 10 m–15 m of the OKL only. The extracted water will be treated in an above ground water purification plant prior to infiltration. That part of the infiltration water charged to the wells inside the contaminated area will additionally be amended with substances supposed to enhance in situ

Fig. 2. Vertical cross section of the aquifer close to the spill area
biodegradation of the chloroethenes. During the anaerobic phase an easily degradable dissolved substrate is provided to allow for the growth of a fermentative and methanogenic bacterial consortium. The anaerobic phase will allow for a reductive dechlorination of the chloroethenes. During the aerobic phase the substrate is replaced by hydrogen peroxide which will serve as an oxygen source for aerobic bacteria which primarily will oxidize the methane produced during the anaerobic phase. The lower chlorinated chloroethenes will be mineralized in this phase by a co-metabolic pathway. Details on the different reaction steps will be discussed in Sect. 2.4 and [19].

2 Reactive Transport Model

As a first step, the groundwater flow was computed for the region indicated by the frame in Fig. 1. The objectives of this regional flow model were twofold: First we examine whether the contaminants will possibly reach one of the adjacent water works in the future and second the regional flow model should allow to embed the smaller scale reactive transport model into the regional flow system, i.e. the regional model should provide meaningful boundary values for the small scale model.

The boundary conditions and the water extraction rates of the water works were interpolated from the exceptional large databases of the Landesanstalt für Umweltschutz Baden-Württemberg and the Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg [10,1] for the Upper Rhine Graben.

The regional flow model shows that the contaminants present in the aquifer today will finally reach the wells of a water works situated about 2 km downstream of the spill area. However, assuming an effective porosity of 20% it will take the contaminants more than 30 years to reach these wells which are filtered in the upper part of the MKL.

2.1 Reactive Transport Model Domain

As mentioned before, the intended in situ remediation measure for the cleanup of the site consists of 2 pumping wells and 6 infiltration wells. These wells cover an area of about 150 m × 150 m. Figure 3 shows the piezometric heads for a part of the regional flow model domain. The reactive transport model domain will be restricted to the OKL as the contaminants have not yet reached deeper layers in that area. The calculated piezometric heads and the data on aquifer geometry and hydrogeological parameters were cut from a refined area of the regional flow model and were transferred to the small scale reactive transport model indicated by the frame in Fig. 3.

The reactive transport model has been implemented in the UG toolbox [3]. Figure 4 shows the domain for the reactive transport model. The blue part represents the layer (ZH1) with a small hydraulic conductivity.
Fig. 3. Part of the regional flow model domain. The light blue rectangular frame denotes the area of the small scale reactive transport model. Its dimensions are 280 m × 280 m. The red lines are isolines for piezometric heads of the OKL. Values are given in m asl. The red and blue rectangles depict the positions of the two pumping and 6 infiltration wells.

Fig. 4. Domain for the reactive transport model. The low conductivity layer ZH1 is shown in blue.

2.2 Initial Conditions for Transport Modelling

Four different chloroethenes will be considered in the reactive model. Three of them, namely tetrachloroethene (PCE), trichloroethene (TCE) and 1,2-cis-dichloroethene (DCE) make up the major part of the contaminant mixture. The concentrations of the fourth one, vinyl chloride (VC), were usually below detection limit in the aquifer (1 μg/l), but it is considered in the model because it might eventually be produced during the reductive steps of the bioremediation.

An important characteristic of the plume is the non-uniform vertical distribution of the chloroethenes. The maximum concentrations were found in the multilevel well B392 with 3545 μg/l total chloroethenes at a depth of 20 m below water table. The quantitatively dominating compound was 1,2-cis-dichloroethene, a typical product of microbially mediated dechlorination of PCE and TCE.
The initial concentrations for the reactive transport model have been interpolated from 35 sampling points including 11 multilevel wells of the measured local data with the so-called Kriging-method of the interpolation software SURFER from Golden Software Inc.. The data were recorded and provided by the IBL Umwelt- und Biotechnik GmbH.

With the same technique the initial oxygen concentration has been estimated. Substrate and methane were initially not present in the aquifer. Very low initial bacteria concentrations were provided in the whole model domain to allow for microbial growth. The initial concentrations of PCE, TCE, DCE and oxygen are shown for a cross-section in the upper part of the model domain in Fig. 5.

2.3 Transport Equation

The transport of a dissolved substance $X$ is governed by advection, dispersion/diffusion and adsorption which leads to the transport equation

$$R_X \frac{\partial c_X}{\partial t} - \nabla \cdot (D \nabla c_X - v c_X) + r = 0,$$

where $c_X$, $R_X$, $D$, $v$ denotes the concentration of the dissolved component $X$, the retardation factor, the dispersion tensor and the pore velocity. $r$ represents the influence of wells and sinks for instance caused by chemical transformation. For a more detailed description, we refer to text books like [4,9,15,16]. The used hydrogeological parameter can be found in [19].

2.4 Biodegradation of Chloroethenes

Two ways for the biodegradation of the chloroethenes tetrachloroethen (PCE), trichloroethen (TCE), cis-1,2-dichloroethen (DCE) and vinyl chloride (VC) (see Fig. 6) are known. The first possibility is the reductive dechlorination under anaerobic condition. The second option is the co-metabolic degradation under aerobic condition. A summary of the degradation of halogenated organic compounds can be found in [14].
Reductive dechlorination The dechlorination of PCE leads to TCE, the dechlorination of TCE to DCE and the dechlorination of DCE to VC. A reductive dechlorination of VC does often not occur. Thus, the final products are usually DCE or VC. The dechlorination of PCE, TCE and DCE can be initiated by the anaerobic degradation of a number of substances (see e.g. [7]).

Our model for the reductive dechlorination is based on the following assumptions (see [6,19]).

- The degradation starts with PCE, continues with TCE and DCE and ends with VC.
- PCE, TCE and DCE compete for H₂. PCE is preferably dechlorinated, then TCE and DCE.
- Bacteria $B_1$ consume a substrate $S_1$ under anaerobic conditions and produce H₂ and methane (CH₄). H₂ is then used for the reductive dechlorination.
- We do not explicitly take into account that the dechlorination is actually catalyzed by different organisms and not by the bacteria $B_1$. However, since both organisms are closely related, the simplification is justified.

Co-metabolic degradation Co-metabolic processes are supported by microbial activity although the bacteria do not gain energy from the degradation process. The bacteria produce enzymes for the consumption of primary substrates which are able to oxidize TCE, DCE and VC ("fortuitous degradation") as well. An aerobic degradation of PCE has not been observed yet. [5,12].

Our model for the co-metabolic degradation uses the following assumptions [19].

- TCE, DCE and VC are completely mineralized (decomposed into anorganic compounds).
- TCE, DCE and VC compete for the enzymes. VC is preferably degraded, then DCE and TCE. An aerobic degradation of PCE does not occur.
- The methane which is produced by the anaerobic consumption of the substrate $S_1$ is used as primary substrate $S_2$.
- Bacteria $B_2$ consume the substrate $S_2$ under aerobic conditions. The bacteria $B_1$ and $B_2$ are different species.
- The growth of the bacteria population $B_2$ causes the oxidation of a certain amount of TCE, DCE and VC.
- A possible competitive inhibition of substrate degradation by TCE, DCE or VC is not taken into account.

According to the model assumption, 9 species are necessary for a mathematical description of the model: The anaerobic bacteria $B_1$, the aerobic bacteria $B_2$, the substrate $S_1$, methane $S_2$, oxygen $O$, PCE, TCE, DCE, and VC. For the bio-chemical model equations we refer to [19].

## 3 Solution Process

The solution of the coupled system of transport reaction equations

$$ R_X \frac{\partial c_X}{\partial t} = \nabla \cdot (D(v) \nabla c_X - v c_X) + \text{Reac}_X(\ldots) + q_X, \quad \text{and BC,} $$

with $X \in \{S_1, S_2, O, \text{PCE, TCE, DCE, VC}\}$ and

$$ \frac{\partial c_B}{\partial t} = \text{Reac}_B c_B - \mu_{dec} c_B, $$

$B \in \{B_1, B_2\}$ is discussed in this section. Reac$_X$, Reac$_B$ stands for the reaction terms described in the previous section.

### 3.1 Discretization

We discretize the time derivative first using implicit Runge-Kutta schemes. For instance, the implicit Euler scheme leads to the following system of equations

$$ \frac{c_X(t) - c_X(t-\Delta t)}{\Delta t} = \nabla \cdot (D(v) \nabla c_X(t) - v c_X(t)) + \text{Reac}_X(\ldots(t)) + q_X $$

and

$$ \frac{c_B(t) - c_B(t-\Delta t)}{\Delta t} = \text{Reac}_B(\ldots(t-\Delta t)) c_B(t-\Delta t) - \mu_{dec} c_B(t). $$

Note that the term describing the exponential grows of the bacteria is explicitly discretized. While implicit methods are more stable for differential equations describing an exponential decay, explicit methods are more stable for differential equations describing an exponential growth. This can easily be seen from the stability function. The size of the time step $\Delta t$ is controlled by the Newton method. Slow convergence of the Newton method leads to a reduction of the time step. The maximal value for $\Delta t$ is 3 days.

The spatial derivatives are discretized using the finite volume method (see e.g. [8]) with a first order upwind scheme for the approximation of the
advection. In order to guarantee mass conservation, the transport reaction system (2)–(3) and the flow equations must be solved on the same grid.

The grid is adapted to the solution in each time step using local refinement and de-refinement. The grid adaption is performed by a dynamic nested iteration which is introduced in [17]. A typical grid with about 44000 elements is shown Fig. 7. The regions in the vicinity of the wells are strongly refined.

3.2 Solution of the Discretized Equations

The nonlinear systems of equations are reduced to linear systems of equations using a Newton method with a line search technique following the paper of Bank and Rose [2]. In each step of the Newton method, a large linear system (about 200000 unknowns) representing the Jacobian of the nonlinear system has to be solved. This linear system includes all transport-reaction equations for the involved species.

Since the performance of multigrid methods is in contrast to classic solvers independent of the number of unknowns, standard and algebraic multilevel methods have been developed for these systems of transport-reaction equations. This has been done using a point-block block. Both methods turned out to efficient solvers for these linear systems [18].

4 Simulation Results

We consider simulation results for two different cases. The simulation time was 2 years.

Case 1 We assume an idealized homogeneous distribution of the hydraulic conductivity for the contaminated area. All wells (two pumping wells and six infiltration wells) are operated as described in Sect. 1.2. The in situ remediation starts with an anaerobic phase, i.e. with the injection of substrate $S_1$ (20 mg/l). After 50 days of anaerobic operation a 50 days period of aerobic
treatment (i.e. injection of hydrogen peroxide (120 mg/l)) is installed. After another 50 days the aerobic mode is replaced by the anaerobic mode an so on. This procedure leads to 13 switches between anaerobic and aerobic treatments during the 2 years remediation period. The bio-chemical production and degradation of the chloroethenes is displayed in Fig. 8.

Figure 9 shows the temporal decrease of the chloroethene mass in the model domain. Most of the reduction of the chloroethene mass is due to classical pump-and-treat removal, especially for the period after 250 days. Table 2 shows the different mass balance terms for case 1.

Table 2 reveals that in situ degradation is an important but not the dominant removal process in the model. The major part of the contaminant mass is either still remaining in the model domain after 2 years or it is extracted by the two wells. As the adsorption coefficient in our model is larger for PCE
Table 2. Mass balance for the contaminants in case 1

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Mass</th>
<th>Degradation (−) or Production (+)</th>
<th>Extraction by wells</th>
<th>Outflow at model boundaries</th>
<th>mass after 2 years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kg</td>
<td>%</td>
<td>kg</td>
<td>%</td>
<td>kg %</td>
</tr>
<tr>
<td>PCE</td>
<td>18.4</td>
<td>100</td>
<td>−2.6</td>
<td>−14.1</td>
<td>−3.2</td>
</tr>
<tr>
<td>TCE</td>
<td>10.4</td>
<td>100</td>
<td>+0.8</td>
<td>+7.7</td>
<td>−4.1</td>
</tr>
<tr>
<td>DCE</td>
<td>92.9</td>
<td>100</td>
<td>−4.5</td>
<td>−4.8</td>
<td>−45.4</td>
</tr>
<tr>
<td>VC</td>
<td>0</td>
<td>+0.9</td>
<td>−0.7</td>
<td>0</td>
<td>−10.8</td>
</tr>
<tr>
<td>Total chloroethenes</td>
<td>121.7</td>
<td>100</td>
<td>−5.4</td>
<td>−4.5</td>
<td>−53.4</td>
</tr>
</tbody>
</table>

than for the other chloroethenes there is relatively more PCE left at the end of the period compared to TCE and DCE. VC production is generally low and all VC is captured by the wells. It has to be noted here that VC production is the consequence of our specific model assumptions. It is possible that no VC at all will be produced in reality.

Throughout the 2 years a total of 5.4 kg chloroethenes were degraded in situ. At the same time 1748 kg of substrate and 6048 kg of oxygen were injected. The major part of the substrate was transformed to methane which in turn was predominantly oxidized by the injected oxygen. The weight ratio of chloroethene degradation to substrate injection is as low as 0.003, and that of chloroethene degradation to oxygen injection is even as low as 0.0009.

The low efficiency of the method is the result of the fact that the injected substrate is always decomposed in the immediate vicinity of the wells. At the beginning of the remediation the chloroethenes around the injection wells are subject to microbial degradation. However, after two or three cycles the aquifer in the neighborhood of the wells is cleaned up, while the substrate is still consumed there and can not spread out into the aquifer.

In consequence, substrate degradation and methane production and therefore the biological activity is no longer linked to the presence of chloroethenes. In other words, substrate degradation and methane production occur at the “wrong” location in the model and thus do no longer support chloroethene degradation in the aquifer. A typical situation is shown in Fig. 10. The vicinity of the injections wells has been cleaned up, but high bacteria and substrate concentrations are restricted to the neighborhood of the injection wells.

Case 2 In case 2 the duration of the two modes is increased from 50 days to 100 days. Additionally, the initial anaerobic/aerobic period (200 days) was followed by a 300 days period where only the extraction wells and the
injection well outside the contaminated area were used. The other injection wells were then re-activated for the following 200 days and again turned off for the remaining 30 days. Although the injection wells were operated for 400 days only compared to 730 days in case 1, the total chloroethene degradation could be increased from 6.6 kg to 7.3 kg. This is caused by the fact that contaminated water reached the locations of the injection wells during downtime from upstream parts of the aquifer and that those contaminants were subject to degradation in the successive injection period. In the standard case the permanent injection forced the contaminated upstream water to pass the injection wells. Total chloroethene removal was also increased in case 2 from 58.8 kg to 66 kg.

**Fig. 10.** Concentrations in the upper part of the aquifer. Maximum values are red, minimum values are blue.
<table>
<thead>
<tr>
<th>Case 1 (50 days cycles)</th>
<th>Total removal [kg]</th>
<th>Extraction</th>
<th>Degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58.8</td>
<td>91%</td>
<td>9%</td>
</tr>
<tr>
<td>Case 2 (intermittent operation)</td>
<td>66</td>
<td>89%</td>
<td>11%</td>
</tr>
</tbody>
</table>

**Table 3.** Summary of the total chloroethene removal and the relative contributions of extraction and degradation

Table 3 shows the simulated removal of chloroethenes for the 2 years of simulation and the relative contributions of extraction and degradation for all ten cases. Please note that the initial chloroethene mass was 121.7 kg.

## 5 Summary of Results and Conclusions

The successful combined application of the standard groundwater flow model MODFLOW [13] and a newly developed reactive transport model implemented in the UG code [3] on a field case of chloroethene transport allowed to predict the trends in the performance of a planned in situ bioremediation.

The point-block multigrid method proved to be a very efficient tool to solve the coupled equation system for multi-species transport and biochemical reactions. The employed one-step or direct solution approach admitted relatively large time steps which were only limited by the convergence of the Newton method. The simulation time for one configuration on a Pentium III (650 MHz) was about one day.

The results from the simulations suggest that in situ biodegradation will take place in the aquifer, however its efficiency could be improved. The main reason for the limited efficiency of the in situ degradation is that after a certain period of well operation the surrounding of the wells is cleaned up, while the chloroethenes in the remainder of the aquifer are outside the constant sphere of bacterial action stimulated by substrate injection.

A possible way to strengthen the role of in situ degradation for total chloroethene removal would be to operate the injection wells intermittently (cf. case 2). In practice, the operation mode of the wells could be adapted to the specific requirements of the field case by continuously monitoring the contaminant concentrations.

Usually about 60% of the initial contaminant mass present in the aquifer was removed after 2 years of operation. This has to be compared to a pure pump-and-treat scheme, where the removal rate was decreased to about 40% in 2 years (case 8).

The simulations show that a minor amount of vinyl chloride might be produced during chloroethene dechlorination. However, this vinyl chloride will be nearly quantitatively removed at the extraction wells and therefore poses no threat to downstream groundwater quality.
As a general conclusion we recommend to optimise the planned remediation measure with respect to in situ degradation efficiency by modifying the actual well configuration and operation scheme.

References


Influence of Surfactants on Spreading of Contaminants and Soil Remediation

P. Knabner, S. Bitterlich, R. Iza Teran, A. Prechtel, and E. Schneid

Institute for Applied Mathematics, University of Erlangen-Nuremberg
Martensstrasse 3, 91058 Erlangen, Germany
E-mail: knabner@am.uni-erlangen.de, URL: http://www.am.uni-erlangen.de/aml

Abstract. Surfactants occur already in undisturbed biological processes in soils but within the development of remediation techniques these substances are of large interest because of their interaction with hydrophobic substances. The standard models for (un-)saturated water flow and solute transport are extended to include the influence of the surfactant transport on the flow of the water phase. Two effects of surfactants on (un-)saturated water flow are included: the modification of the interfacial tension between water and air and the swelling of clay minerals due to sorption of surfactants. Simulations are presented, which exhibit the feedback of surfactant transport on water flow and content. An identification algorithm for hydraulic soil properties has been developed that supplements the simulation tool.

1 Introduction

The numerical modelling of water flow and solute transport has become a standard tool to complement field and laboratory studies in evaluating and forecasting the fate of pollutants. Furthermore advanced simulation is useful to predict the success of site remediation strategies like the treatment of a polluted soil with surfactants (SEAR: surfactant enhanced aquifer remediation).

In the last decade, the modelling of the influence of surfactants on groundwater flow and transport has considerably advanced and can now be incorporated in flow and transport settings. The presence of surfactants in the pore water has an impact on the flow regime and vice versa. Surface active agents enhance aquifer remediation by increasing the solubility of organic substances within the aqueous phase and the mobility of non aqueous phase liquids. On the other hand the effective permeability of the porous medium may vary as a consequence of the aggregation of the surfactant onto the clay fraction of the soil. These aspects have to be taken into account when the risk of groundwater pollution is evaluated and the fate of contaminants is predicted.

To perform practical simulations along with laboratory experiments the identification of hydraulic parameters is essential. In order to describe the water flow in an appropriate manner, the parameters of the water retention curve and the unsaturated hydraulic conductivity have to be specified.
2 Coupled Water–Surfactant Transport

2.1 Model

To describe the migration of a surfactant in a porous medium we have to establish a description of the flow regime as well as a transport model which should include advection, dispersion, diffusion and sorption. Standard formulations exist in form of partial differential equations obeying the principle of conservation of mass. This leads us first to the Richards equation as a model for flow in the saturated and vadose zone

$$\frac{\partial}{\partial t} \Theta(\psi) + \nabla \cdot q = 0 \quad q = -K_s k_r(\psi) \nabla(\psi + z)$$  \hspace{1em} (1)

with parametrizations of the soil water characteristic (water retention curve) and of the unsaturated hydraulic conductivity. Here $t$ denotes the time, $\Theta$ the volumetric water content, $\psi$ the pressure head, $q$ the Darcy flux, $K_s$ and $k_r$ are the saturated and relative hydraulic conductivity and $z$ is the elevation head. The solute transport is described by the following partial differential equation:

$$\frac{\partial}{\partial t} (\Theta c) + \rho_b \frac{\partial}{\partial t} (\phi(c)) + \nabla \cdot w = 0 \quad w = -D \nabla c + qc$$  \hspace{1em} (2)

where $c$ is the dissolved concentration in $x$ at time $t$, $\rho_b$ the bulk density, $\phi$ the equilibrium sorption isotherm, $w$ the mass flux and $D$ the diffusion–dispersion tensor.

These standard equations are now coupled in the following way to take the interaction of fluid flow and surfactant transport into account. According to the work of [5] the surface active agent decreases the surface tension of the fluid with increasing concentrations and thus affects the capillary pressure of an unsaturated soil. Consequently a scaling factor is introduced in the pressure-saturation relation:

$$\Theta(\psi) \rightarrow \Theta\left(\frac{\sigma_0}{\sigma} \psi\right) =: \theta(\psi, c) \quad \frac{\sigma_0}{\sigma} = \frac{1}{1 - b \ln(c/a + 1)}$$  \hspace{1em} (3)

with empirical, surfactant dependent parameters $a$ and $b$. In a recent study [6], this model has proven its applicability in accordance with experimental laboratory results. The variability of the hydraulic conductivity due to the sorption of the surfactant is incorporated in the model according to [4]:

$$K_s \rightarrow K_{\text{eff}}(c) = K_{\text{coarse}}^{1-\nu} K_{\text{clay}}^\nu \quad \nu = \nu_{\text{clay}} + \frac{\rho_b}{\rho_s} \phi(c)$$  \hspace{1em} (4)

$K_{\text{eff}}(c)k_r(\psi) =: K(\psi, c)$

$K_{\text{eff}}$ represents the effective hydraulic conductivity, $K_{\text{coarse}}$ and $K_{\text{clay}}$ the saturated hydraulic conductivities of the coarse respectively the clay fraction.
of the soil, $\nu_{\text{clay}}$ is the volume fraction of the clay component and $\rho_b$ and $\rho_s$ are the densities of the bulk soil and the surfactant, respectively. Surfactants sorb preferentially to the clay and therefore increase the volume fraction of the clay-surfactant conglomerate, therefore the effective permeability decreases due to the sorption of the surfactant. These mechanisms couple the surfactant transport to the water flow. On the other hand, the actual water content and Darcy flux in the transport equation are determined by the Richards equation. As the surfactants have an influence on the flow regime, they do also affect the fate of a contaminant which is transported in the subsurface. The upstanding mutually coupled system of partial differential equations along with its defining coefficient functions has to be discretized to be solved on a computer.

2.2 Discretization

The transport equations (1) and (2) are written in a mixed variational formulation. The time evolution is discretized by the backward Euler scheme. The spatial discretization is introduced by lowest order mixed finite elements on triangles. The water pressure $\psi$ and the surfactant concentration $c$ are approximated in $P^0(T)$, i.e. piecewise constant on elements. The fluxes of water $q$ and of surfactant $w$ are approximated in $RT^0(T)$, i.e. piecewise linear on elements, having continuous normal components over the edges (in the sequel abbreviated by “continuity of fluxes”). To circumvent the resulting saddlepoint problem, we introduce hybridization of the flux ansatz space. We enlarge the discrete space of fluxes to the product space $\prod_{T \in T} RT^0(T)$, neglecting continuity of fluxes. Additional Lagrange-multipliers on edges ($\lambda$, $\mu$) and a new constraint equation are introduced to re-establish the continuity of fluxes. After this extension of the problem formulation and its variables the system is twice reduced by condensation. First the flux variables are explicitly eliminated. The resulting system of equations that describes the evolution of one timestep is then

$$\theta(\psi_T, c_T) + \frac{\tau}{|T|} K(\psi_T, c_T) \sum_{E,E' \subset T} B_{T,EE'}^{-1}(\psi_T - \lambda_{E'}) = \theta(\psi_T^{\text{old}})$$  \hspace{1cm} (5)

$$\sum_{T \ni E} \sum_{E' \subset T} K(\psi_T, c_T) B_{T,EE'}^{-1}(\psi_T - \lambda_{E'} - z_{TE'}) = 0$$  \hspace{1cm} (6)

$$\theta(\psi_T, c_T) c_T + \rho_b \phi(c_T) + \frac{\tau}{|T|} \sum_{E,E' \subset T} A_{T,EE'}^{-1}(c_T - \mu_{E'}) +$$
$$+ c_T \frac{\tau}{|T|} K(\psi_T, c_T) \sum_{E,E' \subset T} B_{T,EE'}^{-1}(\psi_T - \lambda_E) = \theta(\psi_T^{\text{old}}) c_T^{\text{old}} +$$
$$+ \rho_b \phi(c_T^{\text{old}})$$  \hspace{1cm} (7)
\[
\sum_{\substack{T \supset E \\
E' \subset T}} \left( c_T K(\psi_T, c_T) B_{T,EE'}^{-1}(\psi_T - \lambda_{E'} - z_{TE'}) + A_{T,EE'}^{-1}(c_T - \mu_{E'}) \right) = 0
\]

where \(\psi_T, c_T\) are pressure and concentration values on the elements (the degrees of freedom of the discrete ansatz space) and \(\lambda_E, \mu_E\) are the Lagrange-multipliers on edges. \(B_T\) is the element matrix, composed of the \(L^2\)-scalar products of the basis functions from the \(RT^0\) Ansatz space and connecting the edges on that element. \(A_T\) is the same matrix but including the inverse of the diffusion–dispersion tensor \(D\), which itself depends on the Darcy flux \(\mathbf{q}\). All variables but \(\psi_T^{old}\) and \(c_T^{old}\) correspond to the new time level. The second step of variable elimination is done in an implicit way. The elementwise equations (5) and (7) represent the discrete evolution of element values (\(\psi_T, c_T\)) and involve only degrees of freedom from one element. These equations together define an implicit expression of the element values for given Lagrange-multipliers (\(\lambda_E, \mu_E\) for \(E \subset T\)). In this treatment the dependence of \(A_T\) on \(D^{-1}(\mathbf{q})\) and thus on \(\psi_T, c_T\) and \(\lambda_E\) for \(E \subset T\) because of

\[
q_{T,E} = \sum_{\substack{E' \subset T}} K(\psi_T, c_T) B_{T,EE'}^{-1}(\psi_T - \lambda_{E'} - z_{TE'})
\]

is not stated explicitly. Within the iteration process to be used for the solution of the evolving system (see Sect. 2.3) the Jacobians do not include this dependence, but the evaluation of residuals considers the correct form of \(A_T\). The solution of equations (5)–(8) requires only evaluation of these equations and of their derivatives for given variables. Therefore the implicit definition of \(\psi_T\) and \(c_T\) can formally be included in the edgewise equations (6) and (8), describing the continuity of flux across an edge. Finally (6) and (8) represent the global system of equations, that has to be solved for the Lagrange-multipliers in each time step.

2.3 Algorithm

A damped version of Newton’s Method (Armijo’s rule) is used to solve the local ((5), (7)) and the global ((6), (8)) nonlinear equations. Due to the treatment of the \(\mathbf{q}\)-dependence of the dispersion tensor, actually this is a modified Newton’s method, ignoring the corresponding terms in the Jacobian. Within each iteration of the global solution process, the evaluation of (6) and (8) and of their derivatives with respect to the Lagrange-multipliers is preceded by solution of the local problems (5) and (7).

The global system of linear equations is solved by a multigrid method. The system matrix is the Jacobian of (6), (8), the right hand side is the residual of this set of equations and the unknowns are the updates of the Lagrange-multipliers on edges within the Newton iteration. Based on the equivalence of nonconforming and mixed finite elements [1] the multigrid method is built from grid transfer operators, derived for the Crouzeix-Raviart element [2].
The coarse grid matrices are defined by Galerkin approximation. The linear problem on the base level of the grid hierarchy is solved by LU decomposition. On fine grid levels smoothers like Gauss-Seidel or ILU are used.

### 2.4 Simulation Results

The introduced model is used to demonstrate the effects of varying surface tension and permeability in a two dimensional simulation with an artificial data set. All the results are compared to the corresponding flow and transport scenario with constant surface tension and without permeability reduction (labelled ‘Reference’ in Fig. 1 and Fig. 2).

The Richards equation is complemented by exponential parametrizations of the water retention curve and the unsaturated hydraulic conductivity. We impose a stationary unsaturated flow regime from the northern to the south-

![Fig. 1. Water–surfactant transport with (middle and right column) and without (left column) permeability reduction: contour plots of surfactant concentration and water content for successive timepoints $t \in \{5, 10, 30\}$]
ern boundary of a square domain by defining the flux and the corresponding initial pressure profile. Eastern and western boundaries are impermeable (no flow boundaries). Thus, when assuming no impact of the surfactant transport on the water flow, the water content $\theta = 0.295$ (color-coded in Fig. 1 and Fig. 2 with green), the pressure and the flux remain constant throughout the entire time interval.

For the transport of the surfactant, we included the effects of advection, diffusion, longitudinal and transversal dispersion as well as linear equilibrium sorption. The domain is initially void of surfactant. The substance is injected at the western half of the northern boundary within the time interval $[0, 6]$, after which no more surfactant enters the area. At the southern boundary a free outlet is given (homogeneous Neumann condition). See the migration of the contaminant plume at different time steps on the left hand side of Fig. 1 and Fig. 2.

![Fig. 2. Water–surfactant transport with (middle and right column) and without (left column) pressure scaling: contour plots of surfactant concentration and water content for successive timepoints $t \in \{5, 10, 30\}$](image-url)
In a first step we disregard pressure scaling and respect only the permeability reduction through the sorptive capacity of the surfactant as proposed in model equation (4). As the sorbed mass of surfactant per mass of soil $\phi(c)$ is increasing, the effective hydraulic conductivity $K_{\text{eff}}$ will be dominated by the low conductivity of the clay fraction $K_{\text{clay}}$ and thus decrease. As we impose the same water flux at the boundaries as in our reference case, this leads to a retarded water transport in the area of low permeability, thus an increase in the water content (see right column of Fig. 1). As water flow is hindered by the surfactant plume, water content in front of the plume now decreases. The permeability reduction induces the retardation of the surfactant migration, as can be seen in the middle column of Fig. 1.

In a second simulation, we disregard the permeability reduction, but enable the surface active agents to change the surface tension and thus affect the hydraulic pressure head as described by model equation (3). Increasing surfactant concentration $c$ implies lower surface tensions and consequently higher pressure heads. As a consequence of the generated hydraulic gradient, water content decreases (right column of Fig. 2) with the solute front. If surface tension is assumed constant, increases in pressure head indicate increased water contents, i.e. a wetting front. But as demonstrated in a recent experimental study by [6] pressure heads may increase substantially while the water content decreases in the presence of surface active agents. This observation can be clearly reproduced by this two dimensional numerical simulation. Another remarkable effect is the change of the water flow in the eastern half of the domain. As water content decreases where the surfactant concentration is high, it increases in the eastern half, i.e. the flow is deviated towards the east. The constant water content in the reference case is $\theta = 0.295$. This explains the extended spreading of the surfactant plume (middle column of Fig. 2) in the transversal direction. Regarding the travel times, the surfactant migration is not retarded in this case.

A subsequent simulation of contaminant transport strongly depends on the water content and flux distributions that result from coupled water-surfactant transport. For an inert contaminant, i.e. neglecting sorption effects, the migration velocity is given by $q/\Theta$. Not only the direction of the flux is altered but also the modification of water content will change this migration velocity. In addition the micellar pseudophase occurring at surfactant concentrations above a critical value acts as a solubilizer for organic solutes. Taking this solubilization into account, e.g. using the model of carrier facilitation [3], is a next step for the model development. Further influences of surfactants on spreading of contaminants in soils are related to the modification of interfacial properties that govern sorption processes.

3 Identification of Hydraulic Properties

The numerical simulation of (un-)saturated water flow critically depends on the knowledge of the hydraulic soil properties, expressed by the characteris-
tics of water retention $\Theta$ and unsaturated hydraulic conductivity $K$. These hydraulic properties can be identified by measurements obtained from soil column outflow experiments.

The design of a suitable experiment for this purpose is as follows. A vertically oriented soil column with a known initial pressure head equilibrium distribution ($\psi \geq 0, q = 0$) is drained by slowly decreasing the pressure head at the outlet (lower boundary, $x = 0$) and preventing any flow at the top of the column (upper boundary, $x = L$). Mathematically this is modelled by a Dirichlet boundary condition $\psi(0, t) = h(t)$ at the lower boundary and a homogeneous flux boundary condition ($q(L, t) = 0$) at the upper boundary. The experimental setup is complemented by two measurements, $f$ and $g$, over time $t$: The cumulative outflow $f(t) := \int_0^t q(0, \tau) \, d\tau$ is measured at the outlet of the soil column and the pressure head $g(t) := \psi(L, t)$ is measured at the top of the column.

We consider the model of the Richards equation in pressure head form (1) in one spatial dimension. Solving this model equation for given hydraulic functions $\Theta$ and $K$ and assigning the measurements $f(t)$ and $g(t)$ to the hydraulic functions $\Theta$ and $K$ characterizes the direct problem (DP: $(\Theta, K) \mapsto (f, g)$). The inverse problem (IP: $(f, g) \mapsto (\Theta, K)$) consists of determining the hydraulic functions $\Theta$ and $K$ from given measurements $f$ and $g$.

### 3.1 Identifiability and Stability

If the applied suction $h(t)$ at the lower boundary is assumed to be smooth and monotone decreasing in time, the mapping DP is injective for sufficiently smooth hydraulic functions $\Theta$ and $K$. Therefore it is meaningful to assume that the inverse problem IP is uniquely solvable. Then the hydraulic functions are identifiable from outflow experiments.

The inverse problem is stabilized by parametrizing the hydraulic functions. We are looking for an unbiased parametrization, which does not take any a-priori shape information like the van Genuchten-Mualem parametrization into account. A general approach uses splines to parametrize these functions. In this way we obtain a piecewise linear or a piecewise quadratic parametrization. The unknown hydraulic functions are defined by parameter vectors $p^\Theta$ and $p^K$ of dimension $r$. A special aspect of such parametrizations is the fact that the water retention and the hydraulic conductivity are not coupled like in the van Genuchten-Mualem model. The low smoothness of piecewise linear functions applied for $\Theta$ and $K$ leads to low smoothness of the observations $f(t)$ and $g(t)$, corresponding to the measurements. Therefore an approach with quadratic B-splines is profitable. Because of the ill-posedness of the inverse problem it exists a threshold $r_{max}$ depending on the error of the discretization of the direct problem, measurement errors and the type of the parametrization, such that a parametrization with more than $r_{max}$ degrees of freedom leads to instabilities.
3.2 Numerical Identification by Output Least Squares

The solution of the inverse problem is based on the minimization of an error functional

$$
\sum_{i=1}^{M} \alpha_i \left( f(\hat{t}^i) - \hat{f}^i \right)^2 + \sum_{i=1}^{N} \beta_i \left( g(\hat{t}^i) - \hat{g}^i \right)^2
$$

with positive weighting factors $\alpha_i$, $\beta_i$ and measured values $\hat{f}^i$ and $\hat{g}^i$ at time points $\hat{t}^i$ and $\hat{t}^i$, respectively. Efficient optimization algorithms (like quasi-Newton methods) need the value of the error functional and its gradient in every step of the optimization. The computation of the Hessian is not necessary. The gradient of the error functional can be evaluated in two ways.

- Finite Difference Method: We have to solve the direct problem $2r$ times for the one-sided and $4r$ times for the central difference ratio to compute an approximation of the gradient.
- Direct Method: Differentiation of the direct problem with respect to the degrees of freedom leads to a system of linear equations. The combined solution process (direct problem and gradient) needs about twice the CPU time of the solution of the direct problem DP.

An appropriate choice of initial values accelerates the minimization of the error functional and avoids an early abort of the optimization procedure. We obtain appropriate start values for our parametrization of the hydraulic properties $\Theta_n$ and $K_n$ with $r_n$ degrees of freedom by interpolating the functions $\Theta_m$ and $K_m$, which minimize the error functional for $r_m$ degrees of freedom ($r_m < r_n$). The values of the hydraulic functions $\Theta$ and $K$ at saturation ($\psi \geq 0$) are often known from independent experiments and can be used to determine initial values for the parametrization with the least possible number of degrees of freedom.

![Hydraulic properties](image)

**Fig. 3.** Hydraulic properties (dashed) and identified hydraulic properties (piecewise quadratic) for 9 degrees of freedom (solid)
Fig. 4. “Measurements”: original-undisturbed (dashed), disturbed (points), reconstructed (solid)

We simulate an experiment for a column with a van Genuchten-Mualem parametrization of the hydraulic functions and disturb the “simulated measurements” by a gaussian distributed noise (5% of the maximum value). The identification results are shown in Figs. 3 and 4.

References

Improvement of Environment Observing Remote Sensing Devices by Regularization Techniques

P. Maaß¹, Christine Böckmann², and Alexander Mekler¹

¹ Zentrum für Technomathematik, Universität Bremen, PF 330440, 28334 Bremen, Germany
E-mail: pmaass@math.uni-bremen.de, mekler@cnow.de

² Institut für Mathematik, Universität Potsdam, PF 601553, 14415 Potsdam, Germany
E-mail: bockmann@rz.uni-potsdam.de

Abstract. This project is carried out jointly with the support of the company Elight Laser Systems GmbH in Teltow. The goal is to improve the evaluation software of the measurements of so-called lidar (light detection and ranging) devices which are produced in the company. Such lidar devices are used for instance at Departments of Environment for air pollution control. To this end, we developed a capable and efficient regularization method for this inverse ill-posed problem by which future problems of this type may be treated routinely on PC’s.

1 Introduction

Aerosols affect life on earth in several ways. They play an important role in the climate system; the effect of aerosols on the global climate system is one of the major uncertainties of present climate predictions. They play a major role in atmospheric chemistry and hence affect the concentrations of other potentially harmful budget, in particular in the UV-B part of the spectrum. Moreover, one reason for the ozone depletion is the chlorine (Cl) in CFCs in the stratosphere. On the other side, polar stratospheric clouds (PSCs) a type of aerosol particles are believed to play an active role in precursor stages of ozone depletion in the winter-cold stratosphere by catalyzing heterogeneous chemical reactions on their surface and by redistributing HNO₃ through sedimentation (see [TT] and [T]). Such particles can be produced by volcanic eruption: in the stratosphere or by environmental pollution of the air above industrial areas. At ground level, they can be harmful, even toxic, to man, animals, and plants. Because of these adverse effects that aerosols can have on human life, it is necessary to achieve an advanced understanding of the processes that generate, redistribute, and remove aerosols in the atmosphere. A quantitative dataset describing the aerosol vertical, horizontal, and temporal distribution, including its variability on a continental scale, is necessary. Such a dataset could be used to validate and improve models that predict
the future state of the atmosphere and is dependence on different scenarios describing economic development, including those actions taken to preserve the quality of the environment (see [B et al]).

From the optical data it is necessary to determine the microphysical parameters of the particles, e.g. knowledge of the particle size distribution is necessary to model processes involving ozone chemistry (see [MT]). The size distribution of these cloud particles is an important parameter for quantifying those mechanisms, because it relates the total surface to the total mass. This distribution can be determined either by in-situ measurements with optical particle counters or by remote sensing with lidar equipment (see [BS], [B] and [MWA]).

Fig. 1. Lidar device for environmental monitoring by remote sensing (see [W])
The mathematical model for a lidar measurement consists of a system of two Fredholm integral equations of the first kind for the backscatter and extinction coefficients $\beta^{Aer}$ and $\alpha^{Aer}$

$$\beta^{Aer}(\lambda, z) = \int_{r_0}^{r_1} K_\pi(r, \lambda; m) \, n(r, z) \, dr$$

$$= \int_{r_0}^{r_1} \pi \, r^2 \, Q_\pi(r, \lambda; m) \, n(r, z) \, dr , \quad (1)$$

$$\alpha^{Aer}(\lambda, z) = \int_{r_0}^{r_1} K_{\text{ext}}(r, \lambda; m) \, n(r, z) \, dr$$

$$= \int_{r_0}^{r_1} \pi \, r^2 \, Q_{\text{ext}}(r, \lambda; m) \, n(r, z) \, dr , \quad (2)$$

where $r$ is the particle radius, $m$ the refractive index, $r_0$ and $r_1$ represent suitable lower and upper limits of realistic radii, $\lambda$ is the wavelength, $\lambda_0$ the smallest and $\lambda_1$ the largest wavelength, $z$ is the height, $n$ the aerosol size distribution we are looking for, $K_\pi$ the backscatter and $K_{\text{ext}}$ the extinction kernel. The kernel function reflects shape, size, and material composition of particles. We assume Mie particles. The following formulas hold for extinction and backscatter efficiencies (see [BHJ])

$$Q_\pi = \frac{1}{k^2 r^2} \left| \sum_{n=1}^{\infty} (2n + 1)(-1)^n (a_n - b_n) \right|^2 ,$$

$$Q_{\text{ext}} = \frac{2}{k^2 r^2} \sum_{n=1}^{\infty} (2n + 1) Re(a_n + b_n) , \quad (3)$$

where $k$ is the wave number defined by $k = 2\pi/\lambda$ and $a_n$ and $b_n$ are the coefficients which we get from the boundary conditions for the tangential components of the waves. The determination of the aerosol size distribution function $n(r)$ from a small number of backscatter and extinction measurements is a nonlinear inverse ill-posed problem.

2 Ill-Posed Problem and Regularization

The equations (1) and (2) are ill-posed on all three counts (existence, uniqueness, and stability), where stability means a solution that changes only slightly with a slight change in the problem (see [G]). We consider an operator of the form $Tx = y$ where $T : H_1 \to H_2$ is a compact, linear (but not necessarily self-adjoint) operator from a Hilbert space $H_1$ into a Hilbert space $H_2$. For a bounded linear operator $T$ a solution $x \in H_1$ of the equation $Tx = y$ exists
if and only if $y$ belongs to $R(T)$, the range of $T$. Since $T$ is linear, $R(T)$ is a subspace of $H_2$, which in generally does not exhaust $H_2$. We may enlarge the class of functions $y$ for which a type of generalized solution exists to a dense subspace of function in $H_2$. This accomplished by introducing the idea of a least squares solution. A function $x \in H_1$ is called a least squares solution if

\[ ||Tx - y|| = \inf\{||Tu - y|| : u \in H_1\} . \tag{4} \]

The set of all least squares solutions is closed and convex. Therefore, there is a unique least squares solution of smallest norm which we call generalized solution. The mapping $T^\dagger$ that associates with a given $y \in D(T^\dagger) = R(T) + R(T)^\bot$ the unique least squares solution having smallest norm, $T^\dagger y$, is called the Moore-Penrose generalized inverse of $T$. In our scheme $T^\dagger$ is then the mechanism which provides a unique solution for any $y \in D(T^\dagger)$. In this sense, $T^\dagger$ settles the issues of existence and uniqueness for generalized solutions. The generalized Pseudoinverse operator $T^\dagger : D(T^\dagger) \to H_1$ is a closed densely defined linear operator which is bounded if and only if $R(T)$ is closed. Since both lidar integral operators are compact, each of them can have closed range if and only if $R(T)$ is a finite dimensional subspace of $H_2$. This is not the case just under the given lidar integral kernels. Therefore, $R(T)$ is not closed so $T^\dagger$ is unbounded, i.e. $T^\dagger$ is discontinuous. Very small changes in the right hand side $y(\lambda)$ can be accounted for by large changes in the solution $x(r)$. That the instability is fundamental, and not just a consequence of some special form of the kernels, follows from the Riemann–Lebesgue lemma.

If we wish to obtain a well-posed problem we need a so called regularization. In general regularizations are families of operators

\[ T_\gamma : H_2 \to H_1 \text{ with } \lim_{\gamma \to 0} T_\gamma y = T^\dagger y \text{ for all } y \in D(T^\dagger) , \tag{5} \]

i.e. the convergence is pointwise on $D(T^\dagger)$ (see [L]). The parameter $\gamma$ is the so-called regularization parameter. In the case of noisy data $y^\delta$ with $||y - y^\delta|| \leq \delta$ we determine as solution

\[ x^\delta_\gamma \equiv T_\gamma y^\delta . \tag{6} \]

However, the total error consists of two parts, i.e. two summands,

\[ x^\delta_\gamma - x = T_\gamma (y^\delta - y) + (T_\gamma - T^\dagger)y . \tag{7} \]

The first part is the data error and the second part the approximation error or regularization error. If $\gamma \to 0$ the approximation error tends to zero while the data error tends to infinity. Therefore, the total error can never be zero and we are in a dilemma. We have to look for an “optimal” regularization parameter $\gamma$ which minimizes the total error.

### 2.1 Some Properties of the Lidar-Operators Kernels

Put $k_i = 2\pi/\lambda_i$ ($i = 0, 1$), $H_1 = L_2(r_0, r_1)$, $H_2 = L_2(k_0, k_1)$. Thus in (1*) and (2*) two families of integral operators $T^{(m)}_\pi : H_1 \to H_2$ and $T^{(m)}_{ext} : H_1 \to H_2$
are defined which kernels are
\[
K_\pi^{(m)}(r, k) = \pi k^{-2} \left| \sum_{n=1}^{\infty} (2n+1)(-1)^n(a_n - b_n) \right|^2,
\]
\[
K_{ext}^{(m)}(r, k) = 2\pi k^{-2} \sum_{n=1}^{\infty} (2n+1)\text{Re}(a_n + b_n)
\]
\[
= 2\pi k^{-2} \text{Re} \sum_{n=1}^{\infty} (2n+1)(a_n + b_n), \tag{8}
\]
respectively. The restrictions on the real variables \(r, k\) and on a complex parameter \(m\) may be written in the form
\[
0 < x < \infty, \ m^2 \notin M_\varepsilon \tag{9}
\]
where \(x := rk, \varepsilon > 0\) is an arbitrary small fixed number and \(M_\varepsilon\) denotes an \(\varepsilon\)-ring around the circle \(\{m : |m - 1| = 2|m|\}\) on the complex plane.

**Theorem 1.** Both of the series in (8) are uniformly convergent in every bounded subdomain of the definition domain (9).

**Theorem 2.** Both of the functions \(k^2K_{ext}^{(m)}\) and \(k^2K_\pi^{(m)}\) may be represented in form of power series with respect to \(x = rk\) namely \(\sum_{j=0}^{\infty} \gamma_j^{(ext)} x^j, \sum_{j=0}^{\infty} \gamma_j^{(\pi)} x^j\), respectively, where \(\gamma\)'s denote some real coefficients.

**Remarks.**
1) It follows from Theorem 1 that the series in (8) represent analytical functions in every bounded subdomain of the domain (9). It should be noted however that the kernels are not differentiable with respect to \(m\).
2) For the operators \(T_\pi^{(m)}\) and \(T_{ext}^{(m)}\) a kind of the one-to-one property follows from Theorem 2 that allows to apply the methods of [TGSY] to looking for a solution of the inverse problem.
3) Due to Theorem 2 one can discover some connections between analysis of the lidar-operators and so-called classical moment problem.
4) The decay rate of the singular values \(\mu_i\) of the lidar operators is so fundamental for the behavior of ill-posed problems that it makes sense to use this decay rate to characterize the degree of ill-posedness of the problem. Hofmann (see [Ho]) gives the following definition: if there exist a positive real number \(\alpha\) such that the singular values satisfy \(\mu_i = O(i^{-\alpha})\), then \(\alpha\) is called the degree of ill-posedness. The problem is characterized as mildly or moderately ill-posed if \(\alpha \leq 1\) or \(\alpha > 1\), respectively. On the other hand, if \(\mu_i = O(e^{-\alpha i})\), i.e. the singular values decay very rapidly, then the problem is termed severely ill-posed. For our lidar operators (1) and (2) we determined the condition numbers and by a numerical weighted nonlinear least squares method fit an approximation to the degree of ill-posedness. In general, one observes that the lidar operators are moderately ill-posed. In detail, we observe that the degree of the extinction operator is higher than the degree of the backscatter.
one (see [MB]). Moreover, if the absorption of the particles is large then the degree grows. As one expects the condition numbers grow with \( n \) and they show the same behavior as the degree.

### 2.2 Projection Methods as Regularization

Now our aim is to approximate \( T^\dagger y \) in the sense of (5). We know that, ignoring the trivial case in which the kernel \( K(\cdot,\cdot) \) is degenerated, the generalized solution \( T^\dagger y \) depends discontinuously on \( y \), but we would like to make our approximation continuously depending on \( y \). There are a lot of regularization methods, we refer to [EHN]. But if one would like to solve a real practical problem the results of regularization for infinite dimensional spaces are unsuitable. Hence we need a discretization (see [E]). On the one hand it is possible to combine any regularization method with any projection method. On the other hand one observes that pure projection methods into finite dimensional spaces act as regularization themselves without any additional filter, where the regularization parameter is \( n \). In addition, there is a regularization influence by the basis function choice, too. Let \( X_1 \subseteq X_2 \subseteq \ldots \subseteq H_1 \) be finite dimensional subspaces of \( H_1 \) with \( \overline{\bigcup_{n=1}^{\infty} X_n} = H_1 \), i.e. dense in \( H_1 \), and \( T_n := T|_{X_n} \) is the restriction of \( T \) to a subspace \( X_n \) of \( H_1 \). A natural way to generate a finite dimensional approximation is to find the minimal least squares solution of the equation \( T_n x = y \). As an approximation to \( T^\dagger y \) one could use the unique least squares solution, i.e. \( T^\dagger_n y \) or \( T^\dagger_n y^\delta \), respectively, where \( ||y - y^\delta|| \leq \delta \) represents the noise level of the data. The approximate solution \( x_n \in X_n \) minimizes \( ||Tx - y^\delta||^2 \) over \( X_n \). Since \( X_n \) is finite dimensional, \( R(T_n) \) is closed, i.e. \( T^\dagger_n \) is continous. This problem is well-posed. For more details we refer to [E], [Ha] and [G].

### 2.3 Hybrid Method with Variable Projection via B-Splines

From the theoretical studies of the lidar operators we get the knowledge how to develop the regularization method. Because of computer time consuming we decided to use a collocation method which is a special case of the Galerkin method. We propose now a hybrid regularization technique, a combination of a variable dimension projection method with truncated singular value decomposition. To turn into a finite dimensional problem we might simply try to solve the problem over a finite dimensional subspace of \( H_1 \). For example, if \( X_n \) is an \( n \)-dimensional subspace of \( H_1 \) spanned by the linearly independent vectors \( \{\phi_1, \ldots, \phi_n\} \) then the vector \( x_{n,\gamma}^\delta \in X_n \) minimizes \( ||Tx - y^\delta|| \) over \( X_n \) with a truncation level \( \gamma \). The solution \( x_{n,\gamma}^\delta \) can be represent as \( x_{n,\gamma}^\delta = \sum_{i=1}^{n} d_i \phi_i \) where the unknown coefficients \( d_i, i = 1, \ldots, n \), are the generalized solution of the linear equation system

\[
\sum_{i=1}^{n} \int_{r_{\text{min}}}^{r_{\text{max}}} K(r, \lambda_j) \phi_i(r) \, dr \, d_i = y(\lambda_j), \quad j = 1, \ldots, N + M,
\]
which may be underdeterminate or overdeterminate, respectively. The points
\( \lambda_j, j = 1, \ldots, N + M \), are so-called collocation points, in general they are
the measurement points, where \( N \) and \( M \) are the numbers of backscatter or
extinction coefficients, respectively. We solve (10) by using truncated singular
value decomposition with level \( \gamma \). We might call this type of discretization a
“finite element” discretization because the computed numbers \( d_i \) are coeffi-
cients of certain basis functions \( \{ \phi_i \} \) which often will be taken as B-spline
functions on some grid. A spline of degree \( k - 1 \) (order \( k \)) is a function
\( s \in C^{k-2}[r_{\min}, r_{\max}] \) which in every interval \([\tau_i, \tau_{i+1}], \ (i = 0, \ldots, l)\), consists
of a polynomial \( s_i \) with degree \( \leq k - 1 \). The spline space of degree \( k - 1 \) we
denote by \( S_{k, \Delta} \). Obviously the dimension is \( \dim S_{k, \Delta} = k + l =: n \), where
\( l + 2 \) is the number of different nodes. An advantageous basis we get with the
recursion \((X_{[\tau_i, \tau_{i+1}]} \) denotes the characteristic function\)

\[
N_{i+1}(r) := \chi_{[\tau_i, \tau_{i+1}]}(r) = \begin{cases} 1 & \text{if } r \in [\tau_i, \tau_{i+1}] \\ 0 & \text{otherwise} \end{cases}, \quad (11)
\]

\[
N_{ik}(r) := \frac{r - \tau_i}{\tau_{i+k-1} - \tau_i} N_{i,k-1}(r) + \frac{\tau_{i+k} - r}{\tau_{i+k} - \tau_{i+1}} N_{i+1,k-1}(r), \quad (12)
\]

where \( \tau_1 \leq \ldots \leq \tau_n \) are the extended nodes, \( N_{ik}(r) \) are the B-splines of order
\( k, k = 1, \ldots, n; \ i = 1, \ldots, n - k \). The support \( N_{ik} \subset [\tau_i, \ldots, \tau_{i+k}] \) is a local one,
\( N_{ik}(r) \geq 0 \) for all \( r \in \mathbb{R} \), \( N_{ik}(r) \) is a piecewise polynomial of degree \( \leq k - 1 \)
with respect to the interval \([\tau_j, \tau_{j+1}]\); furthermore \( N_{ik}, i = 1, \ldots, n \) are local
linear independent. Moreover, \( B = \{ N_{ik}, \ldots, N_{n,k} \} \) is a well-conditioned basis
of \( S_{k, \Delta} \) (see [DH]). Each spline \( s \in S_{k, \Delta} \) is a unique convex combination of
so-called de Boor-points \( d_i \) of \( s \) where \( s = \sum_{i=1}^{n} d_i N_{ik} \). Now our regularized
solution has the description

\[
x^\delta_{n, \gamma}(r) = \sum_{i=1}^{n=k+l} d_i \ N_{ik}(r) \quad (13)
\]

thus we have three regularization parameters \( k, n \) and \( \gamma \). For more details
see [B].

3 Numerical Results

From numerical point of view, i.e., with the degree of ill-posedness in mind,
it is reasonable to formulate equations (1) and (2) into a more specific form

\[
y(\lambda_j) = \int_{r_{\min}}^{r_{\max}} \widetilde{K}^\nu(r, \lambda_j; m) \ v(r) \ dr \quad (14)
\]

with

\[
\widetilde{K}^\nu(r, \lambda_j; m) := \begin{cases} K^\nu_r(r, \lambda_j; m) & : \lambda_j \in \lambda^\pi \\ K^\nu_{ext}(r, \lambda_j; m) & : \lambda_j \in \lambda^{ext} \end{cases}, \quad (15)
\]
where \( y(\lambda_j) \) are the optical data (whether it is backscatter \( \beta \) or extinction \( \alpha \)) depending on \( \lambda_j \) where \( \lambda^\pi = \{355, 400, 532, 700, 800, 1064\, \text{nm} \} \), i.e. \( N = 6 \) while \( \lambda^{ext} = \{355, 532\, \text{nm} \} \) (two Raman channels), i.e. \( M = 2 \). The \( v(r) \) term is the volume concentration distribution obtaining from \( n(r) \) with multiplying by \( \frac{4\pi r^3}{3} \) and

\[
K_{v/\text{ext}}(r, \lambda; m) = \frac{3}{4r} Q_{\pi/\text{ext}}(r, \lambda; m),
\]

we refer to the equations (1) and (2).

The reconstruction results of the inversion by using synthetically optical data were compared with the input distribution and the input refractive index. Logarithmic-normal distributions are used to describe the particle size distribution spectrum

\[
n(r) = \frac{1}{r} \frac{1}{\sqrt{2\pi} \ln \sigma} \exp(-0.5 \frac{(\ln r - \ln r_{mod})^2}{\ln^2 \sigma}).
\]
The particle parameters are the mode radius $r_{\text{mod}} = 0.3\, \mu m$ and the mode width $\sigma = 1.6$. Moreover, for the inversion we use the refractive index $m_1 = 1.4 + 0.005i$ (see Fig. 2a and 2b) or $m_2 = 1.7 + 0.1i$ (see Fig. 2c and 2d) and the lower and upper integration limits $r_0 = 0.001\, \mu m$ and $r_1 = 2.0\, \mu m$, respectively.

Beside the volume distribution, the complex refractive index is the second unknown quantity in the inversion. This problem is a highly nonlinear one and in general, no explicit information on the refractive index is available. Although we know that the refractive index depends on the wavelength as well as on the particle size, we assume the index to be constant. By using a refractive index grid we try to enclose the area of possible refractive indices and actually with that technique one obtains a much promising result, see Fig. 2a and 2d. The resulting area is the black one with respect to an assumed noise level. The point with an arrow is with respect to a special norm choice, here simple Euclidean norm, the best one in that sense. But we have to remark that in the case of noisy data the point with the arrow must not be the best one. All points of the black area must be considered as a possible solution. Otherwise one observes the same effect as with known refractive index, specifically if the imaginary part of the complex refractive index becomes larger then the problem gets more ill-posed, see Fig. 2d. Because the imaginary part is large, i.e. $0.1i$, the black area obtains a shape so that on the one hand the imaginary part can be estimated, but on the other hand the real part cannot be limited with this method. This is an ongoing work.

Finally, we present a real-life example from the Lindenberg Aerosol Characterization Experiment LACE 98 (Lindenberg/Germany, 54.2°N, 14.2°E): Measurement case from 9 August 1998, 20:00 to 22:00 UTC. During the measurement period the atmospheric boundary layer stretched to approximately 500–1000 m. The prominent feature is an elevated layer in a height from 3000 to 6000 m. Cirrus was present from 7000 to 9000 m. The elevated layer could be observed by lidar. The origin of the elevated layer was northwestern Canada. That could be explained by backtrajectory analysis of the Meteorological Observatory at Lindenberg. Satellite pictures of this area in Canada revealed heavy forest fires, which suggests that the air mass was loaded with soot particles from biomass burning (see [Metal]). The improved inversion with a suitable weighted T-Norm, see Fig. 3, is performed with six backscatter coefficients from the backscatter lidar and two extinction coefficients from the Raman lidar. The interesting microphysical parameters (see [B]) are the effective radius, the surface-area concentration, the volume concentration, the number concentration and the refractive index. We inverted them to

\[
r_{\text{eff}} = 0.24 \pm 0.005\, \mu m,
\quad a_G = 125 \pm 7\, \mu m^2/cm^3,
\quad v_G = 10 \pm 0.5\, \mu m^3/cm^3,
\quad n_G = 685 \pm 314\, cm^{-3}
\quad \text{and}
\quad m = 1.63 \pm 0.05\, + \, 0.0675 \pm 0.0175\, i,
\]

for more details see Fig. 3. The interpretation of the determined physical particle properties and the comparision with good agreements to [M et al] except the number concentration support the assumption of biomass burning. The number concentration in principle is hard to derive, because with the lidar
Fig. 3. Real-life inversion results with unknown refractive index

only optically active particles in the size region > 50 nm can be seen, but high and variable particle number concentrations occur in the entire size range < 100 nm.

Acknowledgments. This work has been supported by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) under grant 03-MA7PO1/5 and 03-MA7BR1/7. The first author is very grateful to Detlef Müller from the Institute of Tropospheric Research, Leipzig, for the preparation of the measurement date.

References


III. Flow, Transport and Reactions in Technological Processes

Stability Analysis for Reactors from Chemical Industry
B. Fiedler, A. Schuppert, M.A. Efendiev, L. Lerman, J. Rademacher

Heterogeneous Dynamic Process Flowsheet Simulation of Chemical Plants
F. Grund, K. Ehrhardt, J. Borchardt, D. Horn

Numerical Simulation of Annular Chromatography
L. Tobiska, A. Thiele

Numerical Methods for Parameter Estimation in Bingham-Fluids
G. Wittum, V. Schulz, B. Maar, D. Logashenko

A Viscoelastic Turbulence Model Based on Renormalization Group Theory
M. Niggemann, M. Holzmann, D. Schmidt, K. Soldner

Modelling and simulation of capacitor impulse welding
D. Homberg, W. Dreyer, F. Duderstadt

Analysis of Transport Processes for Layered Porous Materials Used in Industrial Applications
H. Neunzert, A. Zemitis, K. Velten, O. Iliev

Modelling and numerical simulation of district heating networks with time-saving solution method
R. D. Grigorieff, R. Kocher

Sensitivity and Robustness Analysis for Construction and Monitoring of Turbine-Generator Shafts
D. Pratzel-Wolters, P. Lang, A. Wirsen, S. Kulig
Stability Analysis for Reactors from Chemical Industry

B. Fiedler\textsuperscript{1}, M. A. Efendiev\textsuperscript{2}, L. Lerman\textsuperscript{1}, J. Rademacher\textsuperscript{1}, and A. Schuppert\textsuperscript{3}

\textsuperscript{1} Freie Universität Berlin, Fachbereich Mathematik
Arnimallee 2-6, 14195 Berlin, Germany
\textsuperscript{2} Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstr. 39, 10117 Berlin, Germany
\textsuperscript{3} BAYER AG Leverkusen
ZT-TEZ,E41, 51386 Leverkusen, Germany

Abstract. We present, for a class of industrially relevant chemical reactions the dependence of stability on important chemical parameters, such as coolant, dilution and diffusion rates. Moreover, the influence of the order of reaction rates for stability is considered. Our analysis is based on a generalized upscaling balance condition for the equilibria concentrations.

1 Model Reactor

As an example for the stability of a chemical reactor we consider the following radical reaction

\begin{align}
S &\rightarrow 2S^* \\
S^* + E &\rightarrow E^* + D_1 \\
E + E^* &\rightarrow P^* \\
P^* + E &\rightarrow P + E^* \\
2E^* &\rightarrow D_2 \\
E^* + S^* &\rightarrow D_3
\end{align}

(1) (2) (3)

Here \(S\) is a radical starter, such as hydrogen peroxide or UV-light, and the chain (2) converts the educt \(E\) into the product \(P\). The corresponding radicals are \(S^*, E^*, P^*\). Inconsequential side products are denoted by \(D_1, D_2, D_3\). This reaction scheme plays an important role in chemical industry. Indeed, let us consider

a) Oxidation of Cumene \(E = (C_9H_{12})\) with Azo-bis-isobutyronitril \(S = (C_8N_4H_{12})\). The product \(P = C_9H_{11}OOH\) (hydrogen peroxide) is an important chemical substance for obtaining Phenole \((C_6H_5OH)\) and Acetone \((CH_3COCH_3)\).
b) Polymerisation of Styrene $E = (C_8H_8)$ with Dicyclohexylperoxydicarbonat $S = (C_4H_{22}O_6)$. The important product is Polystyrene $P = C_{16}H_{16}$.

Assuming, aside from a small region of laminarity $\Omega$, the reactor to be well-mixed, we find that the corresponding dynamics for reaction mechanism (1)–(3) will be governed by (see [5])

$$\begin{align*}
\dot{S} &= S_0 - k_1 S - \tilde{D}_1 S \\
\dot{E} &= E_0 - k_2 s E - k_3 e E - k_4 p E - \tilde{D}_2 E \\
\dot{s} &= 2k_1 S - k_2 s E - k_5 e s - \tilde{D}_3 s \\
\dot{e} &= k_2 s E - k_3 e E + k_4 p E - 2k_5 e^2 - k_6 e s - \tilde{D}_4 E \\
\dot{p} &= k_3 e E - k_4 p E - \tilde{D}_5 p \\
\dot{T} &= h_1 k_1 S + h_2 k_2 s E + h_3 k_3 e E + h_4 k_4 p E + h_5 k_5 e^2 + h_6 k_6 e s - \kappa(T - T_0)
\end{align*}$$

Here $T$ denotes temperature, $h_i$ are the reaction enthalpies, and the reaction rates

$$k_i = k_i(T) = k_i^* \exp(-\gamma_i/T)$$

are of Arrhenius type with activation energies $\gamma_i$ and rate constants $k_i^*, i = 1, \ldots, 6$. Externally sustained, space dependent feed concentrations are denoted by $E_0, S_0$. Note that $D = (\tilde{D}_1, \ldots, \tilde{D}_5)$ and $\kappa$ indicates dilution and external cooling rate with coolant temperature $T_0$.

**Remark 1.** For the reaction mechanism mentioned above we have following data for reaction rates for Cumene ($C_9H_{12}$) and Styrene ($C_8H_8$):

**Cumene:**

- $k_1 = 1.4 \cdot 10^{-4}$, $k_3 = 3.5$, $k_5 = 3 \cdot 10^5(T = 80°C)$

**Styrene:**

- $k_1 = 1.5 \cdot 10^{-4}$, $k_3 = 58$, $k_5 = 3.9 \cdot 10^6(T = 60°C)$

In general, for this class of radical chemical reactions the reaction rates can often be determined only up to the order of magnitude [1,3]. More precisely:

$$\begin{align*}
k_1, k_2 & : \text{from } 10^{-9} \text{ to } 10^{-5} \\
k_3, k_4 & : \text{from } 0, 2 \text{ to } 100 \\
k_5, k_6 & : \text{from } 10^5 \text{ to } 10^8
\end{align*}$$

## 2 Generalized Upscaling Balance Condition for Equilibria Concentrations

Following Feinberg [4], we write chemical reactions of species $x_0, \ldots, x_n$ as

$$y_0^{(j)} x_0 + \ldots + y_n^{(j)} x_n \xrightarrow{k_j} \tilde{y}_0^{(j)} x_0 + \ldots + \tilde{y}_n^{(j)} x_n$$

(7)
with integer stoichiometric coefficients \( y^{(j)}, \bar{y}^{(j)} \geq 0 \) and positive real reaction rate coefficients \( k_j \). Then the associated dynamics is given by

\[
\dot{x} = \sum_{j=1}^{M} k_j x^y (\bar{y}^{(j)} - y^{(j)}).
\]  

Here \( x = (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} \) is the vector of concentrations and \( x^y := x_0^y \cdots x_n^y \) represents the usual mass action kinetics.

Note that the external feed concentrations can be incorporated in this model by adding formal reactions

\[
0 \xrightarrow{k_0} x_i.
\]  

For the radical starter reactions considered in the present paper, we can therefore introduce a crucial small parameter \( \epsilon > 0 \), in this sense, by the feed "reaction"

\[
0 \xrightarrow{k_0=\epsilon} S
\]  

with \( \epsilon := S_0 \) denoting the feed concentration of the radical starter. Let \( \alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1} \) denote the scaling defined by

\[
\begin{pmatrix}
x_0 \\
\vdots \\
x_n
\end{pmatrix} =
\begin{pmatrix}
\epsilon^{\alpha_0} & 0 \\
\vdots & \ddots \\
0 & \epsilon^{\alpha_n}
\end{pmatrix}
\begin{pmatrix}
\xi_0 \\
\vdots \\
\xi_n
\end{pmatrix}.
\]  

Substituting (11) into (8) we obtain

\[
e^{\alpha_j} \bar{y}^{(j)} = \sum_{j=1}^{M} k_j \epsilon^{\alpha_0 y_0^{(j)} + \cdots + \alpha_n y_n^{(j)}} \xi_0^{y_0^{(j)}} \cdots \xi_n^{y_n^{(j)}} (\bar{y}^{(j)} - y^{(j)}).
\]  

With the abbreviation \( \alpha \cdot y^{(j)} = \alpha_0 y_0^{(j)} + \cdots + \alpha_n y_n^{(j)} \) for the scalar product, the equilibrium equation for (12) is

\[
\sum_{j=1}^{M} k_j \epsilon^{\alpha \cdot y^{(j)}} \bar{y}^{(j)} (\bar{y}^{(j)} - y^{(j)}) = 0
\]  

for the scaled steady state \( \bar{\xi} = \xi(\epsilon) \).

Taking into account (5) and (6) we can assume that:

\[
k_j = k_{j,0} e^{\beta_j} + o(\epsilon^{\beta_j}),
\]  

where \( \epsilon := S_0 \) denotes the small starter concentration.

**Definition 2.** We call a pair \( (\alpha, \beta) = (\alpha_0, \ldots, \alpha_n, \beta_1 \ldots, \beta_M) \in \mathbb{R}^{n+1} \times \mathbb{R}^M \) a generalized scaling exponent for the steady state concentrations \( x = \)
(x₀, ⋅⋅⋅, xₙ) of (8), if there exists a sequence of scaled equilibria ξ(εₗ) = (ξ₀(εₗ), ⋅⋅⋅, ξₙ(εₗ)), εₗ → 0, ℓ → ∞ of (13), with k_j given by (14), such that for all m = 0, ⋅⋅⋅, n the following nondegeneracy conditions holds:

\[ 0 < \lim_{\ell \to \infty} \inf_{\xi_m(εₗ)} \leq \lim_{\ell \to \infty} \sup_{\xi_m(εₗ)} < \infty. \tag{15} \]

**Theorem 3.** Any generalized scaling exponent 
\[(α, β) = (α₀, ⋅⋅⋅, αₙ, β₁, ⋅⋅⋅, β_M) \in \mathbb{R}^{n+1} \times \mathbb{R}^M \]
satisfies

\[ \min_{j: y[m]^{(j)} = 0} \{(α \cdot y^{(j)}) + β_j\} = \min_{i: y[m]^{(i)} = 0} \{(α \cdot y^{(i)}) + β_i\}, \tag{16} \]

for all fixed m = 0, ⋅⋅⋅, n.

**Proof.** Let \((α, β) \in \mathbb{R}^{n+1} \times \mathbb{R}^M\) be any generalized scaling exponent. We rewrite equation (13) for the scaled steady state \(ξ = ξ(ε)\) in components \(m = 0, ⋅⋅⋅, n\) as

\[ \sum_{j=1}^{M} k_j ε^{α \cdot y^{(j)}} ξ y^{(j)} y_m^{(j)} = \sum_{i=1}^{M} k_i ε^{α \cdot y^{(i)}} ξ y^{(i)} y_m^{(i)} \tag{17} \]

or, equivalently,

\[ \sum_{j: y[m]^{(j)} \neq 0} k_j ε^{α \cdot y^{(j)}} ξ y^{(j)} y_m^{(j)} = \sum_{i: y[m]^{(i)} \neq 0} k_i ε^{α \cdot y^{(i)}} ξ y^{(i)} y_m^{(i)}. \tag{18} \]

Let \(k_j = k_j β_j + o(ε^{β_j})\) and fix \(m\). We sort the terms on each side of (18) by increasing powers \(α \cdot y^{(j)} + β_j, α \cdot y^{(i)} + β_i\) of \(ε\) respectively. Note that all terms are strictly positive by the nondegeneracy assumption (15). Let

\[ \min_{j: y[m]^{(j)} \neq 0} \{(α \cdot y^{(j)}) + β_j\} = α \cdot y^{(j₀)} + β_{j₀} \text{ and } \min_{i: y[m]^{(i)} \neq 0} \{(α \cdot y^{(i)}) + β_j\} = α \cdot y^{(i₀)} + β_{i₀}, \]

in (18). The leading (alias: lowest) power of \(ε\) in the \(m\)-th component of (18) is given by \(α \cdot y^{(j₀)} + β_{j₀}\) on the left, as compared to \(α \cdot y^{(i₀)} + β_{i₀}\) on the right. These leading powers may in fact be realized by several other terms \(β_j + α \cdot y^{(j)} + β_i\) in addition. Comparing coefficients, we divide (18) by \(ε\) to the power min \(α \cdot y^{(j₀)} + β_{j₀}, α \cdot y^{(i₀)} + β_{i₀}\), and let \(ε \to 0\). This immediately yields

\[ α \cdot y^{(j₀)} + β_{j₀} = α \cdot y^{(i₀)} + β_{i₀} \]

by positivity of all terms in the sums. This proves the Theorem 3. \(\square\)

Next, we apply the generalized upscaling condition to the reaction mechanism of the form (7)–(9). Following (9), (10) we augment the reaction mechanism (1)–(3) by adding formal reactions

\[ 0 \xrightarrow{k_0=S_0} S \]
\[ 0 \xrightarrow{k_1=E_0} E \tag{19} \]
with small \( \epsilon := S_0 \) and consider the limit \( \epsilon \searrow 0 \) henceforth. For the reaction mechanisms with \( x = (x_0, \ldots, x_4) = (S, E, s, e, p) \) and external starter concentrations \( S_0 = \epsilon \) the generalized scaling conditions (15) for the scaling exponent \( (\alpha, \beta) \) given by
\[
\alpha = (\alpha_0, \ldots, \alpha_4) := (\alpha_S, \alpha_E, \alpha_s, \alpha_e, \alpha_p), \beta = (\beta_1 \ldots, \beta_6)
\]
and \( D = 0 \) take the form
\[
S : = \min\{\alpha_S + \beta_1\} = 1 \quad (20)
\]
\[
E : = \min\{\alpha_E + \alpha_s + \beta_2, \alpha_E + \alpha_e + \beta_3, \alpha_E + \alpha_p + \beta_4\} = 0 \quad (21)
\]
\[
s : = \min\{\alpha_E + \alpha_s + \beta_2, \alpha_s + \alpha_e + \beta_6\} = \alpha_S + \beta_1 \quad (22)
\]
\[
e : = \min\{\alpha_E + \alpha_e + \beta_3, 2\alpha_e + \beta_5, \alpha_s + \alpha_e + \beta_6\} = \min\{\alpha_E + \alpha_s + \beta_2, \alpha_E + \alpha_p + \beta_4\}
\]
\[
p : = \min\{\alpha_E + \alpha_p + \beta_3\} = \min\{\alpha_E + \alpha_e + \beta_4\} \quad (24)
\]
Of course equations (20)–(24) admit a huge number of solutions, depending on \( \beta = (\beta_1, \ldots, \beta_6) \). For each particular chemical reaction we will select concrete numerical values of \( \beta_1, \ldots, \beta_6 \). Therefore, taking into account the information on the data \( k_j \), that is (5)–(6) for the reaction mechanism (1)–(3), we may assume that \( \beta_S := \beta_1 = \beta_2, \beta_c := \beta_3 = \beta_4, \beta_N := \beta_5 = \beta_6 \).

It is not difficult to prove, that (20) yields
\[
\alpha = \begin{pmatrix} 1, -1, -1, 3, 1, 1 \end{pmatrix}, \quad \text{if } D = 0 \text{ and } \beta = 0 \quad (25)
\]
\[
\alpha = \begin{pmatrix} 1, 0, 1, 1, 1, 1 \end{pmatrix}, \quad \text{if } D = (1, 1, 0, 0, 0) \text{ and } \beta = 0 \quad (26)
\]
\[
\alpha = \begin{pmatrix} 1, 0, 1, 1, 1, 1 \end{pmatrix}, \quad \text{if } D = (1, 1, 1, 1, 1) \text{ and } \beta = 0 \quad (27)
\]
\[
\alpha = \begin{pmatrix} 1 - \beta_s, \beta_N - 2\beta_c - 1, 3 - \beta_N + 2\beta_c - 2\beta_s, 1 - \beta_N, 1 - \beta_N \end{pmatrix}, \quad \text{if } \beta_s - \beta_c < 1, D = 0 \quad (28)
\]
\[
\alpha = \begin{pmatrix} 1, \alpha_E, \alpha_E + \beta_c - \beta_N + 1, -\alpha_E - \beta_c, -\alpha_E - \beta_c \end{pmatrix} \quad (29)
\]
\[
\text{if } \beta_s - \beta_c \geq 1, D = 0, \text{ where } \frac{\beta_N - \beta_c - \beta_s}{2} \leq \alpha_E \leq \frac{\beta_N - 2\beta_c - 1}{2}.
\]
Formulas (25)–(29) show the dependence of the scaling exponent \( \alpha = (\alpha_S, \alpha_E, \alpha_s, \alpha_e, \alpha_p) \) upon order of reaction rates and dilution ones. Note the singularity in the expansion for \( E \), when \( D = 0, \beta = 0 \) and \( D = 0 \) and \( \beta_s = 1, \beta_c = 0, \beta_N = -1 \).

**Remark 4.** Using the expansion in \( \epsilon \) of equilibria concentration, we can easily find the expansion in \( \epsilon \) of the steady state temperature. For example, the case
\[ D = 0, \beta = 0 \] yields an \( \varepsilon \)-expansion of the steady state temperature \( T^*(\varepsilon) \), that is

\[
T^*(\varepsilon) = \frac{h_3 + h_4}{\kappa} \cdot \frac{E_0}{2} + \kappa T_0 + o(\varepsilon)
\]

with higher order terms of order at least \( \varepsilon \).

### 3 Stability Analysis

For stability analysis we make use of asymptotic expansions of equilibria concentrations (25)-(29). First we restrict ourselves to \( D = \beta = 0 \). To this end we consider the following 6 \( \times \) 6 matrix \( A(\varepsilon) = \)

\[
\begin{bmatrix}
-k_1 & 0 & 0 & 0 & 0 & -k'_1 s \\
0 & -k_3 e - k_4 p - k_2 s & -k_2 E & -k_3 E & -k_4 E & \Theta_1 \\
2k_1 & -k_2 s & -k_2 E - k_6 e & -k_6 s & 0 & \Theta_2 \\
0 & k_2 s - k_3 e + k_4 p & k_2 E - k_6 e & -k_3 E - 4 k_5 e - k_6 s & k_4 E & \Theta_3 \\
0 & k_3 e - k_4 p & 0 & k_3 E & -k_4 E & \Theta_4 \\
h_1 k_1 & h_2 k_2 s + h_3 k_3 e + h_4 k_4 p & h_2 k_2 E + h_6 k_6 e & h_3 k_3 E + 2 h_5 k_5 e + h_6 k_6 s & h_4 k_4 & \Theta_5
\end{bmatrix}
\]

where \( \Theta_j, j = 1, \ldots, 5 \) denote

\[
\Theta_1 = -k'_1 s E - k'_2 e E - k'_4 p E; \quad k'_j := \frac{d}{dt} k_j
\]

\[
\Theta_2 = 2 k'_1 S - k'_2 s E - k'_6 e s
\]

\[
\Theta_3 = k'_2 s E - k'_3 e E + k'_4 p E - 2 k'_5 e^2 - k'_6 e s
\]

\[
\Theta_4 = k'_3 e E - k'_4 p E
\]

\[
\Theta_5 = h_1 k'_1 S + h_2 k'_2 s E + h_3 k'_3 e E + h_4 k'_4 p E + h_5 k'_5 e^2 + h_6 k'_6 e s - \kappa.
\]

We consider the following eigenvalue problem

\[
A(\varepsilon)x(\varepsilon) = \lambda(\varepsilon)x(\varepsilon)
\]

and are interested in the stability of the chemical reaction under consideration. We begin our stability analysis by grouping the matrix \( A(\varepsilon) \) according to its leading order in \( \varepsilon := S_0 \). For this purpose we use the scaling exponents (25). Then \( A(\varepsilon) \) can be decomposed as

\[
A(\varepsilon) = \varepsilon^{-1/2}[A_0 + o(\varepsilon^{1/2})].
\]

It remains to compute \( \lambda_j(A_0(\varepsilon)), j = 1, \ldots, 6 \). Note that the \( 5 \times 5 \) matrix which consists of the first five columns and rows of \( A(\varepsilon) \) has the first five negative eigenvalues. We denote this matrix by \( \tilde{A}_{0,5}(\varepsilon) \). It is not difficult to see that some of the eigenvalues of \( \tilde{A}_{0,5}(\varepsilon) \) have zero limit as \( \varepsilon \to 0 \) (remaining negative for \( \varepsilon > 0 \)).

Thus for stability analysis it remains to check the influence of the coolant rate \( \kappa \). For simplicity of computations we assume that \( k_3 = k_4 \). It is not
difficult to see that (using simple transformations) the stability analysis of
the \((6 \times 6)\) matrix \(A(\epsilon)\) is reduced to the \((4 \times 4)\) matrix of the form
\[
\begin{pmatrix}
  -\frac{E_0}{E} & -c & -k_3E & -k_4E \\
  -\frac{b_3+h_4}{2} \frac{E_0}{E} & c - \kappa & h_3k_3E + 2h_5k_5E & h_4k_4E \\
  0 & 0 & -k_3E - 2k_5E & k_4E \\
  0 & 0 & k_3E & -k_4E
\end{pmatrix}
\]
\[(33)\]
Here \(c\) is a positive constant depending on the activitation energy and the
given temperature. Hence, it follows from (33) that there exists \(\kappa_c > 0\) such
that for \(\kappa \geq \kappa_c\) we have stability and \(0 \leq \kappa < \kappa_c\) belongs to the instability
region.

The case \(D = (1, 1, 0, 0, 0)\) and \(\beta = 0\) can be handled in analogous way,
which is based on the scaling exponent (26). In this case, the \((6 \times 6)\) matrix
\(A(\epsilon)\) is decomposed as \(A(\epsilon) = A_0 + o(\epsilon)\) with leading \((6 \times 6)\) matrix, for which
the eigenvalues are determined by
\[
\begin{align*}
  \lambda_1(A_0) &= -k_1 - 1, \\
  \lambda_2(A_0) &= -1, \\
  \lambda_3(A_0) &= -k_2E_0, \\
  \lambda_4(A_0) &= -(k_3 + k_4)E_0, \\
  \lambda_5(A_0) &= 0 \\
  \lambda_6(A_0) &= -k_5 < 0 \text{ and } \\
  \lambda_7(A_0) &= -k
\end{align*}
\]
\[(34)\]
Next we are interested in the role of diffusion for the stability of the chem-
ical reactor (1)–(3). Let \(d = \text{diag} [d_1, \cdots, d_6]\) be the diagonal matrix with
elements consisting of diffusion rates \(d_1, \cdots, d_6\) (see[5]). Then holds

**Theorem 5.** Let \(D = 0\) and \(\beta = 0\). Then the interval
\[
k_c - \mu_1d_6 < k < \infty
\]
provides stability of the chemical reactor (1)–(3), where \(\mu_1\) is a first eigenvalue
of the Laplacian \((-\Delta)\) with Dirichlet boundary condition on \(\partial\Omega\).

**Proof (Sketch).** In this case stability will be governed by the equation
\[
\det(A(\epsilon) - \mu_1 \text{diag} [d_1, \cdots, d_6] - \lambda Id) = 0
\]
\[(35)\]
where \(A(\epsilon)\) is the same \((6 \times 6)\) matrix as above. We show that in the presence
of diffusion the stability condition for (35) is reduced to
\[
(\lambda + \mu_1d_2)(\lambda + \kappa - \kappa_c + \mu_1d_6) = 0,
\]
and takes the form \(\kappa > \kappa_c - \mu_1d_6\). This proves the Theorem 5. \(\square\)

It should be emphasized that the interval \(\kappa_c - \mu_1d_6 < \kappa < \kappa_c\) which provides
stability of the chemical reactor in the presence of diffusion belongs to the
instability interval in the diffusionless case. Analogously in the case \(D =
(1, 1, 0, 0, 0), \beta = 0\) one can prove the stabilizing role of diffusion.
Remark 6. The dependence of stability of chemical reactions from order of reaction rates is analyzed in the same way. Indeed, in each concrete case we have to choose numerical values of $\beta_s, \beta_c, \beta_N$. For example, based on (5) and (6) we can choose $\beta_s = 1, \beta_c = 0, \beta_N = -1$, which in turn – due to (29) – implies that $\alpha = (0, -1, 1, 1, 1)$. Hence, stability analysis for oxidation of Cumene $E = \text{C}_9\text{H}_{12}$ and polymerisation of Styrene $E = \text{C}_8\text{H}_8$ are governed by the $6 \times 6$ matrix $A(\epsilon)$, which has decomposition $A(\epsilon) = \epsilon^{-1}[A_0 + o(\epsilon)]$.

The following figures 1 and 2 show the dependence of eigenvalues from the order of reaction rates for $k_1 = k_2 = \epsilon, k_3 = k_4 = 1, k_5 = k_6 = \epsilon^{-1}$ and from diffusion respectively.

**Fig. 1.** Without diffusion. The curve for eigenvalue EW4 is hidden by EW2

**Fig. 2.** With diffusion. The curves for EW2, EW4, EW5 are hidden by EW6
4 Conclusion

For a class of industrially relevant chemical reactors we have presented a stability analysis, depending on crucial chemical parameters such as coolant, dilution and diffusion rates. Moreover, the role of the order of magnitude of the reaction rates for stability has been considered. The peculiarity of the class of reaction mechanisms under consideration is that they act catalytically even in low concentrations. Therefore, the reaction is usually run with low starter concentration $S$ producing a sufficient quantity of $E^\bullet$. For this class of chemical reactors the stabilizing role of diffusion has been shown. Our stability analysis has been essentially based on a generalized upscaling balance condition for equilibria concentrations developed by us for this purpose. Although we have illustrated the stability analysis for a model reactor (1 educt E), it should be noted that our approach is applicable to more complex reaction mechanisms, such as reactions with several educts (in particular, $E_1 = \text{Cl}_2, E_2 = \text{C}_7\text{H}_8$), reaction mechanism of third order or reaction mechanisms of second order with an unstable intermediate products (see [2]). We think that our analysis can describe phenomena in which chemical reactors remain stable for a while and then suddenly become unstable.

References

Heterogeneous Dynamic Process Flowsheet Simulation of Chemical Plants

Friedrich Grund, Klaus Ehrhardt, Jürgen Borchardt, and Dietmar Horn

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany
E-mail: grund@wias-berlin.de

Abstract. For large-scale dynamic simulation problems in chemical process engineering, a heterogeneous simulation concept is described which allows to distribute the solution of the models of coupled dynamic subprocesses to a computer network. The main principle of such a technique is to solve the submodels of an overall model independently of each other on subsequent time intervals. This is done by estimating the vector of input variables of the submodels, calculating the corresponding time behaviour of the output variables concurrently, and matching the time profiles of the interconnecting variables of the process flowsheet iteratively. Therefore, accelerated waveform iteration methods are considered, using Broyden- and block-Broyden-type updates. The simulation concept is investigated especially in the case that the submodels do not provide input-output sensitivities.

1 Distributed Simulation of Coupled Processes

The numerical simulation of dynamic processes for large-scale plants in chemical industry requires a strongly modular division of modelling tasks with respect to the design of sufficiently encapsulated subprocess models, their mathematical formulation and numerical solution. In [3]–[6] we have shown how such a modular modelling can be used to solve homogeneous dynamic simulation problems for complex industrial distillation plants, achieving considerable speedup factors on parallel computers with shared memory. But the modular modelling concept is even more significant in the case of heterogeneous simulation problems, where different co-workers, possibly working with different simulation tools on different computer platforms, have developed well working codes for subprocesses, which have now to be coupled with other subprocesses to simulate plantwide dynamic process flowsheets. An example for such a process flowsheet, assembled from 3 subprocesses, is shown in Fig. 1.

A general approach of coupling \( p \) subprocesses can be derived in the case that each submodel \( i \), which represents a specific subprocess of the entire plant, is determined by the submodel function which uniquely generates an output function \( v_i(t) \) for each given input function \( u_i(t) \), i.e.

\[
v_i(t) = G_i(u_i(t)), \quad i = 1, \ldots, p \quad t \in [t_0, t_E].
\]
Here the submodel functions $G_i$ are only given implicitly by applying a numerical solution procedure to the respective submodel equations. Frequently, the internal model of subprocess $i$ is described by a large system of differential algebraic equations (DAEs)

\[ F_i(t, y_i(t), \dot{y}_i(t), u_i(t)) = 0, \]

\[ y_i(t_0) = y_i^0, \quad (2) \]

with $y_i = (x_i, v_i)^T$. These internal submodel equations (2), the dimension of the resulting discretized model as well as their numerical implementation are not known outside the subprocess simulation. But it is worth mentioning that the dimensions of the vectors of input and of output variables $u_i$ and $v_i$ are usually small compared with the large number of internal variables $x_i$.

In a global process flowsheet definition, each component of the overall input vector $u(t) = (u_1, ..., u_p)^T$ corresponds uniquely to one component of the overall output vector $v(t) = (v_1, ..., v_p)^T$. All feed and energy streams entering the plant from the outside as well as the products, byproducts, and energy streams leaving the plant are balanced within the appropriate submodels, so that they do not need to be considered here. Therefore, the system of equations describing the coupling between the submodels can be defined by

\[ f(u(t)) \equiv u(t) - PG(u(t)) = 0, \quad t \in [t_0, t_E], \quad (3) \]

where $P = ((p_{ij}))$, with $p_{ij} \in \{0, 1\}$, is a permutation matrix which allocates the individual input and output variables of all subprocesses, and $G$ represents all submodel functions $G_i$, cf. (1).
A self-evident extension of the coupling equations (3) results from the assumption of a time-delay in submodel coupling which may be described by some simple residence-time model of material and heat flow within the interconnecting pipes like

\[ \frac{d u(t)}{dt} = \omega \left( PG(u(t)) - u(t) \right), \quad t \in [t_0, t_E] \]  

(4)

or more sophisticated plug-flow assumptions. This approach is not exploited directly within this paper, because it is expected that such effects have to be modelled within the related subprocess models only. But the equation (4) can be used to derive a natural relaxation strategy with \( \omega \rightarrow 0 \) for the iterative solution of (3) as it will be shown later on.

In order to investigate the dynamics of the coupling equations (3) in a strongly modular manner, one has to guarantee that within each submodel \( i \) there is no direct feedback of outputs \( v_i(t) \) to its own inputs \( u_i(t) \). The behaviour of the time-dependent system (3) is then mainly determined by the corresponding Jacobian

\[ J = I - P \frac{\partial G}{\partial u}, \]  

(5)

where \( \partial G / \partial u \) denotes the overall block diagonal matrix of the dynamic input-output-sensitivities \( \partial G_i / \partial u_i \) of the submodels. Generally, these dynamic sensitivities are not provided by today's commercial simulation tools. Even an initial estimation of the Jacobian at \( t_0 \) will be difficult to obtain with reasonable effort. Additionally, discontinuities of the solution and of parameter functions may arise in subprocesses at predefined time points or sporadically. Furthermore, a truly heterogeneous process simulation environment may require to handle different submodels on different computer platforms and to retain proven software for submodel solution with its own internal step size and accuracy control. Under these stringent modelling restrictions the subsequent algorithmic details for solving (3) have to be assessed.

2 Waveform Iteration Methods

It is obvious that for the distributed simulation of coupled subprocesses an adapted waveform iteration method, cf. [8] and [13], can be used. The basic idea of waveform iteration is to solve submodels of an overall model independently of each other on subsequent time intervals, so called windows. For that, on the current window, the time behaviour of the vector of input variables of the submodels is estimated, the corresponding time behaviour of the output variables is computed concurrently for all submodels, and the interconnecting variables of the flowsheet are matched iteratively. An overview of this approach with applications in chemical engineering and a discussion of preparatory literature has been given in [11].
2.1 The Discretized Problem

Due to the need to simplify communication and synchronization between subprocess simulations, the time variable \( t \in [t_0, t_E] \) is discretized within each window \([t_0^s, t_m^s]\) equidistantly. With \( t_0^s = t_0 \), one gets for the \( s \)'th window of length \( T^s = t_m^s - t_0^s \)

\[
t_j^s = t_{m}^{s-1} + j \times T^s / m, \quad j = 0, 1, \ldots, m.
\] (6)

A sufficient degree of adaptivity is still preserved by a flexible window length \( T^s \) as well as by the possibility to reduce the window size, after some iteration, by shifting the left boundary of the window from \( t_0^s \) to some \( t_l^s, 0 < l < m \), if convergence is achieved for all \( t_j^s, j = 0(1)l \) already.

Keeping in use the same symbols for notational simplicity, now \( U_i \) and \( V_i \) denote the vectors of all discretized input and output variables at all internal time points \( t_j^s, j = 0, \ldots, m \) of the current window in an appropriate order, and \( U \) and \( V \) the related overall vectors \( U = (U_1, ..., U_p)^T \) and \( V = (V_1, ..., V_p)^T \), respectively.

2.2 Iterative Solution of Coupling Equations

A Picard-type iteration, which solves (3) at the grid points \( t_j^s, j = 0(1)m \), of the current window, is given by

\[
v_i^{k+1} = G_i(U_i^k), \quad i = 1, ..., p \tag{7}
\]
\[
u^{k+1} = P \nu^{k+1}, \tag{8}
\]

where \( k \) denotes the iteration index. Obviously, the evaluation of (7) can be done in parallel for all subprocesses, followed by a common step (8) of allocating the output variables to the related inputs. It is noteworthy to state that it is nowhere required at all that the elements of \( U \) and \( V \) have to be discrete function values, but instead they can be parameters of an appropriate continuous approximation of the trajectories of the interconnecting variables. In this paper we use a linear interpolation in order to get continuous piecewise linear approximations of the inputs \( U_i \) in (7). This is realized by calling an additional interpolation subroutine within each function evaluation of the associated systems (2). It avoids the permanent reinitialization of the integration procedure, as it has to be done in the case of piecewise constant inputs, and increases the approximation accuracy.

To accelerate the convergence of the Picard iteration, (8) can be substituted by a quasi-Newton approach

\[
u^{k+1} = \nu^k - (B^k)^{-1} (\nu^k - PG(u^k)), \tag{9}
\]

where \( B^k \) should be an approximation of the Jacobian, preferably

\[
I - PG_u^k. \tag{10}
\]
Here \( G_u^k \) represents a reasonable approximation of the overall block diagonal matrix of the dynamic input-output sensitivities of submodels at all internal time points of the current window. Generally, a trusted estimation of \( G_u^k \), even at \( t_0 \), will be difficult to get. Therefore, we have been looking for an update procedure in the case that no initial estimation is available, i.e. the iteration starts with an iteration matrix \( B^0 = I \), as it is discussed in [12], or in the case that only some of the submodels provide such information.

**Broyden update:** Numerous versions and modifications of Broyden update formulas have been derived in the past, mainly focussed on improving a given iteration matrix in (9) by calculating an adjacent matrix which still satisfies the quasi-Newton condition with the most recent function values, cf. [9], [12],

\[
B^{k+1} \Delta u^k = \Delta f^k. \tag{11}
\]

In this paper the classical Broyden method is adapted to the investigated waveform iteration technique and a possible extension to it is discussed.

With abbreviations

\[
\Delta u^k \equiv u^{k+1} - u^k \tag{12}
\]
\[
\Delta v^k \equiv v^{k+1} - v^k \tag{13}
\]
\[
\Delta f^k \equiv f^{k+1} - f^k = \Delta u^k - P \Delta v^k \tag{14}
\]

the conventional Broyden approach

\[
B^{k+1} = B^k - \frac{(B^k \Delta u^k - \Delta f^k)(\Delta u^k)^T}{(\Delta u^k)^T \Delta u^k}, \quad (\Delta u^k)^T \Delta u^k \neq 0 \tag{15}
\]

is applied even in the case that no initial estimate of the sensitivity matrix is available at all, i.e. \( G_u^0 = 0 \) or \( B^0 = I \). Obviously, this update strategy (15) can be applied to \( G_u^k \) directly to get

\[
G_u^{k+1} = G_u^k \left( I - \frac{\Delta u^k(\Delta u^k)^T}{(\Delta u^k)^T \Delta u^k} \right) + \frac{\Delta v^k(\Delta u^k)^T}{(\Delta u^k)^T \Delta u^k}, \quad (\Delta u^k)^T \Delta u^k \neq 0. \tag{16}
\]

As demonstrated later on, this update strategy improves the iteration (9) significantly even in the case \( G_u^0 = 0 \).

**Block-Broyden update:** In the case that at least some submodels provide input-output sensitivity information, with

\[
G_{u_i}^{k+1} = G_{u_i}^k \left( I - \frac{\Delta u_{i}^k(\Delta u_{i}^k)^T}{(\Delta u_{i}^k)^T \Delta u_{i}^k} \right) + \frac{\Delta v_{i}^k(\Delta u_{i}^k)^T}{(\Delta u_{i}^k)^T \Delta u_{i}^k}, \quad (\Delta u_{i}^k)^T \Delta u_{i}^k \neq 0, \tag{17}
\]

a block-Broyden update of (10) is proposed which maintains the global block pattern, but is still in conflict with the sparsity pattern of sensitivity submatrices, cf. also [16]. Here \( G_{u_i}^k \) denotes the \( i \)'th diagonal block of \( G_u^k \) and
\[ \Delta v_i^k = v_i^{k+1} - v_i^k \] the vector of output variations of the submodel calculated for an input change \[ \Delta u_i^k = u_i^{k+1} - u_i^k \], at all discrete time points of the current window.

The denominator \( (\Delta u_i^k)^T \Delta u_i^k \) instead of \( (\Delta u^k)^T \Delta u^k \) ensures that this update is approximating the dynamic sensitivity matrix of submodel \( i \) independently of the input changes \( \Delta u_j^k \) \((j \neq i)\) of the other submodels. Additionally, the local condition

\[ G_{u_i}^{k+1} \Delta u_i^k = \Delta u_i^k \] (18)

is fulfilled, which gives a finite difference approximation of the sensitivity matrix of the corresponding submodel in the case of an orthogonal sequence \( \Delta u_i^k \), \((k = 1, 2,...)\).

**Relaxation Strategies:** In the case of smooth trajectories, a relaxation strategy can be introduced by formally discretizing the time delayed coupling equations (4) with an implicit Euler formula simultaneously for all internal grid points of the current window. This gives a quite natural relaxation method

\[ u^{k+1}(t_j) = (1 - \lambda_j)u(t_j^0) + \lambda_j P G(u^k), \quad \lambda_j = \frac{1}{1 + \frac{m\omega}{T^2}}, \quad j = 1(1)m \] (19)

where the desired solution is obtained for \( \omega \rightarrow 0 \), i.e., \( \lambda_j \rightarrow 1 \). This embedding technique is especially recommended if a submodel fails to initialize the integration procedure at a given inlet stream variation. Additionally, a component specific relaxation parameter \( \omega \) can be introduced, if an experienced user knows internal dependencies of the process. With

\[ u^{k+1} = u^k - (I - \Lambda P G_u(u^k))^{-1} [u^k - \Lambda P G(u^k) - (I - \Lambda)u^0] \], \( \Lambda = \Lambda(\omega) \), with \( \lim_{\omega \rightarrow 0} \Lambda(\omega) = I \).

### 3 Implementation and Results

This project is primarily intended to develop the algorithmic details of an improved waveform iteration method for large-scale dynamic process simulation. The implementation of a communication software for distributed simulation tasks are beyond the scope of this project, because our cooperation partners at Bayer AG, Leverkusen, have already developed such a software tool called *Simulation Manager* [7]. This tool is mainly used for distributed simulation tasks of stationary process models, but it also allows the distributed dynamic simulation based on a simple Picard-type iteration (7,8) with a piecewise constant approximation of the interconnecting variables. A graphical user
interface is provided for deriving the set of coupling equations by drawing the process flowsheet and for communication and control of subprocesses which can run on different computer platforms using different commercial simulation tools. We have successfully implemented dynamic processes within this environment by using the commercial simulator SPEEDUP™[2] to formulate and solve the subprocesses. Additionally, an own software environment has been implemented which is still limited to the case of spreading the subprocesses within a network of coupled workstations with a common NFS file system. This allows a more rapid testing and an immediate access to the model equations of subprocesses. The following examples have been tackled within both environments by using our block-oriented process simulator BOP [5] to solve the submodels, or the simulator SPEEDUP™ within the Simulation Manager tool, respectively.

Details of a large-scale model of three interconnected reactive distillation columns have been reported previously [10]. Here, each subprocess model consists of 600 cell models with different parameter settings, which results into an DAE system with an overall number of 45 600 differential algebraic equations. Each submodel consists of 15 200 DAEs while the total number of connecting variables is only 40. This is not an untypical ratio in industrial applications. Although the process flowsheet of this model has non-sequential reflux streams, the iteration of the Picard-type method converges quite fast, and the Broyden update does not give a significant additional progress. Obviously, the number of iterations per window and the inlet stream variations are too small, so as to alter the initial iteration matrix significantly.

A different behaviour has been observed for the process flowsheet in chapter 1 (Fig. 1), where a reactor model (Subprocess P1, 507 DAEs) is interconnected with a heat-exchanger (P2, 779 DAEs) and a simplified separation unit (P3, 48 DAEs). The plant serves for the production of 1,2-dicloroethane. Reactants are pure chlorine and ethene diluted by ethane. The exothermal reaction is performed in a tubular plug-flow reactor. To get a sufficient reaction rate, the fresh reaction mixture is preheated by the hot reactor output stream in a counter-current manner. Afterwards the reactor output stream is further cooled down in the separator (P3) to condense the desired product which leaves the condenser as liquid. The remaining gas stream is then partly returned to the input of the heat exchanger (P2). By varying input concentrations, flow rates or temperatures of both feed streams as well as by changing internal parameters, the model allows to perform different dynamic scenarios like ignition or extinguishing of the reaction as well as very rapid transitions between steady states. The exemplary curves in Figs. 2, 3 have been achieved by switching the recycle ratio within the splitter in P3. Figure 2 represents the resulting connecting stream variables between P3 and P2, while Fig. 3 shows the dynamics of the corresponding product stream leaving the reactor P1 towards the heat exchanger P2. The presented scenario shows rapid changes in the very beginning and a smooth tail towards the new steady state, and is therefore well suited for testing
purposes. In both figures a window length of 8 is indicated while the inner grid size of the windows is equal to 1 for the linear approximation of all trajectories.

Table 1 shows the comparison of the Picard-type iteration (7)-(8) versus its modification with Broyden update (16) at a constant window length of 16 seconds. In fact, the number of iterations, i.e. the number of subsequent subprocess integration steps, is significantly reduced by Broyden update in the dynamic part of the trajectories (window 1) compared with the smoother tail afterwards. Furthermore, comparative calculations have been done within the simulation interval $t \in [0,64]$ by choosing different window lengths of 1, 2, 4, 8, 16, and of 32 seconds to demonstrate the even more dramatic influence of the length of the time interval on the total number of iterations. Figure 4 shows the cumulative number of iterations needed with Broyden update (16)

<table>
<thead>
<tr>
<th>Window number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1–4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time interval (sec.)</td>
<td>0–16</td>
<td>16–32</td>
<td>32–48</td>
<td>48–64</td>
<td>0–64</td>
</tr>
<tr>
<td>Number of iterations:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>without Broyden update</td>
<td>39</td>
<td>29</td>
<td>23</td>
<td>18</td>
<td>109</td>
</tr>
<tr>
<td>with Broyden update</td>
<td>23</td>
<td>20</td>
<td>14</td>
<td>13</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 1. Convergence acceleration of Picard-type iteration by Broyden update
Fig. 4. Influence of the window length on the number of iterations

and applying this windowing strategy on the same subgrid, in order to achieve a comparable approximation accuracy. Apparently, the number of iterations is reduced remarkably to less than 10% by increasing the length of the time window from 1 to 32 seconds. This is mainly caused by the fact that the number of coupling systems to be solved iteratively as well as their dimensions are changing significantly with the window length.

The preceding results reveal the inherent potential of a heterogeneous modelling approach based on an appropriate waveform iteration technique. Basic requirements are a correct modelling of the input-output behaviour of each subprocess over small time intervals, as well as a continuous at least piecewise linear approximation of all trajectories of the connecting variables in order to avoid a permanent re-initialization during submodel integration. A well adapted Broyden update strategy may accelerate the solution procedure. Furthermore, the presented approach allows the modular treatment of subprocess model formulation as well as their parallel numerical solution.

Acknowledgements. We thank Bayer AG, Leverkusen, for the valuable support and Aspen Technology, Inc. for providing us a free SPEEDUP™ license for academic use. The assistance of P. Čapek, Institute of Chemical Technology Prague, for test model formulation is kindly acknowledged.

References

Numerical Simulation
of Annular Chromatography

Lutz Tobiska and Angela Thiele

Institut für Analysis und Numerik, Otto-von-Guericke-Universität Magdeburg,
PF 4120, 39016 Magdeburg, Germany
E-mail: lutz.tobiska@mathematik.uni-magdeburg.de,
URL: http://www.math.uni-magdeburg.de

Abstract. This project is carried out jointly with Prof. Seidel-Morgenstern from
the Institute for Process Engineering of University of Magdeburg and with the
support of Schering AG at Berlin. Preparative chromatography attains increasing
importance for the isolation and purification of value added products in the phar-
maceutical industry and in biotechnology. The goal of this project is the numerical
simulation of the separation process in an annular chromatograph to study the
influence of the main design parameters on the efficiency of the multicomponent
separation. Simulations to optimize the separation process with the aim of a high
efficiency of pure products are of special interests. On the basis of the mass balance
equations a mathematical model is proposed and an efficient solution strategy is
developed using streamline diffusion method, adaptive refinement and multigrid
solver.

1 Model Formulation

Chromatographic separation processes attain currently increasing attention
for preparative separation and purification of different products in the bio-
chemical and pharmaceutical industries, because the required purities can be
often achieved with these processes.

One way to perform a continuous separation is offered by the concept of
annular chromatography (see Fig. 1). A continuous rotating annular chro-
matograph was developed for preparative multicomponents separation [3].
Here the stationary phase, a porous medium, is fixed between two concentric
cylinders and this bed is rotating along a fixed feed port. The multicomponent
mixture is dosed continuously on top of the apparatus at one rotating point
and the mobile phase, an eluent fluid, is supplied everywhere else. The mixture
is separated on its way downwards, because of their different affinity for
the stationary phase [1]. Thus pure components can be collected continuously
at the bottom of the apparatus at different angles relative to the stationary
feed point.

The main objective of the project is the evaluation of the potential of
annular chromatography in real separation processes.
2 Mathematical Model

The annular chromatograph can be essentially considered as a plane fixed bed with periodic boundary conditions. Based on the mass conservation the mathematical model consists of a coupled system of nonlinear convection diffusion equations in two dimensions:

\[
\frac{\omega}{\partial \theta} \frac{\partial c_i}{\partial \theta} + \omega F \frac{\partial q_i(c)}{\partial \theta} + u \frac{\partial c_i}{\partial x} = D_{ax} \frac{\partial^2 c_i}{\partial x^2} + D_{tan} \frac{\partial^2 c_i}{\partial \theta^2} \quad \text{in } \Omega. \tag{1}
\]

Nonlinear adsorption isotherms according to the Multi-Langmuir equation are used to describe the equilibrium distribution between the mobile phase and the stationary phase:

\[
q_i(c) = \frac{a_i c_i}{1 + \sum_{j=1}^{n} b_j c_j} \quad i = 1, \ldots, n. \tag{2}
\]

The variables \(x\) and \(\theta\) denote the axial coordinate and the angular coordinate respectively. Also, \(c_i\) is the concentration in the mobile phase and \(q_i\) the loading on the stationary phase, \(u\) is the axial velocity, \(\omega\) the rotation velocity and \(D_{ax}, D_{tan}\) are the Dispersion coefficients in axial and tangential direction. Finally \(F\) is given by \(F = (1 - \epsilon)/\epsilon\) with the porosity \(\epsilon\).

The \(n\) components of the mixture, dosed on top of the apparatus, are separated from each other in the annulus on their way downward, because of
the different adsorption parameters \(a_i\). Dependent on these different affinities to the stationary phase helical concentration bands are formed, which are characteristic for each substance. Thus the individual components leave the annulus at characteristic fixed angular regions and can be collected at the outway. These helical concentration profiles are developed and broadened due to the unavoidable diffusion and dispersion effects. In particular, the contribution of the tangential direction on the band broadening was not evaluate in previous numerical studies. But, experimental studies already revealed that dispersion effects in an annular chromatograph are larger than in conventional packed beds. Thus, special attention has been given on the influence of the tangential dispersion to the separation process.

For completing the mathematical model (1)–(2) we apply the periodic boundary condition

\[ c_i(0, x) = c_i(2\pi, x), \]

and the inflow and outflow boundary condition

\[ c_i(\theta, H) = \begin{cases} c_{i,\text{Feed}} & : \theta \in [\gamma_1, \gamma_2] \\ 0 & : \text{otherwise}, \end{cases} \quad \text{and} \quad \frac{\partial c_i(\theta, 0)}{\partial x} = 0 \]

on top \( x = H \) and on bottom \( x = 0 \) of the chromatograph. The width of the inflow intervall \( \gamma = \gamma_2 - \gamma_1 \) will be 1 percent of the circumferential.

The main difficulties of the described model are the nonlinear coupling of the concentrations \( q_i \) and \( c_i \), small dispersion coefficients \( D_{ax} \) and \( D_{\tan} \) in front of the highest derivatives and the small width of the inflow \( \gamma \).

3 Numerical Solution

From the mathematical point of view (1)–(4) is singular perturbed with respect to \( D_{ax}, D_{\tan} \) and \( \gamma \). The streamline diffusion method with conforming piecewise linear finite elements is used as stabilized discretisation method. Adaptive mesh refinement is applied to guarantee a high accuracy of the numerical solution with reasonable numerical cost.

For simplicity the case of two components \((n = 2)\) is considered. The adsorption isotherms \( q_i \) (see (2)) describe a nonlinear connection between the two components, therefore the partial derivatives \( \partial q_i / \partial \theta \) can be expressed in the following form:

\[ \frac{\partial q_i}{\partial \theta} = \Psi_i(c_1, c_2) \frac{\partial c_i}{\partial \theta} - \Phi_i(c_1, c_2), \]

with the functions \( \Psi_i \) and \( \Phi_i \) given by (with \( i \neq j \))

\[ \Psi_i(c_1, c_2) = \frac{a_i(1 + b_j c_j)}{(1 + b_1 c_1 + b_2 c_2)^2} \quad \text{and} \quad \Phi_i(c_1, c_2) = \frac{a_i(c_i b_j \frac{\partial c_j}{\partial \theta})}{(1 + b_1 c_1 + b_2 c_2)^2}. \]
Now, we replace in (1) $\partial q_i/\partial \theta$ by (5) and get the following outer-inner iteration to solve the nonlinear system of equations (1):

**outer iteration:** $k = 1..m$ with: $c_1^0 = c_2^0 = 0$

- first inner iteration: $l = 1..n_1$ with: $c_1^{k,0} = c_1^{k-1}$

\[-D_\theta \frac{\partial^2 c_1^{k,l}}{\partial \theta^2} - D_x \frac{\partial^2 c_1^{k,l}}{\partial x^2} + \omega(1 + F\Phi_1(c_1^{k,l-1}, c_2^{k-1})) \frac{\partial c_1^{k,l}}{\partial \theta} + u \frac{\partial c_1^{k,l}}{\partial x}
\]

\[= \omega F\Phi_1(c_1^{k,l-1}, c_2^{k-1}) \tag{7}\]

Define: $c_1^k = c_1^{k,n_1}$, $c_2^k = c_2^{k,n_2}$.

- second inner iteration: $l = 1..n_2$ with: $c_2^{k,0} = c_2^{k-1}$

\[-D_\theta \frac{\partial^2 c_2^{k,l}}{\partial \theta^2} - D_x \frac{\partial^2 c_2^{k,l}}{\partial x^2} + \omega(1 + F\Phi_2(c_1^k, c_2^{k,l-1})) \frac{\partial c_2^{k,l}}{\partial \theta} + u \frac{\partial c_2^{k,l}}{\partial x}
\]

\[= \omega F\Phi_2(c_1^k, c_2^{k,l-1}) \tag{8}\]

Define: $c_2^k = c_2^{k,n_2}$.

The outer-inner iteration (7)–(8) works as follows: Starting with a fixed concentration of the second component $c_2$ the equation (7), a nonlinear convection diffusion equation, is solved in a simple iteration manner, to get a new concentration $c_1$. The same procedure is applied to (8), where the first component is fixed and the equation (8) yields a new solution for $c_2$. Between the two inner iteration steps (7) and (8) the results are always updated. Then the algorithm is repeated until the residual is within a given tolerance.

For performing the iteration method (7)–(8) described above a linear convection diffusion equation has to be solved in each inner iteration step. The streamline diffusion method [4] has been successfully applied on layer adapted grids in order to guarantee a high accuracy of the discretization with a minimal number of nodes. Since the inflow interval and the direction of the flow are known in advance, a-priori adapted meshes are a powerful refinement strategy. In this way, no additional cost arises from calculating error indicators.

The resulting large algebraic system of equations is solved by an efficient multigrid strategy [2], based on solving the discretized equations on a set of hierarchical meshes. The continuity of the finite element trial functions on the hierarchy of locally adapted meshes is achieved by the concept of hanging nodes. This allows an efficient implementation of the corresponding data structure.

### 4 Numerical Tests

The developed algorithm has been used to investigate the influence of several model parameters on typical performance criteria of the chromatographic separation process.
The separation of two components was considered. In the figures dimensionless coordinates \( x \) and \( \theta \) are used which corresponds to a height \( H = 1.0 \) and a circumference of 1. The following values for the parameters are used if not stated otherwise: \( D_{ax} = 0.00114 \), \( D_{tan} = 0.0 \), \( u = -2 \), \( \omega = 2.0 \), \( c_i,\text{Feed} = 1.0 \), \( a_1 = 1.0 \), \( a_2 = 1.2 \), \( b_1 = 1.0 \), \( b_2 = 1.0 \) and \( F = 1.0 \).

Figures 2 and 3 show the two components on their way down and at the outflow boundary. In Fig. 2 the band broadening of both components can be seen clearly. The shape of the inflow concentrations is small and high, on the way down the height of the concentration profiles decreases and the shapes become broader. Caused by the different affinity of the single components, the individual components leave the apparatus at different positions. Figure 3 shows the outlet of the annulus, there are regions with pure components and a non-separated fraction.

A typical performance criterion of a chromatographic separation process is the yield of pure substances. The yield of pure products is a measure for the efficiency of the separation process. To determine the yield of a separation process we have to calculate intervals \([\theta_{i,\text{begin}}; \theta_{i,\text{end}}]\) as pure fraction of the component \( i \). These intervals are dependent on a given purity requirement.
Numerical Simulation of Annular Chromatography

Fig. 4. Pure fractions: $i = 1$ (green), $i = 2$ (blue)

$PUR_i$. Then the absolute and relative yield of a component is determined as

$$Y_{a,i} = \int_{\theta_i, \text{begin}}^{\theta_i, \text{end}} c_i \, d\theta \quad \text{and} \quad Y_{r,i} = \frac{Y_{a,i}}{c_i, \text{Feed} \cdot \gamma} \cdot 100\%.$$  \hspace{1cm} (9)

The yield $Y_{r,i}$ describes the ratio between the collected and the dosed amount for a component $i$ in a fixed interval $[\theta_i, \text{begin};\theta_i, \text{end}]$.

With these numbers we can evaluate the simulation, shown above in Figs. 2–3 for three purity demands: $PUR_i = 90\%$, $99\%$, $99.9\%$.

Figure 4 shows the intervals for the collection of pure components for the above mentioned purity demands. The pure fraction of the first component is green and the pure fraction for the second component is blue. With a higher purity demand the width of the pure fraction decrease dramatically.

We want to optimize the separation process to obtain a high efficiency of pure products. One of the main production parameters is the rotation velocity. So, the influence of the rotation velocity $\omega$ on the separation process is investigated. Figure 5 shows simulations for various values of $\omega$, the other parameters be as before. An increasing rotation velocity leads to smaller peak maxima, broader outflow intervals and better separations. But, if the rotation speed is too high, the pure components will start to overlap again and the efficiency of separation decreases.
Fig. 5. Outflow boundaries $c_i(\theta, x = 0)$: for $\omega = 0.3, 0.8, 1.5, 3.0$

The following Fig. 6 shows the efficiency of the first component dependent on the rotation velocity for two various adsorption parameters of the second component $a_2$ for a purity demand of 90%. In both cases a maximum of the relative yield exists. Normally, a greater adsorption parameter $a_2$ yields larger efficiency, but the pure components begin to overlap for smaller rotation speeds as in the case of a smaller adsorption parameter $a_2$. For $a_2 = 1.5$ nearly complete separation is reached and yields are closed to 100% in a great range of rotation speeds.

Until now the tangential dispersion $D_{\text{tan}}$ was equal to zero in the numerical simulations. We continue to examine the effects of a nonvanishing tangential dispersion coefficient on the separation result, supposing that it is unavoidable to consider that dispersion direction for quantitative prediction.

Figure 7 shows the efficiency for the same simulations as in Fig. 6, but now for the case $D_{\text{tan}} = D_{\text{ax}}$ instead of $D_{\text{tan}} = 0.0$. The comparison between Fig. 6 and 7 confirms the effect of a nonvanishing tangential dispersion. The same tendency for increasing rotation speeds can be seen. There is a range of maximal yield and a decreasing of the yield for higher rotation speeds. Larger adsorption parameters $a_2$ yields better separations. But the efficiency becomes smaller under consideration of the tangential dispersion coefficient especially for slow rotation speeds and a small adsorption parameter $a_2$. Basically the quality of the separation process is improved with an increasing separation factor. In the case of small adsorption factors $a_2$ and nonvanishing tangential dispersion the yields drop down considerably.
Another main production parameter of the separation process in an annular chromatograph is the inflow concentration $c_{i,\text{Feed}}$. An increasing $c_{i,\text{Feed}}$ leads to a higher nonlinearity of the process (see (2)). Let the parameter of the adsorption isotherm be $a_2 = 1.5$ and of the rotation velocity $\omega = 0.5$. Figure 8 shows the outlet of the chromatograph for various inflow concentrations $c_{i,\text{Feed}}$. For a small feed concentration an almost complete separation of the components is reached. Pure but diluted fractions can be collected, and only a small region in the middle is not separated. For larger feed concentration the height of the peak maxima and also the peak deformation caused by the nonlinearity of the Langmuir isotherm increase. The non-separated region between the pure fractions increases, too. The shapes become more and more asymmetric and the retention angles move to earlier regions of $\theta$. 
Figures 9 shows the influence of the inflow concentration and the value of tangential dispersion on the relative and absolute yield. The purity demand is chosen 90 and 99.9 percent.

For higher feed concentration at the inflow the relative yield decreases for every purity demand and for all tangential dispersion coefficient. Despite of the fact, that the non-resolved fraction decreases, higher feed concentrations appear to be more attractive in terms of absolute yield. For each purity constraint and fixed tangential dispersion a maximum of absolute yield exists. With decreasing purity demand or decreasing $D_{\tan}$ the maximum yield moves to higher feed concentrations and the value of this maximum becomes higher.

In Fig. 9 the great influence of a nonvanishing tangential dispersion on the separation process can be seen. Therefore, accurate qualitative and quantitative predictions of the separation process need an exact knowledge of the tangential dispersion.
5 Conclusion

A mathematical model describing multicomponent separations in an annular chromatograph is presented and solved numerically. The model is based on a stabilized discretization method to guarantee a high accuracy of the numerical solution. A considerable reduction of computing time and used memory could be achieved by using a multigrid approach and an adaptive mesh refinement. The accuracy of the numerical results can be increased without limitations by increasing the computational work.

Tests explain important effects of various parameters as adsorption parameters, feed concentrations and rotation speeds dependent on the tangential dispersion. The numerical simulations show the great influence of the tangential diffusion on the peak profiles and the yields of the solutions, especially for difficult separations and for slow rotation speeds.

Simulations to optimize the separation process with the aim of a high efficiency of pure products are of special interests. The main design parameters, the rotation speed and the feed concentration, have a strong influence on the shape of the helical bands developing in the apparatus. The implementation of the described algorithm admits qualitative and quantitative predictions of the efficiency of the annular chromatograph for special separations.

Acknowledgements. The authors like to thank the BMBF for supporting the research through grant no. 03-T07MA1-6. The numerical simulations have been done by the software package MOONMD. The visualization package GRAPE enabled us to produce some of the figures.

References

Numerical Methods
for Parameter Estimation in Bingham-Fluids

Gabriel Wittum\textsuperscript{1}, Volker Schulz\textsuperscript{2}, Bernd Maar\textsuperscript{1}, and Dmitriy Logashenko\textsuperscript{1}

\textsuperscript{1} Interdisciplinary Center for Scientific Computing, University of Heidelberg, INF 368, 69120 Heidelberg, Germany
\textsuperscript{2} Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany

Abstract. Bingham models are frequently used for describing the flow of pastes. In this project together with Braun GmbH at Friedrichshafen, we develop a parameter estimation method for the automatic determination of certain model parameters. The result of this project is a software tool for the simultaneous determination of all model parameters by using data from a single experiment sweep.

1 Formulation of the Practical Problem

Pastes are used, e.g. in the production of bricks from clay or bodies of catalytic converters from ceramic pastes. Usually they are extruded, where the quality of the extrusion product depends on the velocity distribution of the flow within the extrusion device. Recently, substantial progress has been achieved in the development of numerical simulation techniques for paste extrusion based on Bingham models [2]. However, in practice these numerical techniques can only be used, if certain parameters of the underlying flow model are known. These parameter values cannot be accessed by direct measurements, but are determined up to now in a process involving a rather high empirical effort and using analytical approximation approaches.

Therefore we have developed parameter estimation techniques in a joint effort together with our industrial partner, Braun GmbH (Friedrichshafen), who already uses the numerical simulation techniques from [2] in the design of extrusion machines. Outlets for these extrusion machines are the main products of our industrial partner. The aim of this project is the development of a model based measurement technique, which allows the simultaneous determination of all model parameters from one experiment and which is fast enough for online-usage. Since analytical approaches are not sufficient, a numerical parameter identification method is developed and implemented as software based on the \textit{ug} toolbox [1].

In addition to the pure parameter estimation itself also a statistical sensitivity analysis of the resulting parameters is provided. This reveals that some of the parameters cannot be determined with sufficient reliability by existing measuring devices. In order to overcome this problem, geometry variations of the measuring devices are investigated. This work is still in progress and will be reported on in further publications.
2 Bingham Fluids

The non-Newtonian flow of ceramic pastes can be described by the following PDE system from continuum mechanics for the case of incompressible fluids and low Mach number velocities (s. [4]): the continuity equation

\[ \text{div } \mathbf{u} = 0, \]  

and the momentum equation

\[ \rho \frac{\partial \mathbf{u}}{\partial t} = \text{div } \mathbf{T} + \mathbf{f}, \]  

where \( \rho \) is the (constant) density of the paste, \( \mathbf{u} \) is the velocity field of the flow, \( \mathbf{f} \) is a vector of the body forces and \( \mathbf{T} \) is the stress tensor. According to the angular momentum conservation law, \( \mathbf{T} \) is symmetric. It is usually assumed that it depends only on the pressure and the strain tensor

\[ \mathbf{D} = \frac{1}{2} (\nabla \mathbf{u} + (\nabla \mathbf{u})^T). \]

In general form this dependence can be written as follows:

\[ \mathbf{T} = -p \mathbf{I} + \mathbf{T}^E(\mathbf{D}), \]  

where \( p \) is the pressure and \( \mathbf{T}^E \) the extra stress tensor describing the viscous forces in the fluid. The special choice of the extra stress tensor specifies the mathematical model of the ceramic pastes.

For the materials we are interested in, the following form of the extra stress tensor is used (s. [12,2]):

\[ \mathbf{T}^E(\mathbf{D}) = 2\mu(\mathbf{D})\mathbf{D} = 2 \left( \eta_B + \tau_F(2\mathcal{I}_D)^{-\frac{1}{2}} \right) \mathbf{D}. \]  

(Here \( \mathcal{I}_D \) is the second invariant of \( \mathbf{D} \), \( \mathcal{I}_D = \frac{1}{2}(\text{Tr } \mathbf{D}^2 - (\text{Tr } \mathbf{D})^2) \). But this form is assumed to be true if and only if

\[ |\mathcal{I}_{\mathbf{T}^E}| > \tau_F^2. \]  

Otherwise, i.e. if the inner stresses are small enough, the material is rigid and does not flow so that

\[ \mathbf{D} = 0 \quad \text{for } |\mathcal{I}_{\mathbf{T}^E}| \leq \tau_F^2. \]  

Material whose flow is described by equations (1–2) with the stress tensor (3–4), under the condition (5), and (6) is called a Bingham fluid. The parameters \( \eta_B \) and \( \tau_F \) are said to be Bingham viscosity and yield stress respectively.

We shall consider only flow of ceramic pastes and assume that condition (5) is true respectively. Besides we shall consider only stationary flow, so that the time derivative in (2) is zero. The gravitation, being the only body
force, does not play any essential role in our situation as well, because it is negligible in comparison with the viscous forces. Thus, further we omit also \( \mathbf{f} \). This simplification leads to the following system describing our situation:

\[
\begin{align*}
\text{div} \, \mathbf{u} &= 0, \\
-\text{div} \left( 2\mu(\mathbf{D}) \cdot \mathbf{D} \right) + \nabla p &= 0.
\end{align*}
\] (7)

In this work, we consider only a two-dimensional model.

We would encounter difficulties when trying to get an approximate solution of system (7). The matter is that the extra stress tensor tends to infinity as the term \( \mathcal{II}_D \) tends to zero. In reality this is not the case for the restriction (5) “switching” to the other equation, but omitting this condition we always come to such a situation. To avoid it, system (7) should be regularized. This means that we involve a small parameter \( \delta \) in the generalized viscosity function \( \mu \) so that

\[
\mu(\mathbf{D}) = \eta_B + \tau_F(\delta + 2\mathcal{II}_D)^{-\frac{1}{2}}. 
\] (8)

In this way we get an approximation of the system (7) at which \( \mu \) is bounded. The solution of the regularized system deviates from the unregularized Bingham solution. An exploration of the regularization error in a model problem is carried out in [2]. There, it is shown that this error tends to zero like \( O(\sqrt{\delta}) \).

**Remark:** As for incompressible fluids \( \text{Tr} \mathbf{D} = 0 \), \( \mathcal{II}_D = \frac{1}{2} \text{Tr} \mathbf{D}^2 \). Thus (8) attains the form

\[
\mu(\mathbf{D}) = \eta_B + \tau_F(\delta + \text{Tr} \mathbf{D}^2)^{-\frac{1}{2}}. 
\] (9)

The stress tensor (3) attains the form

\[
\mathbf{T} = -p\mathbf{I} + 2\mu(\mathbf{D})\mathbf{D}. 
\] (10)

In addition to these equations, the model of the ceramic pastes requires special boundary conditions. The reason for this is a phenomenon called wall sliding. Microscopically we have a two-phase flow on the boundary yielding a lubrication effect. Macroscopically this is described by sliding, resulting in a Navier-Type boundary condition.

\[
\mathbf{n}^T \mathbf{T} \mathbf{t} = k \mathbf{u}^T \mathbf{t} + \text{sign}(\mathbf{u}^T \mathbf{t}) \cdot \tau_G, \\
\mathbf{u}^T \mathbf{n} = 0. 
\] (11)

where \( \mathbf{t} \) and \( \mathbf{n} \) are the unit normal and unit tangent vectors to the boundary respectively. These condition include two additional scalar parameters: a wall sliding factor \( k \) and a sliding limit \( \tau_G \).

Thus for the mathematical description of the flows of the ceramic pastes we have the system of PDEs (7) with the boundary conditions (11) on walls. (Besides inflow and outflow boundary conditions can be imposed on some parts of boundary). The whole model involves four parameters: \( \eta_B, \tau_F, k \).
and $\tau_G$. The aim of the parameter estimation procedure described here is to find these values by using experimental data obtained by the device described in the next section. The nonlinearity of the PDE system together with its boundary conditions poses quite a challenge to the numerical treatment of the resulting parameter identification problem. It is necessary to note as well that the pressure $p$ is defined by this system only up to a constant. Methods for the discretization of this system and numerical solution of the resulting nonlinear discrete equations are considered in detail in [2,12].

3 Parameter Identification Technique

The parameters are estimated by using data obtained from a device whose scheme is shown on Fig. 1. This is a conical channel with rigid walls. The paste is pressed through this channel with constant velocity in the direction of the large arrow. During this process we measure the normal stress at seven fixed points on the upper wall (further referred to as measurement points). The values of the normal stress, as well as the inflow velocity of the paste, are then used for the parameter identification.

For the experiments we used the device with the following sizes: $H = 30\,\text{mm}$, $h = 10\,\text{mm}$, $L = 244\,\text{mm}$, so $\alpha = 2.35^\circ$. The inflow velocity was about $80\,\text{m/s}$ (s. the experimental data below). Fig. 2 shows a photo of the device.

To model the flow of the paste inside the device we consider the interior (the polygon) as a region $\Omega$ on which we impose the PDE system (7) with the generalized viscosity function (9). On the part $\Gamma_0 \subset \partial\Omega$ of the boundary corresponding to the rigid walls we assume boundary conditions (11). On the inflow boundary $\Gamma_{in}$ we impose Dirichlet boundary conditions for the velocity specifying a parabolic inflow profile with average inflow velocity $v_0$. On the outflow boundary $\Gamma_{out}$ we impose zero vertical velocity as boundary condition and require $\int_{\Gamma_{out}} p\,ds = \text{const.}$ For every set $q = (\eta_B, \tau_F, k, \tau_G)^T$ of four parameters this system defines the velocity field $u$ and the pressure

![Scheme of the measurement device](image)

Fig. 1. Scheme of the measurement device
Further we write all the partial differential equations comprising the system (7) and the boundary conditions in the form

\[ c(u, p, q) = 0. \]  

Denote the measurement points by \( P_1, P_2, \ldots, P_K \) (in our case \( K = 7 \)), in the order from the outflow. For every \( P_i \) we have a measured value \( \hat{\pi}_i \) of the normal stress. In the same time, for every given set \( q \) of the parameters we can get the fields \( u \) and \( p \) from equation (12) and compute the normal stress \( \pi_P(u, p, q) \) at every point \( P \in \Gamma_0 \):

\[ \pi_P(u, p, q) = \mathbf{n}_P^T \mathbf{T}_P(u, p, q) \mathbf{n}_P, \]

where \( \mathbf{n}_P \) and \( \mathbf{T}_P(u, p, q) \) respectively are the unit normal vector to the boundary and the stress tensor (3) at the point \( P \). The “correct” parameters are then determined as the solution of the nonlinear constrained optimization problem

\[
\begin{align*}
  f(u, p, q) &= \sum_{i=2}^{K} \frac{1}{\sigma_i^2} \left( (\pi_{P_i}(u, p, q) - \pi_{P_1}(u, p, q)) - (\hat{\pi}_i - \hat{\pi}_1) \right)^2 \\
  \text{s. t.} \quad c(u, p, q) &= 0,
\end{align*}
\]

where \( \sigma_i = 0.08(\hat{\pi}_i + \hat{\pi}_1) \) are the standard deviations for the difference evaluations, if all measurements are assumed to be independently normally distributed with expectation \( \hat{\pi}_i \) and standard deviation \( 0.08 \hat{\pi}_i \). Although the model defines the normal stress only up to a constant the differences of the
stresses are defined exactly and should approximate the differences of the measured normal stresses. The numerical solution is carried out in a direct approach, i.e. by discretization of the model and the objective functional. This leads to a finite dimensional non-linearly constrained optimization problem of a very large size that requires the application of structure exploiting methods to reduce the computation time.

4 Discretization of the Model

We discretize the PDE system (7) by the finite volume method from [2]. The application of this method to Bingham equations is similar to the case of a simple Stokes equation and inherits the similar problems. The implementation is based on ug-modules described for Navier–Stockes equations in [5].

We discretize the velocity and pressure fields on quadrilateral elements using a collocated scheme. Such discretizations require stabilization (cf. [6]). For the stabilization the idea described in [7] is used. This technique introduces a new term into the continuity equation, yielding:

\[-\gamma^2 \text{div } \nabla p + \text{div } u = 0\]

with a scalar stabilization parameter \( \gamma \). The influence of this stabilization term to the discretization of the Bingham model is discussed in [2]. The stabilization from [7] does not require any additional stabilization parameters.

Discretizing the outflow boundary conditions we use so called natural boundary conditions of our finite-volume scheme, namely \( \int_{r_{\text{out}}} p \, ds = \text{const.} \). That determines the pressure completely. The constant is given by the stabilization.

Further we shall denote the whole discretized system by

\[ c_h(u, p, q) = 0. \] (14)

Here \( u \) and \( p \) are respectively two- and one-dimensional grid functions defined on the same index sets. We underline that this is a non-linear system for all variables. Further it will be also convenient to denote a set \((u, p)\) by the single letter \( x \). So we shall write \( c_h(x, q) \) instead of \( c_h(u, p, q) \).

Since we aim at fast solution techniques we employ multigrid methods for the solution of (14). However, the direct application of multigrid methods (with an ILU smoother) to the linearized system from a Newton approach fails. Therefore we rewrite the discretized system (14) in the form

\[ A(x, q) x = f, \] (15)

where \( A \) is a sparse matrix. Based on this form we can easily apply a fixed-point method with inner linear multi-grid solvers.
The grid levels were obtained by uniform refinement of a coarser grid containing less than hundred quadrilaterals. The matrices on each grid level were constructed by discretization of the PDE system. In Fig. 3 we show an example of numerical results obtained by this discretization.

For the discretization of the whole system the strain tensor \( D \) and the generalized viscosity function were also defined using finite dimensional grid functions \( u \). This allows to discretize the normal stress functions \( \pi_{P_i} \). We denote the discrete variants by \( \pi_{h,i} \) respectively. This completes the discretization of the whole optimization problem which now reads

\[
f_{h}(x, q) = \sum_{i=2}^{K} \frac{1}{\sigma_i^2} \left( (\pi_{h,i}(x, q) - \pi_{h,1}(x, q)) - (\tilde{\pi}_i - \tilde{\pi}_1) \right)^2 \rightarrow \min
\]

s. t. \( c_h(x, q) = 0 \).

(16)

This is a nonlinearly constrained finite dimensional optimization problem. In the next section we describe a method for its approximate solution.

5 The Optimization Procedure

Here we consider the optimization problem (16) in the abstract form, omitting the subscript \( h \):

\[
f(x, q) \rightarrow \min,
\]

s. t. \( c(x, q) = 0 \),

(17)

with \( f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R} \) and \( c : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^n \) and the Jacobian, \( \mathbf{J} = \frac{\partial \mathbf{c}}{\partial \mathbf{x}} \), which is assumed to be nonsingular. In our case the number of parameters is 4. However, \( n \), the dimension of the grid functions, can be very large.

For the solution of the problem (17) we use a reduced SQP method. A detailed discussion of this approach can be found in [9,3]. Here we sketch only the idea.
Reduced SQP methods are related to projected Lagrangian methods (cf. [10]) and are most advantageous in the case that the number of degrees of freedom (here the parameters) is small compared to the number of state variables. The constraints are linearized by a Taylor expansion up to first order terms, so that all steps \((\Delta x, \Delta q)\) lie in the tangent space of \(c\) of the current approximation \((x, q)\):

\[
c(x, q) + J(x, q) \Delta x + \frac{\partial c}{\partial q}(x, q) \Delta q = 0.
\]

Then the problem is projected to this tangent space and approximated by a quadratic problem with the projected Hessian of the Lagrangian

\[
L(x, q, \lambda) = f(x, q) - \lambda^T c(x, q).
\]  

In this formulation the algorithm reads:

**Algorithm 1**: The RSQP method.

1. Set \(k := 0\); start at some initial guess \(x_0, q_0\).
2. Compute the adjoint variables from the linear system
   \[
   J^T(x_k, q_k) \lambda_{k+1} := \nabla_x f(x_k, q_k);
   \]
   compute the reduced gradient
   \[
   \gamma_k := \nabla_q f(x_k, q_k) - \left( \frac{\partial c}{\partial q}(x_k, q_k) \right)^T \lambda_{k+1};
   \]
   determine some approximation \(B_k\) of the projected Hessian of the Lagrangian.
3. solve \(B_k \Delta q_k = -\gamma_k\).
4. compute step on \(x\) from the linear system
   \[
   J(x_k, q_k) \Delta x_k := -\frac{\partial c}{\partial q}(x_k, q_k) \Delta q_k + c(x_k, q_k).
   \]
5. Set \(x_{k+1} := x_k + \Delta x_k, q_{k+1} := q_k + \Delta q_k\).
6. \(k := k + 1\); go to (1) until convergence.

The computationally expensive operation of evaluation of the projected Hessian is avoided by using update formulas: at the first iteration some initial approximation, for example

\[
B_0 = \alpha I
\]

with a positive scalar \(\alpha\), is taken. Then every next approximation is computed from the previous one by a formula

\[
B_{k+1} = B_k + \text{Update}(B_k, s_k, v_k).
\]

There exist different update strategies (s. [10,3]), for instance the BFGS update formula reads:

\[
\text{Update}(B, s, v) = \frac{vv^T}{v^Ts} - \frac{(Bs)(Bs)^T}{s^TBs}.
\]

For the arguments of this update one can take vectors

\[
s_k := q_k - q_{k-1}, \quad v_k := \gamma_k - \gamma_{k-1}.
\]
It can be proven that under mild conditions the reduced SQP method described by Algorithm 1 with the BFGS update formula (20–22) shows 2-step superlinear local convergence (s. [3]).

In the Algorithm 1, it is necessary to invert the Jacobian $J$ of the constraints. As we mentioned in the previous section, this approach is not recommendable. The following consideration allows to replace $J$ with the fixed point iteration matrix $A$.

To this end we incorporate the reduced SQP algorithm within the previously mentioned fixed point iteration instead of a Newton iteration – in the spirit of inexact reduced SQP methods as considered in [9]. Carrying out the fixed point iteration instead one obtains an algorithm that does not involve inversion of $J$ or $J^T$. Nevertheless the Jacobian is still used for computation of the correct adjoint variables and the correct state increments in the sense of defect correcting iterations:

**Algorithm 2:** The RSQP method with an approximate Jacobian.

(0) Set $k := 0$; start at some initial guess $x_0, q_0$.

(1) Compute the increment of the adjoint variables from the linear system

$$A^T(x_k, q_k) \Delta \lambda_k := \nabla_x f(x_k, q_k) - J^T(x_k, q_k) \lambda_k;$$

compute the reduced gradient

$$\gamma_k := \nabla_q f(x_k, q_k) - \left( \frac{\partial c}{\partial q}(x_k, q_k) \right)^T (\lambda_k + \Delta \lambda_k);$$

determine some approximation $B_k$ of the projected Hessian of the Lagrangian.

(2) solve $B_k \Delta q_k = -\gamma_k$.

(3) compute step on $x$ form the linear system

$$A(x_k, q_k) \Delta x_k := -\frac{\partial c}{\partial q}(x_k, q_k) \Delta q_k + c(x_k, q_k).$$

(4) Set $x_{k+1} := x_k + \Delta x_k$, $q_{k+1} := q_k + \Delta q_k$ and $\lambda_{k+1} = \lambda_k + \Delta \lambda_k$.

(5) $k := k + 1$; go to (1) until convergence.

A step of this method can be also interpreted as an approximate Newton step for the necessary conditions of extremum for the problem (17) since the updates of the variables are computed according to a linear system

$$
\begin{pmatrix}
0 & 0 & A^T \\
0 & B_k & \left( \frac{\partial c}{\partial q} \right)^T \\
A & \frac{\partial c}{\partial q} & 0
\end{pmatrix}
\begin{pmatrix}
\Delta x \\
\Delta q \\
\Delta \lambda
\end{pmatrix}
= \begin{pmatrix}
\nabla_x L \\
\nabla_q L \\
-c
\end{pmatrix}.
$$

Some notes on the convergence properties of the reduced SQP methods can be found in [9]. The positive definiteness of the approximations $B_k$ should be preserved. The BFGS update formula (20–21) yields a positive definite $B_{k+1}$ for a positive definite $B_k$ if and only if $v_k s_k > 0$ (s. [10]). In the RSQP method started with an arbitrary initial guess this property can be violated that can lead to indefinite approximations of the Hessian. As a remedy we used the Powell modification of the BFGS update formula [11].
The next important issue is that these methods converge only locally and starting with an arbitrary initial approximation Algorithm 2 always requires some damping in step 4. The damping factor is usually chosen by a line search. It is required that the damping parameter supplies sufficient descent of a merit function that involves the constraint as well as the value of the objective function. This is done to provide also the approximate feasibility of the guess. We used the merit function

\[ \psi(x, q) = f(x, q) + \sum \mu_i |c_i(x, q)|, \]  

(24)

where \( \mu_i > 0 \) are estimated upper bounds for the absolute values of the adjoint variables \( |\lambda_i| \) at the optimal solution. This merit function supplied good convergence properties up to the practical precision.

For the computations we implemented Algorithm 2 with the Powell modification of the BGFS update formula. The initial approximation to the Hessian was \( B_0 = I \). For the initial guess of the parameters we took some reasonable values. As the initial guess for \( x \) we took an approximate solution of the constraint for the initial parameters obtained by the fixed point iteration. For the solution of the large sparse linear systems with the matrices \( A \) and \( A^T \) we applied the above mentioned multigrid method with ILU smoothing. This provided an approximate feasibility in the first iteration. The initial values of the adjoint variables were merely \( \lambda_0 = 0 \). In the next section we describe the results of the numerical experiments.

6 Experiments

Here we show results of one of the experiments. For this experiment we used the device with the following sizes: \( H = 30 \text{ mm}, h = 10 \text{ mm}, L = 244 \text{ mm}, \alpha \approx 2.35^\circ \) (s. Fig. 1). The paste was pressed through the device at the average inflow velocity \( v_0 = 80 \text{ mm/s} \). The measured normal stresses on the upper wall are listed in the Table 1. The measurement points are numbered here from the outlet. The precision of the measurements is \( \pm 8\% \).

<table>
<thead>
<tr>
<th>Measurement point</th>
<th>( P_1 )</th>
<th>( P_2 )</th>
<th>( P_3 )</th>
<th>( P_4 )</th>
<th>( P_5 )</th>
<th>( P_6 )</th>
<th>( P_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal stresses (bar)</td>
<td>2.66</td>
<td>5.2</td>
<td>7.45</td>
<td>9.3</td>
<td>11.2</td>
<td>12.4</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Table 1. Measured normal stresses

Because of the comparatively low measurement precision the normal stresses obtained from the model do not coincide with the measured ones. The measured stresses do not correspond to real material parameters, so the minimum value of the objective function that can be expected is not equal to 0.
Fig. 4. The differences of the normal stresses. Empty circles denote the relative measured stresses ($\pi_i - \hat{\pi}_i$), and the black circles show the stresses $\pi_{h,i}(x, q) - \pi_{h,1}(x, q)$ computed for the final guess $(x, q)$ of the optimization procedure. The difference $\pi_{h,1}(x, q) - \hat{\pi}_1$ is 11.22.

The optimization procedure was started with the initial parameters

$$\eta_B = 0.0005 \text{ bar} \cdot \text{s}, \quad \tau_F = 3 \text{ bar}, \quad k = 0.5 \frac{\text{bar} \cdot \text{s}}{\text{m}}, \quad \tau_G = 0.15 \text{ bar}$$

on a grid with 969 nodes (896 quadrilaterals) on the finest level. The initial guess of the pressure and the velocity field was obtained by the fixed point iteration. As the stopping criterion we used the relative variation of the merit function: the procedure halted after iteration $k$ if

$$\frac{\psi_k - \psi_{k-1}}{\psi_{k-1}} < 0.01.$$

where $\psi_i$ is the value of the merit function after $i$-th iteration and the Euclidean norm of the constraint (the PDE) is below 0.001. In this experiment the procedure stopped after 12 iterations. The final value of the objective function was 0.034332, the derivative w. r. t. $\eta_B$ = --0.0057970 and the derivative w. r. t. $\tau_F$ = --0.00073115. The final Euclidean norm of the constraint was 0.000565912. The final parameters are

$$\eta_B = 0.515479 \text{ bar} \cdot \text{s}, \quad \tau_F = 2.89747 \text{ bar}, \quad k = 0.139043 \frac{\text{bar} \cdot \text{s}}{\text{m}}, \quad \tau_G = 0.275458 \text{ bar}.$$

The normal stresses computed on the final approximation of the pressure and the velocity in comparison with the measured ones are depicted in Fig. 4. As one can see they have the appropriate precision.

This numerical experiment have been preformed on a SGI Indigo 2 computer with a R4400 Processor (200 MHZ). The cpu-time has been 767 s.
A simple computation of the velocity field and the pressure in the measuring device for the initial parameters and the same grid took 318 s. That means, our simultaneous approach to parameter identification needs only approximately twice as much time as a pure solution of the model equations.

But the numerical experiments also show that one can obtain quite similar normal stresses evaluations for quite different sets of model parameters. This is reflected in the 95%-confidence intervals for the parameters: $\eta_B$ is determined up to a value $\pm 0.0764$, $\tau_F$ up to $\pm 0.2416$, $k$ up to $\pm 2.416$ and $\tau_G$ up to $\pm 0.2416$. That means especially $k$ and $\tau_G$ cannot be well determined by the current measurements.

As a remedy one can propose to change the shape of the measuring device. A more complicated geometry providing greater accuracy of the computations can be found by form optimization. This subject is currently carried out and will be reported on in the nearest future.

References

A Viscoelastic Turbulence Model
Based on Renormalization Group Theory

M. Niggemann, M. Holzmann, D. Schmidt, and K. Soldner

Fachhochschule Wuerzburg-Schweinfurt, Germany

Abstract. The aim of this project, which is in cooperation with DaimlerChrysler AG, Rhodia Acetow AG, the Steinbeis transfer center HTCO and Prof. H. Kielhoefer at the University of Augsburg, is to improve the numerical simulation of turbulent separated flows. Flows of that type arise in a variety of technically important situations. An example from automotive industry is the flow behind a car. By means of the renormalization group theory (RNG), we have constructed a Reynolds stress model with model coefficients calculated from theory. A first version of our turbulence model has been tested with the aid of the finite element code FIDAP with encouraging results for the flow over a backward facing step. Within this project, we have also addressed some more theoretical aspects, ranging from questions regarding the “epsilon expansion” up to the question of existence and regularity of solutions of the stochastically forced Navier Stokes equation. The latter work, which will be essential for a mathematically rigorous foundation of our model, has been done by Dipl.-Math. Ch. Gugg at the University of Augsburg. The ultimate test case for our model shall be the flow behind a car with data from wind tunnel experiments provided by the DaimlerChrysler AG.

1 Introduction

A Reynolds stress model being quadratic in the mean strain rate and additionally containing its Oldroyd derivative has been shown to arise naturally from the mode reduction process, when renormalization group theory (RNG) is applied up to second order in the “ε expansion” to the suitably stochastically forced Navier Stokes equations. The mathematical structure of the resulting anisotropic turbulence model is similar to the two-scale DIA model of Yoshizawa [23], but our model coefficients have been explicitly calculated from theory and they depend smoothly on the local turbulence Reynolds number. So the model may be used over a wide range of different flow regimes. In the high Reynolds number limit, our model becomes identical with the constitutive equation of a viscoelastic fluid (second order fluid), thus giving support to a proposal of Rivlin [16], that a (fully developed) turbulent Newtonian fluid may be regarded as a non-Newtonian fluid. In this case our turbulence model is similar to the model of Speziale [18], which was established on a purely constitutive basis with empirically adjusted coefficients. The closure of our Reynolds stress model is up to now achieved by means of the RNG $K - ε$ model of Yakhot and Orszag [22]. From a methodological point of view, there are some similarities between our work and the work of...
Rubinstein and Barton [17], so our result may be compared with theirs. Our model does contain convective transport terms of the strain rate tensor, which are absent from the model of Rubinstein and Barton, but which they state to be essential in the model of Speziale and which are surely important in the low-Reynolds number regime; their model fails to fulfill the requirement of material objectivity from continuum mechanics, ours does; and at last, where comparable, our model coefficients differ from theirs in magnitude. Finally we want to note that our model is also applicable to turbulent flow of viscoelastic fluids, e.g. dilute polymer solutions (as far as they can be described as second order fluids).

The paper is organized as follows. In Sect. 2 we present our model, Sect. 3 is devoted to the application of RNG to the model equations, in Sect. 4 we present our high-Reynolds number result, which we believe to be universal and in Sect. 5 we examine a numerical test case.

2 The Model

2.1 The Model in Physical Space

Our aim is to establish a model for strained turbulence. Decomposing the instantaneous velocity field \( u \) into a mean part \( U \) and a fluctuating part \( u' \),

\[
u = U + u',
\]

we assume that the turbulence is driven by the mean flow, i.e. that the velocity fluctuations \( u' \) are generated by the strain rate of the mean flow. The model will be capable of general mean flows, provided that the induced turbulence may be regarded as locally homogeneous. Then we may imagine the whole flow domain as being covered by the union of sufficiently small balls and, because of the local homogeneity, one-point-correlations of the fluctuating field or its derivatives (like the turbulent kinetic energy per unit mass, \( K \), or the energy dissipation rate per unit mass, \( \epsilon \), which will be needed in our model) are constant with respect to the space variables in each ball. We want to model the turbulence in a ball by assuming that this model possesses a common (universal) structure for all balls. The variation of \( K \) and \( \epsilon \) from ball to ball is then taken into account by means of suitable transport equations for these quantities (by now, we use the RNG \( K - \epsilon \) model of Yakhot and Orszag [22] for this purpose). For modelling the turbulent flow within a ball, we adopt the equivalence principle of Yakhot and Orszag [22] and assume that the effect of the mean strain on the small scale eddies can be imitated by a suitably chosen artificial random force \( F \), i.e. that the small scale turbulence may be equivalently generated by \( F \). We imagine the ball as blown up to full space, and then our starting point is the forced incompressible Navier Stokes equation

\[
\begin{align*}
    u_t + (u \cdot \nabla) u &= -\frac{1}{\rho} \nabla p + \nabla \cdot S + F, \\
    \nabla \cdot u &= 0
\end{align*}
\] (1)
with the extra stress tensor $S$,

$$S_{ij} = \nu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)$$

and with an isotropic random force (per unit mass) $F$ chosen as to be compatible with Kolmogorov scaling in the inertial range of turbulence. Based on the work of Forster-Nelson-Stephen [7] and Fournier and Frisch [8], Yakhot and Orszag [22] applied a simplified version of RNG theory (without rescaling) to (1). RNG enabled them to successively remove the small scales from the equation, thereby taking their effect on the large scales into account by means of an effective viscosity $\nu_{eff}$. Thus they established on a systematic way a turbulence model of the eddy viscosity type for prediction of the large scale flow field. For deriving our result, we follow the ideas of Yakhot and Orszag, but in our analysis we do not use the simplified RNG procedure as mentioned above, but instead the full Wilson-type RNG theory as developed for the dynamical case by Ma and Mazenko [12]. In deriving their result, Yakhot and Orszag (see [22] and [14]) treat the small scale flow field iteratively, which amounts to the use of a perturbation expansion with a Reynolds number $\overline{\lambda}$ characteristic of the eliminated eddies being the expansion parameter. Their expansion then stops at first order in $\overline{\lambda}$. In the work presented here, we have taken into account also the terms of the order $O(\overline{\lambda}^2)$. Furthermore, in the course of the elimination process for small scale eddies, which is performed in wavenumber-frequency space, the kernelfunctions of several convolution-type integrals have to be expanded with respect to wavenumber $k$ and frequency $\omega$. Here, Yakhot and Orszag used an approximation being of order 1 in $k$ and of order 0 in $\omega$. For our purposes, it proves to be necessary to take an approximation being 2nd order in $k$ and 1st order in $\omega$. Using the higher order expansions with respect to $\overline{\lambda}$, $k$ and $\omega$, we are faced with the fact that the system (1)/(2) does not remain structurally invariant under the RNG procedure, but that additional terms quadratic in the velocity field and its derivatives are generated. On the other hand, if (2) is imbedded in the more general structure

$$S_{ij} = \nu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \frac{\partial^2 u_i}{\partial t \partial x_j} + \frac{\partial^2 u_j}{\partial t \partial x_i} +$$

$$+ C_1 \left( \sum_k \frac{\partial^2 u_i}{\partial x_k \partial x_j} + \sum_k \frac{\partial^2 u_j}{\partial x_k \partial x_i} \right) +$$

$$+ (C_1 + C_2) \left( \sum_k \frac{\partial u_i}{\partial x_k} \frac{\partial u_k}{\partial x_j} + \sum_k \frac{\partial u_j}{\partial x_k} \frac{\partial u_k}{\partial x_i} \right) +$$

$$+ C_3 \sum_k \frac{\partial u_k}{\partial x_i} \frac{\partial u_k}{\partial x_j} + C_4 \sum_k \frac{\partial u_i}{\partial x_k} \frac{\partial u_j}{\partial x_k},$$

invariance under the RNG procedure within the order of our approximation is achieved. Then elimination of the fluctuating small scale field induces correc-
tions of the parameters \((\nu, C_0, \ldots, C_4)\). After having eliminated all fluctuating modes, the parameters \((\nu, C_0, \ldots, C_4)\) have approached their "fixed point" values (see 4.1) which may then be used in (3) as the model coefficients in the high Reynolds number limit; moreover, the velocity field in (3) has to be interpreted as being the mean field \(U\). Defining \(S'_{ij} = S_{ij} - \frac{1}{3}(tr S_{ij})\delta_{ij}\), we incorporate the isotropic part of \(S_{ij}\) into the pressure. Then the system (1)/(3), with \(S_{ij}\) replaced by \(S'_{ij}\) and with \(u\) replaced everywhere by \(U\), constitutes a mean field equation, which is of the type of the above mentioned Yoshizawa model [23]. As will be seen in 3.3 and 4.2, the model coefficients are subject to the symmetry constraints \(C_0 = C_1\) and \(C_1 + C_2 = (C_3 + C_4)/2\), which are imposed by the principles of Galileian invariance and material objectivity, respectively. Furthermore, \((\nu, C_0, \ldots, C_4)\) will be shown in 4.1 to depend on \(K\) and \(\epsilon\), so the RNG \(K - \epsilon\) model will be used for closing the mean field system.

2.2 The Model in Fourier Space

Rearranging the equations (1) and (3) and using \(\nabla \cdot u = 0\), we get

\[
\frac{\partial}{\partial t} u - \nu \Delta u - C_0 \frac{\partial}{\partial t} \Delta u = F - \frac{1}{\rho} \nabla p - (u \cdot \nabla) u + \nabla \cdot \overline{S} \tag{4}
\]

where \(\overline{S}_{ij} = S_{ij} - \nu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - C_0 \left( \frac{\partial^2 u_i}{\partial \partial x_j} + \frac{\partial^2 u_j}{\partial \partial x_i} \right)\) with \(S_{ij}\) from (3).

Without loss of generality, we may assume the force field \(F\) to be divergence free (a possible gradient component may be incorporated in \(p\)). Taking the divergence in (4), we get the corresponding Poisson equation for the pressure \(p\) (PPE). Now we formally take the space-time Fourier-transform of (4) together with the PPE, and using the latter for eliminating the pressure from the equation of motion, yields to

\[
L u_l := (i\omega + [\nu + i\omega C_0] |k|^2) u_l = f_l + \lambda \sum_{j,m=1}^{d} M_{ijm} N_{jm}(u, u) \tag{5}
\]

for the \(l\)-th component of the Fourier-transformed velocity field \(u\) \((l = 1, \ldots, d;\) in this paper, the focus is on \(d = 3\)). Here, \(f = f(\omega, k) = \tilde{F}(\omega, k)\) is the space-time Fourier-transform of the force field \(F\), \(\lambda\) has been introduced for later use as an expansion parameter (finally, we have to put \(\lambda = 1\) again), the operator \(M_{ijm}(k)\) is given by

\[
M_{ijm}(k) = \frac{1}{2i} (k_j D_{lm}(k) + k_m D_{ij}(k)) \tag{6}
\]

with \(i = \sqrt{-1}\) and

\[
D_{lm}(k) = \delta_{lm} - \frac{k_l k_m}{|k|^2} \tag{7}
\]
being a projection operator, which projects on divergence free vector fields; it stems from the elimination of the pressure, and \( M_{ljm} \) additionally contains the Fourier-transform of the divergence operator on the right-hand side of (4). Note that throughout this section \( u = u(\omega, k) \) denotes the space-time Fourier-transform of the instantaneous velocity field. Furthermore, in the linear operator \( L \) we have introduced the complex viscosity \( \nu^* = \nu + i\omega C_0 \), and the nonlinearity \( N_{jm}(u, u) \) is defined by the bilinear operator \( N_{jm}(u, v) \),

\[
N_{jm}(u, v) = u_j \ast v_m + C_1 A_{jm}(u, v) + C_2 B_{jm}(u, v) + C_3 C_{jm}(u, v) + C_4 D_{jm}(u, v)
\]  

(8)

with

\[
A_{jm}(u, v) = \sum_k \{ k_m k_k (u_k \ast v_j) + k_j k_k (v_k \ast u_m) \}
\]

(9)

\[
B_{jm}(u, v) = \sum_k \{ k_m k_m u_j \ast v_k + k_k v_m \ast k_j u_k \}
\]

\[
C_{jm}(u, v) = \sum_k \{ k_j u_k \ast k_m v_k \}
\]

\[
D_{jm}(u, v) = \sum_k \{ k_k u_j \ast k_m v_m \},
\]

which are built up by integrals of convolution type, e.g.

\[
(u_j \ast v_m)(\omega, k) = \frac{1}{(2\pi)^d+1} \int \int \int_{\mathbb{R}^d} u_j(\omega, k) v_m(\omega - \omega', k - k') d\omega' dk'.
\]

(10)

Following Forster-Nelson-Stephen [7] and Yakhot Orszag [22], we choose the random stirring force \( F \) as being Gaussian distributed with mean value 0, homogeneous and isotropic in space and white noise in time. The assumptions on \( F \) imply the following structure of the two-point correlation function of \( f \):

\[
\langle f_i(\omega, k) f_j(\omega', k') \rangle = 2 (2\pi)^{d+1} W(|k|) D_{ij}(k) \delta(\omega + \omega') \delta(\omega + k').
\]

(11)

Here, the Dirac \( \delta \)-distributions guarantee homogeneity in space and time (the latter being equivalent to statistical steadiness), the expression \( W(|k|) D_{ij}(k) \) with the projector \( D_{ij}(k) \) according to (7) results from isotropy, and the fact, that \( W \) does not depend on \( \omega \), reflects \( F \) being white noise in time. From (11) we see, that the function \( W(|k|) \) may be understood as being (up to a numerical factor) the spectral density of the “strength” of the random force \( F \),

\[
\int \int \int_{\mathbb{R}^d} W(r) r^{d-1} dr;
\]

this last integral measures the energy input per unit mass and time, which is caused by the action of the stirring force \( F \) on the fluid. Like DeDominicis and Martin [2] and Yakhot–Orszag [22], we choose the power law

\[
W(|k|) = W_0 |k|^{-y}
\]

(12)
with $W_0$ and $y > 0$ being specified in 3.5 and 4.1, respectively. The singularity of $W(|k|)$ at $k = 0$ reflects the fact, that the energy input into the system stems from the large scale eddies (i.e. the mean flow field in our picture).

3 The RNG Procedure

The RNG method has been established and applied to critical phenomena in statistical mechanics by K. Wilson in the seventies [20] (see also e.g. [6] or [10]). Its extension to dynamical systems is due to Ma and Mazenko [12]. RNG is a coarse-graining procedure, which may be used in the context of turbulence for deriving equations governing the large scale dynamics of the flow. It consists in principle of two major steps: i) elimination of the small scale eddies (i.e. the short-wavelength modes) and taking their net effect into account by means of modified model coefficients (the result is what we call the “intermediate momentum equation” in the following); ii) rescaling of the intermediate equation of motion, so that the surviving modes from step i) become short wavelength modes now. This process of elimination plus rescaling is iterated and self-similarly approaches a “fixed point”. The equation of motion with the fixed point values of the model coefficients is then taken to be the equation governing the dynamics of the large scale eddies (the mean flow field) in the high Reynolds number limit.

3.1 Elimination of the Short-Wavelength Modes

Our model is the system (5) to (12). For wavevectors $k$ with $|k|$ near the reciprocal $\Lambda_d$ of Kolmogorov’s dissipation wavelength, we may assume the motion of the eddies as being critically damped by viscous dissipation, so that the corresponding Fourier modes $u(\omega, k)$ will vanish. Therefore we may introduce an ultraviolet cutoff of the velocity Fourier spectrum, which is denoted by $\Lambda$: $u(\omega, k) = 0$ if $|k| > \Lambda$ and $\Lambda \approx \Lambda_d$. Furthermore, we introduce a smaller wavenumber $\Lambda'$, $\Lambda' = e^{-r}\Lambda$ ($r > 0$), and decompose the velocity field, $u = u^< + u^>$, with $u^< = u^<\chi^<$ and $u^> = u^\chi^>$, where $\chi^<$($k$), $\chi^>$($k$) are the characteristic functions of the wavenumber range $|k| < \Lambda'$ and $\Lambda' < |k| < \Lambda$, respectively. We want to calculate the effect of the “short-wavelength modes” $u^>$ on the “long-wavelength modes” $u^<$. To this end, we project the equation of motion (5) onto the long-wavelength range ($|k| < \Lambda'$) and the short-wavelength range ($\Lambda' < |k| < \Lambda$), respectively; e.g. for the latter case,

$$Lu^> = f^>_t + \lambda \sum_{j,m} M_{ijm}[N_{jm}(u^<, u^<) + 2N_{jm}(u^<, u^>) + N_{jm}(u^>, u^>)]$$

(13)

with $M_{ijm}^>(k) = M_{ijm}(k)\chi^>(k)$, and similarly for $u^<$. The equations for $u^>$ and $u^<$ are coupled via the nonlinear operator $N_{jm}$. In the next step, we try
a perturbation series expansion for \( u^> \) with the expansion parameter \( \lambda \),

\[
u^> = u^{<0>/>} + \lambda u^{<1>/>} + \lambda^2 u^{<2>/>} + O(\lambda^3)
\]

(14)

and with coefficients calculated from (13), e.g.

\[
u^<0/>_i(\omega, k) = G(\omega, k)f_i^>(\omega, k)\text{ etc.}
\]

(15)

where

\[
G(\omega, k) = L^{-1}(\omega, k) = 1/(i\omega + [\nu + i\omega C_0]|k|^2).
\]

(16)

(Writing (13) in dimensionless form, \( \lambda \) would be replaced by a Reynolds number \( \overline{\lambda} \) being characteristic for the small scale eddies (i.e. the modes \( u^>(\omega, k) \)). During the RNG procedure, this Reynolds number will be renormalized and then it proves to remain finite, even if the “bare” Reynolds number goes to infinity, i.e. when fully developed turbulence is reached. Moreover, if the parameter \( \varepsilon := 4 + y - d \) (\( y = 1 \) in three space dimensions) with \( y \) from (12) is small, then \( \overline{\lambda} \) turns out to be small too. So in this case the expansion (14) can be justified a posteriori.) Substituting (14) into the projected momentum equation for \( u^< \) (i.e. the analog of (13)), we are able to eliminate the small-wavelength modes \( u^> \) from this equation. They are expressed now in terms of \( u^< \) and \( f^> \).

3.2 Averaging the Momentum Equation

Using the statistics of the random stirring force \( f \), the momentum equation for the long-wavelength modes \( u^< \) resulting from the preceding step will be small-scale force averaged; i.e. the average will be taken over those realizations, which have the same large scale fields, but different small scale fields. So, this averaging process is assumed to leave long-wavelength modes invariant, \( \langle f^< \rangle = f^< \) and \( \langle u^< \rangle = u^< \).

On the other hand, terms like \( N_{jm}(u^<, u^{<0>/>}) \) in the momentum equation for \( u^< \) will be averaged out,

\[
\langle N_{jm}(u^<, u^{<0>/>}) \rangle = N_{jm}(u^<, \langle u^{<0>/>} \rangle) = N_{jm}(u^<, G\langle f^> \rangle) = 0,
\]

(17)

because of \( \langle f^> \rangle = 0 \). Using arguments like this, the averaged equation of motion for \( u^< \) becomes

\[
Lu^<_i = f^<_i + \lambda \sum_{j,m} M^<_j N_{jm}(u^<, u^<) + T_1 + T_2 + T_3 + \\
+ \text{terms at least trilinear in } u^< + O(\lambda^4)
\]

(18)

with

\[
T_1 = 4\lambda^2 \sum_{j,m,i,n} M^<_j N_{jm}(u^{<0>/>}, GM^>_m N_{in}(u^{<0>/>}, u^<))
\]

(19)
\[ T_2 = 8\lambda^3 \sum_{j,m,i,n,r,s} M_{ijm}^< \langle N_{jm}(u^{0<}), GM_{r,s}^> N_{in}(u^<, GM_{r,s}^> N_{rs}(u^{0<}, u^{0<})), N_{jm}(GM_{r,s}^> N_{in}(u^{0<}), GM_{r,s}^> N_{rs}(u^{0<}, u^<)) \rangle \]  

\[ T_3 = 4\lambda^3 \sum_{j,m,i,n,r,s} M_{ijm}^< \langle N_{jm}(GM_{r,s}^> N_{in}(u^{0<}), u^<), GM_{r,s}^> N_{rs}(u^{0<}, u^<)) \rangle. \]  

### 3.3 The Intermediate Momentum Equation

Equation (18) is the starting point for constructing the small-scale induced corrections of the model coefficients in the momentum equation for \( u^< \). The \( O(\lambda^2) \)-term, \( T_1 \), is linear in \( u^< \) and will be shown to correct the complex viscosity \( \nu + i\omega C_0 \). The \( O(\lambda^3) \)-terms, \( T_2 \) and \( T_3 \), are bilinear in \( u^< \) and provide us with contributions to the bilinear operators \( A_{jm}, ..., D_{jm} \) from 2.2, so they result in corrections to the model coefficients \( C_1, ..., C_4 \). Neglecting in (18) the higher order nonlinear terms in \( u^< \), is in the spirit of “approximate renormalization”, which has been shown for some model problems to lead to surprisingly good results [15], [3]. In some more detail, the term \( T_1 \) has the following structure of nested convolutions:

\[ T_1 = 4\lambda^2 \sum_{j,m,i,n} M_{ijm}^< \langle u_j^{0<} \rangle \times GM_{min}^> (u_i^{0<} \times u_i^<) \rangle + ... + \]

\[ + 4\lambda^2 \sum_{j,m,i,n,k,r} C_4^2 M_{ijm}^< \langle k_k u_j^{0<} \rangle \times k_k GM_{min}^> (k_k u_i^{0<} \times k_k u_i^<)) \rangle. \]

In order to show, along which lines the viscosity correction is established, we want to examine as an example only the first term in (22), calling it \( T_{11} \).

\[ T_{11} \sim \lambda^2 W_0 \sum_{j,m} M_{ijm}^< \sum_{i,n} \langle \int \int |G(\omega', k')|^2 G(\omega - \omega', k - k')d\omega' \]

\[ \times M_{min}^> (k - k') |k'|^{-y} D_{jn}(k')dk' \rangle u_i^<(\omega, k) \]

Note that in (23) the correlation function of \( f \) according to (11) has been used. For correctly evaluating the above integral, we have to observe several technical points. At first, that because of the existence of \( M_{min}^> \) in the integrand, the domain \( D \) of integration is the intersection of the wavenumber shells \( \Lambda' \leq |k'| \leq \Lambda \) and \( \Lambda' \leq |k' - k| \leq \Lambda \). \( D \) possesses a plane of reflection through \( k/2 \). Making use of this symmetry, the transformation \( k' = k/2 + q \) will be used. Now the inner frequency integral may be evaluated using e.g. the residue theorem, then the whole integrand of the \( k' \) integral is expanded in
powers of $k/|q|$, under the assumption that $|k|/|q| << 1$ ("distant interaction approximation"). As the bilinear terms $T_2$ and $T_3$ in (18) contain space derivatives of second order, we have to expand their corresponding integrands $O(|k|^3/|q|^3)$ and, in order to be consistent, we have to do the same for $T_1$. For taking into account the term $i\omega C_0$ in the complex viscosity $\nu + i\omega C_0$, we have the additional requirement, that the integrand above has to be approximated $O(\omega^2)$. Furthermore, we note that after having performed the transformation $k' = k/2 + q$ in the $k'$ integral, the integration with respect to $q$ may be carried out simply over the spherical shell $A' \leq |q| \leq A$, because the $k$-dependent corrections of the integration domain vanish iff evaluated at $\varepsilon := 4 + y - d = y + 1 = 0$. The latter has to be done usually when RNG is applied to turbulence problems using the "$\varepsilon$-expansion" (see [1]). Finally, a peculiarity regarding the evaluation of the terms $T_2$ and $T_3$ should be mentioned, which throws some additional light onto the "$\varepsilon$-expansion". As there arise integrals similar to (23), but with different kernel functions in $T_2 + T_3$, we have to introduce different $\varepsilon_i$-parameters (which have, for $d = 3$, the values $y - 3$, $y - 1$, $y + 1$ and $y + 3$). Again we find that $k$-dependent corrections of the domain of integration vanish iff evaluated at $\varepsilon_i = 0$ (we have shown for a onedimensional model problem that this choice is in some sense the best approximation of the kernel functions). Coming back to $T_{11}$, the overall result of the evaluation procedure for this term is of the form

$$T_{11} = -\nu \bar{\lambda}^2 \left( f(C_0, \Lambda) + i\omega g(\nu, \Lambda) \right) (r + O(r^2)) |k|^2 u_i^{<}(\omega, k)$$

(24)

with real functions $f(C_0, \Lambda)$, $g(\nu, \Lambda)$ and $\bar{\lambda} := \lambda W_0^{1/2}/(\nu^{3/2} A^{\varepsilon/2})$ being a Reynoldsnumber characteristic for the eliminated modes $u^{<}$. From (24) it is obvious, that the term $T_{11}$ contributes to the correction of the complex viscosity by adding the following expressions to $\nu$ and $C_0$, respectively,

$$\Delta \nu = \nu \bar{\lambda}^2 f(C_1, \Lambda) r + O(r^2)$$

$$\Delta C_0 = \nu \bar{\lambda}^2 g(\nu, \Lambda) r + O(r^2).$$

(25)

A similar calculation has to be performed for the other terms in (22) and for $T_2$ and $T_3$ according to the rules of evaluation stated above. The overall result is the intermediate momentum equation for the long-wavelength modes $u^{<}$,

$$L(C'; \nu') u_i^{<} = (i\omega + [\nu' + i\omega C_0]) |k|^2 u_i^{<}$$

$$= f_i^{<} + \lambda \sum_{j,m} M_{ijm}^{<} N_{jm}(C'; u^{<}, u^{<})$$

(26)

which holds for $|k| \leq \Lambda' = e^{-r} \Lambda$. In (26), $\nu' = \nu + \Delta' \nu$ and $C' = (C_0', ..., C_i')$, $C_i' = C_i + \Delta' C_i$ (with $\Delta' \nu$, $\Delta' C_i$ the full corrections similar to (25)) denote the corrected model coefficients, and we have used $L(C'; \nu')$, $N_{jm}(C'; u^{<}, u^{<})$ for indicating the dependence of the linear and nonlinear operators $L$, $N_{jm}$, respectively, on the corrected set of coefficients. Several points regarding the
selfconsistency of our model should be noticed here. At first, the corrections for the coefficient \( C_0 \) stem from the term \( T_1 \) only, whereas the corrections for \( C_1 \) stem from the terms \( T_2 \) and \( T_3 \). Nevertheless, it turns out that the corrections for \( C_0 \) and \( C_1 \) are identical. This is necessary for selfconsistency, because the symmetry condition \( C_0 = C_1 \) is equivalent to the Galileian invariance of the momentum equation. Furthermore, we have shown that the bilinear terms in \( u^< \) (i.e. \( T_2 \) and \( T_3 \)) sum up to 0, when evaluated at order 0 in \( k, k' \). This means that no corrections of the form \( \lambda u_j * u_m \) in (8) are generated, and therefore the bare coupling parameter \( \lambda \) remains unaltered. This again is necessary for selfconsistency, because by the Galileian invariance of the momentum equation, \( \lambda = 1 \) is implied. And at last, all terms generated from \( T_1 \) and \( T_2 + T_3 \), respectively, fit into the structure of (8).

### 3.4 Rescaling the Momentum Equation

Note that (26) is of the same structure as the original momentum equation for \( u \), but it holds in the restricted wavenumber domain \( |k| \leq \Lambda' = e^{-r} \Lambda \). In order to get the equation structurally invariant under the action of the RNG, it has to be rescaled, so that its domain will again be \( |k| \leq \Lambda \). Introducing the scale transformation \( \tilde{k} = e^r k \), \( \tilde{\omega} = \omega e^{a(r)} \), \( \tilde{u}(\tilde{\omega}, \tilde{k}) = u(\omega, k) e^{-b(r)} \) and defining the rescaled parameters \( \tilde{\nu}(r) = \nu' e^{a(r) - 2b(r)} \), \( \tilde{\lambda}(r) = \lambda e^{b(r) - (d+1)r} \), \( \tilde{C}_i(r) = C_i e^{-2r} \) as well as the rescaled force \( \tilde{f}_i^< (\tilde{\omega}, \tilde{k}) = e^{a(r) - b(r)} f_i^< (\omega, k) \), leads to the rescaled momentum equation

\[
(i \tilde{\omega} + \left[ \tilde{\nu}(r) + i \tilde{\omega} \tilde{C}_1(r) \right] |\tilde{k}|^2) \tilde{u}_i^< = \tilde{f}_i^< + \lambda \sum_{j,m} M_{ijm}^< (\tilde{k}) N_{jm} (\tilde{u}_i^<, \tilde{u}_j^<)
\]

with \( k \) varying in the full wavenumber range \( |k| \leq \Lambda \). Choosing the parameters \( a(r), b(r) \) in the scale transformation according to the relation \( 3a(r) - 2b(r) + (d + y)r = 0 \) (\( y \) from (12)) makes the correlation function of the random force \( f \) invariant under the scale transformation, \( \langle f_i^< (\omega, k) f_j^< (\omega, k) \rangle = \langle \tilde{f}_i^< (\tilde{\omega}, \tilde{k}) \tilde{f}_j^< (\tilde{\omega}, \tilde{k}) \rangle \). The further constraint that, once \( \tilde{\lambda} \) has reached a "fixed point" for \( r \to \infty \), also \( \tilde{\nu} \) and \( \tilde{\lambda} \) should become steady, determines the exponents \( a, b \) as to be \( a(r) = (2 - \varepsilon/3) r \), with \( \varepsilon = 4 + y - d \) and \( b(r) = (d + 1) r \) \((d = 3)\). Note that the latter condition implies \( \tilde{\lambda}(r) = \lambda e^{b(r) - (d+1)r} \equiv \lambda \), so that we now may take \( \lambda = 1 \), in order to be compatible with the Galileian invariance of the rescaled momentum equation.

### 3.5 The Recursion Relations

The process of elimination plus rescaling may now be iterated. To this end, step by step elimination of modes from wavevector shells of infinitesimal thickness is considered, \( e^{-r} \Lambda \leq |k| \leq \Lambda \) and formally \( r \) is replaced by \( dr \)
Iteration then leads to differential recursion relations for the model parameters, which are derived from the expressions for the corrected model coefficients similar to (25) combined with rescaling according to 3.4. In that way we arrive at a coupled system of nonlinear ordinary differential equations for determining the renormalized dimensionless model coefficients

\[
\bar{\lambda}(r) := \frac{\bar{\lambda} W_0^{1/2}}{\nu^{3/2} \rho^{\epsilon/2} A c^{2/3}} = \frac{W_0^{1/2}}{\nu^{3/2} \rho^{\epsilon/2} A c^{2/3}}, \quad \bar{C}_i(r) := \bar{C}_i A^2
\]  

as functions of the parameter \( r \), which controls the elimination process. This system (recursion relation of the RNG), which depends only on one external parameter, \( \epsilon = 4 + y - d \) (\( d = 3 \)), is of the following form,

\[
\frac{d\bar{\lambda}}{dr} = \frac{1}{2} \bar{\lambda} \left( \epsilon - \frac{1}{20\pi^2} \bar{\lambda}^2 \frac{1}{1 + \bar{C}_1} P_0(\bar{C}_1, \ldots, \bar{C}_4) \right)
\]  

\[
\frac{d\bar{C}_1}{dr} = \frac{1}{120\pi^2} \bar{\lambda}^2 P_1(\bar{C}_1, \ldots, \bar{C}_4) - 2\bar{C}_1
\]  

\[
\frac{d\bar{C}_i}{dr} = \frac{1}{420\pi^2} \bar{\lambda}^2 \frac{1}{1 + \bar{C}_1} P_i(\bar{C}_1, \ldots, \bar{C}_4) - 2\bar{C}_i, \quad i = 2, 3, 4.
\]

In (29)/(31), \( P_0 \) and \( P_1 \) are second order polynomials in \( \bar{C}_1, \ldots, \bar{C}_4 \) with integer coefficients. \( P_i \) with \( i \geq 2 \) are third order polynomials in \( \bar{C}_1, \ldots, \bar{C}_4 \), with integer coefficients. A discussion of the bifurcation behaviour of the solution shall be presented elsewhere. (It may be of interest for examining the effect of intermittency, even if there may be some doubts as to whether the self-similar RNG solution is still valid in that case, see the discussion in [8]). Here we are only interested in solutions of (29)/(31) for the choice \( \epsilon = 4 \) (i.e. \( y = 3 \)), which corresponds to the Kolmogorov energy spectrum. The system (29)/(31) may then be used in a twofold way. For finite Reynolds numbers (we call this the “low-Reynolds-number” case), which corresponds to finite values of the parameter \( r \), the model coefficients \( \bar{\lambda}(r), \bar{C}_i(r) \) have to be evaluated by solving the differential equation. Because of \( \Lambda' = e^{-r} \Lambda \), this means of course, that the model parameters are cutoff-dependent, \( (\bar{\lambda}, \bar{C}_i) = (\bar{\lambda}(\Lambda'), \bar{C}_i(\Lambda')) \). On the other hand, the cutoff \( \Lambda' \) can be related to the local turbulence Reynolds number, \( R_t = K^2/(\nu_0 \epsilon) \) (with \( K \) the turbulent kinetic energy, \( \epsilon \) its dissipation rate and \( \nu_0 \) the “bare” (=molecular) viscosity). So the model coefficients may be determined from (29)/(31) (or a suitable approximation of this system) as being \( R_t \)-dependent. Due to the fact that \( (\bar{\lambda}(r), \bar{C}_i(r)) \) depends on the initial value \( (\bar{\lambda}(0), \bar{C}_i(0)) \), no universality can be expected in the “low-Reynolds-number case”. The opposite is true in the “high Reynolds number limit”, which corresponds to \( r \to \infty \). In that case, we are looking for the equilibrium points of (29)/(31). For \( \epsilon = 4 \), we have found

\[
(\bar{\lambda}, \bar{C}_1, \ldots, \bar{C}_4) = (12.7365, -0.3054, -0.0243, -0.3734, -0.2861),
\]

which is attractive, and additionally two unstable equilibrium points.
4 The High-Reynoldsnumber Limit

4.1 The Model Coefficients

From (32) and (28), we have for \( r \to \infty \),
\[
W_0^{1/2}/(\gamma^{3/2} A'^{3/2}) = W_0^{1/2}/(\nu^{3/2} A'^{3/2}) \to 12.7365,
\]
by which the asymptotic relation
\[
\nu(A') \sim \frac{1}{(12.7365)^{2/3}} W_0^{1/3} A'^{-\varepsilon/3}, \quad r >> 1 \quad \text{(i.e.} \quad 0 < A' << 1), \tag{33}
\]
is implied as being the cutoff-dependent expression for the effective viscosity in the high-Reynoldsnumber limit. Following the arguments in [1], we get from an EDQNM-type equation the forcing “strength” \( W_0 \) as being proportional to the turbulence dissipation rate \( \varepsilon \). Then we calculate \( C_K = 1.50 \) as the numerical value of the Kolmogorov constant, which is well within the generally accepted range from 1.2 to 2.2. From the Kolmogorov energy spectrum, the cutoff-dependent expression
\[
K(A') \approx \frac{3}{2} C_K \varepsilon^{2/3} A'^{1-2/3} \tag{34}
\]
for the turbulent kinetic energy is deduced. With the aid of (34), the “fixed point” values \( (\overline{C}_1, ..., \overline{C}_4) \) from (32) are transformed into
\[
C_i = \overline{C}_i (2/3)^3 \frac{K^3}{\mathcal{C}_K^2 \varepsilon^2}, \quad i = 1, ..., 4. \tag{35}
\]
Numerically, we get
\[

\nu = 0.08460 \frac{K^2}{\varepsilon}; \quad C_1 = -0.02696 \frac{K^3}{\varepsilon^2}; \quad C_2 = -0.00215 \frac{K^3}{\varepsilon^2}; \tag{36}
\]
\[

C_3 = -0.03296 \frac{K^3}{\varepsilon^2}; \quad C_4 = -0.02525 \frac{K^3}{\varepsilon^2},
\]
which constitutes the set of our model coefficients in the “high-Reynolds number limit”.

4.2 The Viscoelastic Model Fluid

From the detailed structure of the system (29)/(31) it follows, that the coefficients \( (\overline{C}_1, ..., \overline{C}_4) \) are subject to the constraint
\[
\overline{C}_1(r) + \overline{C}_2(r) = \frac{1}{2} (\overline{C}_3(r) + \overline{C}_4(r)), \tag{37}
\]
provided that this symmetry is fulfilled by the initial values \( \overline{C}_i(0) \); the latter is evidently the case, if the fluid under consideration is Newtonian \( (\overline{C}_i(0) = \overline{C}_i(0) \)


0) or of second order type. (Of course, (37) may then be used to eliminate one equation from (29).) There is a further simplification in the high-Reynolds number limit, i.e. in fully developed turbulent flow. In our model, the high-Reynolds number limit means \( r \to \infty \), i.e. all fluctuating modes \( u' \) of the instantaneous flow field, \( u = U + u' \), have been averaged out; what is left, is the mean flow field \( U \). Now we want to examine the behaviour of the second and third terms in \( S_{ij} \) from (3) in the limit \( r \to \infty \). Because of \( C_0 = C_1 \) they are responsible for the convective transport of the strain rate tensor \( A_1 = \nabla u + (\nabla u)^T \) of the instantaneous flow field,

\[
\left( \frac{\partial^2 u_i}{\partial t \partial x_j} + \frac{\partial^2 u_j}{\partial t \partial x_i} \right) + \left( \sum_k u_k \frac{\partial^2 u_i}{\partial x_k \partial x_j} + \sum_k u_k \frac{\partial^2 u_j}{\partial x_k \partial x_i} \right) = \frac{\partial A_1}{\partial t} + (u \cdot \nabla) A_1
\]

(38)

We have the decomposition \( A_1 = A_{1M} + A_{1F} \) with \( A_{1M} = \nabla U + (\nabla U)^T \) and \( A_{1F} = \nabla u' + (\nabla u')^T \) representing the strain rate of the mean flow and the fluctuating part, respectively. Therefore in (38),

\[
\frac{\partial A_1}{\partial t} + (u \cdot \nabla) A_1 = \frac{\partial A_{1M}}{\partial t} + (u \cdot \nabla) A_{1M} + \frac{dA_{1F}}{dt} = \frac{dA_{1F}}{dt},
\]

(39)

where \( d/dt \) denotes the material derivative and \( (u \cdot \nabla) A_{1M} = 0 \) follows from our homogeneity assumption (see 2.2), which means that the mean strain rate tensor must be spatially uniform; the mean strain rate also being steady implies \( \frac{\partial A_{1M}}{\partial t} = 0 \). Now taking the limit \( r \to \infty \) means that the fluctuating part has been fully removed, so the right-hand side in (39) vanishes. Therefore, in the high Reynolds number limit the second and third term in the extrastress \( S_{ij} \) from (3) may be dropped, and then (37) implies that our model fulfills the requirement of material objectivity from continuum mechanics (see [13], [11]). Furthermore this means that locally homogeneous fully developed turbulence may be described by the constitutive equation of a viscoelastic second order fluid,

\[
S = \nu A_1 + \alpha_1 A_2 + \alpha_2 A_1^2
\]

(40)

with \( A_1, A_2 \) being the first and second order Rivlin–Ericksen tensors

\[
A_1 = (\nabla u + (\nabla u)^T)
\]

(41)

\[
A_2 = \frac{\partial A_1}{\partial t} + (u \cdot \nabla) A_1 + A_1 \cdot \nabla u + (A_1 \cdot \nabla u)^T.
\]

Comparing (40) with (3) results in

\[
\alpha_1 = -0.00385 \frac{K^3}{\varepsilon^2}
\]

(42)

\[
\alpha_2 = -0.02525 \frac{K^3}{\varepsilon^2}
\]
with the constraint \( \alpha_1 + \alpha_2 = C_1 + C_2 \). The latter is implied by (37), so consistency is checked. The parameters from (42) are the Rivlin–Ericksen coefficients of the generic second order model fluid describing locally homogeneous fully developed turbulent flow. So the proposal of Rivlin [16], who conjectured nearly fifty years ago, that a turbulent Newtonian fluid could be regarded as a non-Newtonian fluid, seems to have received some support. In cooperation with the Rhodia Acetow AG, Freiburg, we try to understand the rheological properties of our model fluid by comparing them with those of "real" viscoelastic fluids like polymer solutions. Especially, the parameters \( \alpha_1, \alpha_2 \) are related to the first and second normal stress coefficient of a viscoelastic fluid \( (n_1, n_2) \) by \( n_2 = 2\alpha_1 + \alpha_2 \). So we will observe normal stress effects in turbulent shear flows.

### 5 Numerical Test Case

We expect the existence of normal stress differences in shear flows to be an essential feature in turbulent separated flows, like the flow behind a car. The shape and extent of the wake region has a major effect on the drag of a vehicle, so it is important to predict it reliably. The ultimate goal of this project is the simulation of the flow behind a car and the comparison of the computed results with data from wind tunnel experiments provided by DaimlerChrysler. As a first step on this way, we have used the backward facing step problem with experimental data of Eaton and Johnston [4] as a test case. The computations have been performed by means of the finite element code FIDAP. We took the high-Reynolds number data set of [4], i.e. with a Reynolds number based on the step height \( H \) of \( 3.95 \times 10^4 \). The outlet channel height is \( 2.5H \), so downstream the step there is a sudden expansion with a ratio of 1.66. By means of a channel flow calculation, inlet profiles for \( U, K \) and \( \epsilon \) have been computed such as to meet the inlet data of [4]. The profiles have been prescribed \( 6H \) upstream the step. As Thangam [19] mentions that significant errors may be induced by imposing a fully developed outflow condition not far enough downstream, we used the natural stress free outflow condition \( 30H \) downstream the step. This distance was recommended by Thangam [19]. It is known that the logarithmic law of the wall is not well suited for separated flows and within reattachment zones. But following Yakhot et al. [21], it is not necessary to use at the upper and lower wall boundary conditions different from the log law, as test calculations have shown that the effect of different wall treatment is small in the backward facing step problem, provided that the flow field has been resolved fine enough. So we used our high-Reynolds number model from above together with the special wall elements from FIDAP (which use Reichardt's law for interpolating between the viscous sublayer via the buffer layer to the log law layer and \( u_\tau = C^{1/4} K^{1/2} \) as the friction velocity, in order to overcome the above mentioned problem at the reattachment point). Regarding the wall distance, at very few wall elements
behind the step only, $y^+$ values of 10 have been accepted, everywhere else $y^+$ exceeded 30. In order to study a possible grid-dependence, we used two different meshes (with 7000 and 27000 elements, respectively, both with 9-node quadrilaterals). We used the mixed pressure solver [5] and a coupled calculation on the smaller mesh (at the beginning successive substitutions, then Newton-Raphson), with upwinding for $K$ and $\epsilon$. On the finer mesh, we used the segregated solver [5]. Some typical results are shown in the figures. In Fig. 1 the streamlines are depicted, Fig. 2 shows isolines of the x-component of the velocity. On the left side of the figures, the isotropic standard $K - \epsilon$ model has been used for comparison. The right side shows our high-Reynoldsnumber viscoelastic model. The different shape and extent of the separation zones is obvious. The reattachment length from experiment is $7.95H \pm 1H$ (this uncertainty is because of a low frequency motion of the impingement point [4]). This compares with a value of $6.65H \pm 0.15H$ from our computation with the isotropic standard $K - \epsilon$ model and a value of $8.45\pm0.25H$ from the calculation with our high-Reynoldsnumber viscoelastic model. Whereas the reattachment length is clearly underpredicted by the standard $K - \epsilon$ model, the value given by our model fits into the scatter range of the experimental data. The next test case we want to address, is the flow around a spheroid, which has been used recently as an ECARP (European Computational Aerodynamics Research Project) test case [9]. One of the additional challenges contained in this problem is, that the separation line is not "prescribed" a priori by the geometry data, so the low-Reynoldsnumber treatment will become important in this case.
Acknowledgements. We would like to thank DaimlerChrysler AG, Rhodia-Acetow AG and the Steinbeis transfer center HTCO for their fruitful cooperation, Dr. R. Rubinstein from ICASE, NASA Langley, for valuable hints and discussions and Prof. Dr. K. Kirchgaessner and Prof. Dr. H. Kielhoefer for their permanent interest in our work and for many helpful and encouraging discussions.

References

Modelling and Simulation of Capacitor Impulse Welding

D. Hömberg, W. Dreyer and F. Duderstadt

Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, 10117 Berlin, Germany, E-mail: hoemberg@wias-berlin.de,
URL: http://www.wias-berlin.de

Abstract. We present a numerical simulation of capacitor resistance welding. The model accounts for electrical, thermal and mechanical effects, which are non-linearly coupled by the balance laws, constitutive equations and boundary conditions. Assuming only plane strains, we develop a two-dimensional FE approximation with simplicial elements, which is especially suited for the simulation of workpieces with longitudinal projections or thin-walled rings with annular projection. In the melted zone the material is described as an isotropic, electrically conducting thermo-viscoelastic Maxwellian body. In the solid it is assumed to behave non-linearly thermoelastic.

1 Introduction

Compared to other resistance welding technologies, capacitor impulse welding is characterized by an extremely short weld time and a very small heat-affected zone (HAZ). Therefore, this process is well-suited for the welding of different materials or coated materials. Owing to the smallness of the HAZ, the joint is cooled down rapidly by the surrounding material, leading to fine grained structure of the melted zone.

Although there are numerous national and international standards for conventional resistance welding, there is nothing comparable for impulse welding. Hence, it is the goal of this research project to provide our industrial partner with numerical simulations to optimize this technology and to improve and widen its range of applicability.

In the next section, we describe the weld process. Further details can be found in [2] and [3]. In Sect. 3 we derive a mathematical model for the process, which can be viewed as a mechanical extension of the classical thermistor problem (see e.g. [1], [10], and the references therein).

Since the welding process is governed by the discharge of a capacitor, it is important to account also for the electrical properties of the weld machine. This is done in Sect. 3 by a discrete model, coupled to the potential equation by a nonlinear boundary condition. In Sect. 4 we discuss some questions about mathematical results for the model. Numerical results and some details about the numerical algorithm are presented in Sect. 5. The last section is devoted to some concluding remarks.
2 Description of the Weld Process

Figure 1 depicts a sketch of the impulse welding process. First, the capacitor (3) is charged by the transformer (1). For fixed capacity the charging voltage defines the stored energy that can be supplied to the welding process. Then the welding process is started by applying a force $F$ onto the electrode (5). If the initial deformation is too small, the local increase in temperature will be too high, possibly leading to spurring of liquid metal, thereby decreasing the joint quality. On the other hand, if the initial deformation is too big the contact resistivity is diminished and higher welding energy is needed for producing the joint, leading to a larger HAZ [8].

When the primary circuit between capacitor and impulse transformer (4) is closed, current flows through the contact area between the weld pieces (6), (7) in the secondary circuit. In this damped resonant circuit the capacitor energy is completely transformed into heat during a short period of time. Owing to the geometric singularity, the biggest increase in temperature takes place in the projection tip.

The impulse transformer is adjusted to achieve a discharging current below the aperiodic limit. In this case the efficiency of the welding process is higher than in the case of a discharging current that develops as a damped oscillation, because the amount of energy that can dissipate by heat conduction is smaller, leading also to a smaller size of the heat affected zone.

When the melting temperature is reached in the contact area, the joint begins to grow. The increasing deformation of the projection corresponds to a growing contact area and a decreasing resistance at the joint. This effect is amplified by an increasing specific heat capacity. Finally, heat conduction and a decreasing discharging current prevent a further rise of temperature and the process stops with a fast cooling of the welded pieces.

![Fig. 1. Sketch of a capacitor impulse weld machine](image-url)
3 The Model

To derive the physical model we make the following assumptions and simplifications: The electrical properties of the welding machine are accounted for by a discrete model. Further resistances, capacities and inductances caused by the electrically conducting parts of the machine are neglected.

The workpieces to be joined are assumed to be isotropic, deformable, electrically and thermally conducting materials. More precisely, we describe the melted zone as a Maxwellian thermo-viscoelastic body. The solid part, including the heat affected zone is assumed to behave thermoelastically. All material parameters depend on temperature.

To obtain basic features of the process, we make further simplifications. Assuming small deformations we formulate the balance laws in the undeformed region. Therefore one can only expect reasonable results from the simulations as long as the overall deformations, especially in the projection tip, remain small. Finally, we also neglect the initial plastic deformation of the projection tip.

To obtain the displacement \( u(x, t) \) or the velocity \( v(x, t) \), respectively, the electric potential \( \phi(x, t) \) and the temperature \( T(x, t) \), we evaluate the quasistatic balance laws of momentum and electrical charge as well as the balance law of internal energy:

\[
\begin{align*}
\text{div} \, (\sigma) &= 0, \quad \text{(1a)} \\
\text{div} \, (J) &= 0, \quad \text{(1b)} \\
\rho \dot{\varepsilon} + \text{div} \, q &= \sigma : \dot{\varepsilon} + E \cdot J. \quad \text{(1c)}
\end{align*}
\]

Here \( \rho \) is the mass density, \( \sigma \) the stress tensor, \( E = - \text{grad} \, \phi \) the electric field, \( J \) the electric current density, \( q \) the heat flux, \( \varepsilon \) the specific internal energy and \( \varepsilon = \frac{1}{2} (Du + DT \, u) \) the symmetric part of the strain tensor. The scalar products in \( \mathbb{R}^n \) and \( \mathbb{R}^{n \times n} \) are denoted by ‘·’ and ‘:’, respectively.

The constitutive equations for \( J \), \( q \) and \( \sigma \) are

\[
\begin{align*}
J &= \dot{J}(T, \text{grad} \, \phi) = -\gamma(T) \, \text{grad} \, \phi, \quad \text{(2a)} \\
q &= \dot{q}(T, \text{grad} \, T) = -\lambda(T) \, \text{grad} \, T, \quad \text{(2b)} \\
\sigma &= \dot{\sigma}(T, \varepsilon, \varepsilon^c) = K(T) \left( \varepsilon - \varepsilon^c - \alpha(T)(T - T_R)I \right). \quad \text{(2c)}
\end{align*}
\]

Equations (2a) and (2b) are the laws of Ohm and Fourier, \( \gamma \) is the specific electric conductivity, and \( \lambda \) is the heat conductivity, (2c) is Hooke’s law for small displacements of a thermoelastic Maxwellian body and \( K = \{ K_{ijkl} \} \) is the isotropic stiffness tensor with temperature dependent elastic coefficients. \( T_R \) is a reference temperature, and \( \alpha \) denotes the thermal expansion coefficient. For the creep strain \( \varepsilon^c \), which is the internal variable of the Maxwellian body, we assume an evolution equation similar to the Norton creep law,

\[
\varepsilon^c = B(T, \sigma) = \eta_1(T)|S|^{|\eta_2-1}S, \quad \text{(3)}
\]

where \( S = \sigma - \frac{1}{n} \left( \text{tr} \, \sigma \right) I \) is the trace-free part of the stress tensor.
Combining (2c) and (3), we obtain
\[ \dot{\sigma} + K(T)B(\sigma, T) - \dot{T}K_{,T}K^{-1}\sigma = K(T)\left(\dot{\varepsilon} - (\alpha(T) + \alpha,T(T - T_R))\dot{T}\right). \quad (4) \]

To derive a constitutive equation for the internal energy we exploit the entropy inequality with the entropy flux \( h = \frac{1}{T} q \),

\[ \rho \dot{s} + \text{div} \left( \frac{1}{T} q \right) \geq 0. \quad (5) \]

The evaluation of this inequality can be considerably simplified by introducing Lagrange multipliers \( \Lambda^v, \Lambda^T \) and \( \Lambda^c \) and considering

\[ \rho \dot{s} + \text{div} \left( \frac{1}{T} q \right) - \Lambda^v \cdot (\rho \dot{v} - \text{div} \sigma) - \Lambda^T \left( \rho \dot{e} + \text{div} q - \sigma : \dot{\varepsilon} \right) - \Lambda^c : \left( \dot{\varepsilon}^c - B(T, \sigma) \right) \geq 0. \quad (6) \]

Here we replaced the quasistatic momentum balance (1a) with the complete one. While (5) is valid only for solutions to the field equations (1a)–(2c), (6) must hold for all fields, according to I Shi Liu's theorem [9].

Using the Helmholtz free energy \( \Psi \), we have

\[ e = \Psi + s T. \]

Now we assume that \( s \) and \( \Psi \) only depend on \( \varepsilon, \varepsilon^c, T, \) \( \text{grad} T \) and that they are continuously differentiable with respect to these quantities, i.e.

\[ \dot{s} = \frac{\partial s}{\partial \varepsilon} : \dot{\varepsilon} + \frac{\partial s}{\partial \varepsilon^c} : \dot{\varepsilon}^c + \frac{\partial s}{\partial T} \dot{T} + \frac{\partial s}{\partial \text{grad} T} \cdot \left( \text{grad} T \right), \quad (7) \]

\[ \dot{\Psi} = \frac{\partial \Psi}{\partial \varepsilon} : \dot{\varepsilon} + \frac{\partial \Psi}{\partial \varepsilon^c} : \dot{\varepsilon}^c + \frac{\partial \Psi}{\partial T} \dot{T} + \frac{\partial \Psi}{\partial \text{grad} T} \cdot \left( \text{grad} T \right). \quad (8) \]

Since (6) is linear in \( \dot{v} \), and also in \( \text{div} q \), we obtain \( \Lambda^v = 0 \) and \( \Lambda^T = \frac{1}{T} \), leading to the inequality

\[ -q \cdot \text{grad} T - \frac{1}{T} \left( \rho(\dot{\Psi} + s \dot{T}) - \sigma : \dot{\varepsilon} \right) - \Lambda^c : \left( \dot{\varepsilon}^c - B(T, \sigma) \right) \geq 0. \quad (9) \]

Next, in view of (8) we see that (9) is linear with respect to \( \dot{T} \) and obtain \( s = -\frac{\partial \Psi}{\partial T} \). In the same way we conclude that \( \psi \) is independent of \( \text{grad} T \), \( \sigma = \rho \frac{\partial \psi}{\partial \varepsilon} \) as well as \( \Lambda^c = -\rho \frac{\partial \psi}{\partial \varepsilon^c} \). Invoking (2b) and (3) we finally arrive at

\[ -\frac{\rho}{T} \eta_1(T) S |S|^{\eta_2 - 1} \frac{\partial \Psi}{\partial \varepsilon^c} : S \geq 0, \quad (10) \]
which can only be satisfied if \( \frac{\partial \psi}{\partial \epsilon} = \kappa S \), with a scalar \( \kappa < 0 \). Assuming \( \psi \) is twice continuously differentiable and using Hooke’s law (2c), we obtain \( \kappa = -1/\rho \). Finally, we gather up all these relationships to conclude

\[
\rho \dot{\epsilon} = \rho c_v \dot{T} + (\sigma - T \sigma, T) : \left( \dot{\epsilon} - B(T, \sigma) \right) .
\]

Here, \( c_v = e_{,T} \) is the specific heat capacity at constant volume and we have

\[
\sigma_{,T} = K_{,T} K^{-1} \sigma - \left( \alpha(T) + \alpha_{,T}(T - T_R) \right) K I.
\]

### 4 Initial and Boundary Conditions for a Symmetric Workpiece with Longitudinal Projection

In this section we consider initial and boundary conditions for one half of the cross-section of a weld piece with translational symmetry, cf. Fig. 2). Initially, the body is at rest, it is stress free, of homogenous room temperature \( T_0 \) and has a vanishing electrical potential \( \phi \). Only plane strains can occur and the \( z \)-components of the gradients of temperature and electric potential, respectively, are zero. The boundaries \( \Gamma_4 \) are completely stress free, whereas only shear stresses vanish at the remaining boundaries. \( \Gamma_3 \) is a symmetry boundary on which we assume \( v_1 = 0 \). On \( \Gamma_1 \) we assume \( v_2 = 0 \). The normal stress on \( \Gamma_2 \) is governed by the temporal evolution of the electrode force \( F \), more precisely, we have

\[
\int_{\Gamma_2} \sigma_{22} ds = F(t),
\]

where \( F \) is explicitly prescribed. Owing to the short weld time of few milliseconds, we neglect heat conduction across the boundary, i.e. we assume \( q \cdot \nu = 0 \), where \( \nu \) is the outer normal vector.

**Fig. 2.** Definition of boundaries (left) and circuit diagram of the resonant circuit during welding (right)
For the electric potential we assume \( \phi = 0 \) on \( \Gamma_1 \) and \( J \cdot \nu = 0 \) on \( \Gamma_3 \) and \( \Gamma_4 \). On \( \Gamma_2 \) we assume
\[
J \cdot \nu = -\frac{1}{|\Gamma_2|} \dot{Q}_2(t).
\] (12)

Here, \( |\Gamma_2| \) indicates the length of \( \Gamma_2 \) and \( \dot{Q}_2 \) is the current intensity in the secondary circuit as depicted in Fig. 2. \( Q = (Q_1, Q_2)^T \) is obtained as the solution to the following system of ordinary differential equations
\[
\begin{pmatrix}
L_1 & L_{12} \\
L_{12} & L_2
\end{pmatrix}
\begin{pmatrix}
\dot{Q}_1 \\
\dot{Q}_2
\end{pmatrix}
+ \begin{pmatrix}
C_1^{-1} & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
Q_1 \\
Q_2
\end{pmatrix}
= - \begin{pmatrix}
0 & 0 \\
0 & R_2(t)
\end{pmatrix}
\begin{pmatrix}
\dot{Q}_1 \\
\dot{Q}_2
\end{pmatrix}
= \begin{pmatrix}
0 \\
-U_F
\end{pmatrix},
\]
or equivalently
\[
\mathcal{L} \ddot{Q} + C^{-1} Q = -\mathcal{R}(t) \dot{Q},
\] (13)

with initial conditions \( \dot{Q}(0) = (0, 0)^T \), \( Q(0) = (C_1 U_L, 0)^T \). \( U_L \) is the charging voltage according to Fig. 1 and \( R_2 \) is the discrete equivalent resistance of the weld pieces \( \Omega \). The mutual inductivity \( L_{12} \) is given by \( L_{12} = k \sqrt{L_1 L_2} \) with a coupling factor \( k \in (0, 1) \). The voltage drop \( U_F \) on \( \Omega \) can be derived from the identity
\[
U_F \dot{Q}_2 = J \cdot E = \int_{\Omega} \gamma(T) |\text{grad } \phi|^2 dx.
\]

Testing (1b) with \( \phi \) using (2a) and (12), we obtain
\[
\int_{\Omega} \gamma(T) |\text{grad } \phi|^2 dx = \dot{Q}_2 \frac{1}{|\Gamma_2|} \int_{\Gamma_2} \phi ds.
\]

In other words, we have
\[
U_F = \ddot{\phi} := \frac{1}{|\Gamma_2|} \int_{\Gamma_2} \phi ds.
\] (14)

5 Some Remarks on the Mathematical Analysis of the Model

If we assume the stiffness tensor and the thermal expansion coefficient to be independent of temperature, we obtain the following system of equations:

\[
\text{div } \sigma = 0, \quad (15a)
\]
\[
\dot{\sigma} + KB(\sigma, T) = K \left( \dot{\epsilon} - \alpha \dot{T} I \right), \quad (15b)
\]
\[
\rho c_V \dot{T} - \text{div} \left( \lambda(T) \text{grad } T \right) = -\alpha (n \mu_1 + 2 \mu_2) T \text{div } v
+ n_1(T) |S|^{n_2+1} + \gamma(T) |\text{grad } \phi|^2, \quad (15c)
\]
\[
- \text{div} \left( \gamma(T) \text{grad } \phi \right) = 0, \quad (15d)
\]
together with the initial and boundary conditions

\[ \sigma_{ij} n_j = g_i, \quad \text{on } \Sigma_1, \]  
\[ v = 0, \quad \text{on } \Sigma_2, \]  
\[ T(\cdot, 0) = T_0, \quad \text{in } \Omega, \quad \frac{\partial T}{\partial n} = 0, \quad \text{on } \partial \Omega, \]  
\[ \phi = 0, \quad \text{on } \Sigma_2, \]  
\[ -\gamma(T) \frac{\partial \phi}{\partial n} = -\frac{1}{|\Sigma_1|} \dot{Q}_2(t), \quad \text{on } \Sigma_1, \]  

where \( Q = (Q_1, Q_2) \) is the solution to

\[ \mathcal{L} \dot{Q} + C^{-1} Q = \left( 0, -\frac{1}{\Sigma_1} \int_{\Sigma_1} \phi \, dx \right)^T, \]  
\[ Q(0) = (C_1 U_L, 0)^T, \quad \dot{Q}(0) = (0, 0)^T. \]

Here, \( \partial \Omega = \Sigma_1 \cup \Sigma_2, \mu_{1,2} \) are the Lamé constants and \( n = \dim \Omega \).

Neglecting the terms in the stress deviator \( S \) and the velocity \( v \), the system (15c,d) is the classical thermistor problem, coupled however with the non-standard boundary condition (16e), (17a,b). But thanks to a regularity result by Gröger [5], at least in two space-dimensions, the results of [10] can be applied to obtain existence of a weak solution to this problem, whereas uniqueness is still an open problem, cf. e.g. [1].

In [6] the existence of a weak solution to a thermoelastic problem, coupled with (15d) through the Joule heating term \( \gamma(T) | \text{grad } \phi |^2 \) has been proved. In [7] we have investigated an electro thermo-conductive problem with a linear creep law and a linear heat equation.

Note that even in the linear version of the Norton law (3), i.e. \( \eta_2 = 1 \), one obtains a quadratic term in \( \sigma \) on the right-hand side of the energy balance.

### 6 Numerical Approximation and Results

We discretize the weak formulation of the balance laws (1a–c) using simplicial finite elements. In each time-step the system is decoupled using a semi-implicit Euler scheme. The resulting subsystems are solved by Newton’s method, where the stress increment is eliminated to avoid saddle point problems. For the implementation we used the toolbox \textit{pdelib} (cf. [4]), which has been developed at WIAS, and extended it by components for the treatment of multidimensional mechanical problems. For details, we refer to [2] and [3].

Figures 3–5 show results of a numerical simulation for two weld pieces made of steel. Figure 5 depicts the beginning of the melting process. The right-hand picture shows that there is already a region close to the projection tip with hardly any shear stresses. Roughly speaking, this is the region that will be squeezed together by the applied force \( F \). To take care of this effect
Fig. 3. Initial condition $T_0$, $\sigma_{22}$ and $\sigma_{12}$

Fig. 4. A snapshot of the discharging current $Q_2$, plotted against time, the electric potential $\phi$ and temperature $T$ during an early stage of the heating process

Fig. 5. $T$, $\sigma_{22}$ and $\sigma_{12}$ during the start of the melting
Table 1. Variation of the maximal width of the liquid zone depending on the projection shape parameters

<table>
<thead>
<tr>
<th>Width of projection</th>
<th>Height of projection</th>
<th>projection tip angle</th>
<th>max. width of liquid zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 mm</td>
<td>1.6 mm</td>
<td>60°</td>
<td>0.7 mm</td>
</tr>
<tr>
<td>0.7 mm</td>
<td>1.3 mm</td>
<td>55°</td>
<td>1.4 mm</td>
</tr>
<tr>
<td>1.2 mm</td>
<td>1.1 mm</td>
<td>43°</td>
<td>0.2 mm</td>
</tr>
<tr>
<td>1.6 mm</td>
<td>1.0 mm</td>
<td>28°</td>
<td>not melted</td>
</tr>
</tbody>
</table>

seriously, we will have to update the grid topology during the process and to evaluate the fields on the deformed topology. The corresponding changes of the algorithm and their implementation are subject of current research in our group.

However, already in the present stage of implementation, it is possible to make some parameter studies. Table 1 describes the effect of changing the shape of the projection, characterized by its height, contact width and the angle of the projection tip, while keeping its volume constant. This means increasing the contact surface is accompanied by decreasing of the projection height and angle. The results show that there is an optimal configuration where the width of the liquid zone (i.e. \( T > T_m \)) is bigger than in all the other configurations.

References

8. Meyer, A., Grundlegende Untersuchungen zum Buckelschweißen von un­
legierten Stählen unter besonderer Berücksichtigung der Elektrodenbewegung,
Analysis of Transport Processes for Layered Porous Materials Used in Industrial Applications

H. Neunzert, A. Zemitis, K. Velten and O. Iliev

Institute of Industrial Mathematics, Gottlieb-Daimler-Str. 49, 67663 Kaiserslautern, Germany

Abstract. This work was aimed at the development of mathematical models and corresponding numerical solution and parameter estimation procedures which are needed as a basis for the computer-aided design of layered porous materials for industrial applications (e.g., hygienic products, technical textiles). The applications lead to nonlinear partial differential equations which must be solved in complex 3D geometries in many cases. Additionally, they may involve saturated/unsaturated flow, coupled flow and deformation problems, swelling particles, large jumps of the material parameters at the interfaces, convection dominance and complex boundary conditions. We introduce a generic mathematical model for layered porous materials, discuss some of the numerical aspects with an emphasis on 3D geometry description and present example applications.

1 The Need to Understand Layered Porous Materials

There are abundant industrial applications of layered porous materials: in the textile and filter industry, in various composite materials contexts etc. The great variety of applications reflects a great number of design alternatives, which is probably the most important characteristic of layered porous materials design. In fact, there is virtually no limit to the number of porous materials which can be combined in many ways to produce layered materials for each specific application. On the one hand, this is certainly an advantage in the sense that even for products which have been engineered over a long period of time there is still room for further substantial improvements. On the other hand, any strategy based on conventional trial-and-error methods is likely to end up with suboptimal solutions due to the complexity of the design problem. A real optimization of layered porous materials thus imperatively requires the application of professional tools.

It is, obviously, hard to decide about promising new product designs as long as it is not clear what kind of materials and material combinations generate what kind of local flows, transport processes and deformations, and how these local events contribute to the overall performance of the layered material. Therefore, the most important ingredient of any professional approach must be a thorough understanding of how layered porous materials work and how they achieve what they are expected to achieve in each of their industrial applications.
Mathematical models of layered porous materials provide this kind of understanding. Based on quantitative formulations of the relevant physical principles as well as a quantification of the materials in terms of material parameters, they allow to explore the working mechanisms of layered porous materials right in their process environment where they are used. In filter materials, for example, simulations based on mathematical models allow the development engineer to zoom into the filter during operation, to see how the liquid saturation in the filter develops, how it is influenced by the choice of materials, and how it influences filter efficiency.

In the porous media group of the ITWM, software tools for layered porous materials design based on mathematical models are being developed for various applications. The mathematics behind these software tools is quite challenging since it involves the solution of partial differential equations in complex 3D geometries in many cases, as well as saturated/unsaturated flow, coupled flow and deformation problems, swelling particles, large jumps of the material parameters at the interfaces, convection dominance and complicated boundary conditions.

Some of these numerical aspects have been treated in cooperations with the groups of A. Buikis (Latvian University, Riga) and R. Ciegis (Technical University of Vilnius, Lithuania). In our cooperation with A. Buikis, the emphasis was on some problems arising at the boundaries and interfaces of multi-layered porous materials, while in the cooperation with R. Ciegis an acceleration of the 3D-computations was achieved using specifically designed MPI-based parallel algorithms. In this report, we will shortly sketch our modelling and numerical approaches and give a few example applications.

2 A Generic Mathematical Model

We have formulated a generic mathematical model that includes all aspects of layered porous materials that we have found to be relevant for the understanding of layered porous materials performance in industrial applications. While a thorough treatment of this model is beyond the scope of this text, we will at least sketch some of its most important features.

In many applications, the generation and control of fluid flow within layered porous materials is critical for their performance. Frequently, saturated flow conditions (in areas where all pores of the material are liquid-filled) as well as unsaturated flow conditions (in areas where the pores of the material are partially liquid-filled) must be taken into account simultaneously. External forces and/or capillary forces generate the liquid flow. If the resistance to gas phase flows is negligible, a suitable mathematical model can be expressed e.g. in the form of the Fokker–Planck-equation [1]:

\[ C_i \frac{\partial \psi_i}{\partial t} + \nabla \cdot (K_i(w_i) \nabla (\psi_i - z)) = f_i, \quad \text{in} \quad \Omega_i \]
Here, \( \Omega_i \) is the domain occupied by the \( i \)-th material domain (e.g., a material layer or a porous particle embedded into a material layer), \( \psi_i = -\frac{p_i(w_i)}{\rho g} \) is the macroscopic suction in the \( i \)-th material domain (\( g \): acceleration of gravity, \( \rho \): liquid density), \( p_i(w_i) \) is the average liquid pressure in the \( i \)-th material domain (its dependence on the saturation \( w_i \) is a constitutive property of the materials and may change at the interfaces between the material domains), \( w_{0i} \leq w \leq m_i \) is the liquid content in the \( i \)-th material domain \( (w_{0i}: \text{irreducible wetting fluid content}, m_i: \text{porosity}), K_i(w_i) \) is the effective hydraulic conductivity in the material domain \( i \) (its dependence on liquid content \( w_i \) is a constitutive property of the materials and may change at the interfaces between the material domains), \( z \) is the distance to some fixed horizontal plane, \( f_i \) is a source or sink term, and \( C_i = \frac{d w_i}{d \psi_i} \) is the water capacity (it can be derived from the constitutive relation \( p_i(w_i) \)).

Denoting the interface between two material domains by \( \Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j \), the coupling conditions can be formulated as \( \left( \frac{\partial}{\partial n} = \text{normal derivative} \right): \)

\[
[K(w) \frac{\partial}{\partial n}(\psi - z)]|_{r_{ij}} = 0 \quad (1) \\
[\psi]|_{r_{ij}} = 0 \quad (2)
\]

These conditions must be fulfilled on all \( \Gamma_{ij} \neq \emptyset \). Note that \([f]|_{r_{ij}}\) signifies the jump of \( f \) across the interface \( \Gamma_{ij} \). Additionally, appropriate initial and boundary conditions must be supplied to complete the formulation of the flow problem.

Another important aspect is the interaction between deformation and flow. Deformations may be caused either by external forces (e.g. a baby sitting in its diaper) or by changes of the interior structure of the material (e.g. swelling absorber particles in hygienic products). They can be critical for the functionality of layered porous materials. Materials containing fractions of textile and fleece materials, for example, are frequently being subjected to big deformations. Under operational conditions, these materials sometimes go through the entire range of porosities, between very porous (90 % porosity and more) and very much condensed states (porosities of a few percent only). As a consequence, the material parameters also undergo big changes. E.g., the range of the conductivity parameter \( K_i \) may cover several orders of magnitude.

The mechanics of layered porous materials may exhibit the whole range of linear and nonlinear behavior [2]. Therefore, the key step in the modelling of layered porous materials mechanics is a careful selection of appropriate constitutive laws. The coupling between the flow and deformation models is then established using some equation expressing the appropriate equilibrium condition between the forces exerted by the solid and liquid parts of the material (e.g. some version of the so called Terzaghi’s principle, cf. [3]).
3 Treatment of Complex Geometries in 3D

In many applications, the three-dimensional starting configurations of layered porous materials are parallelepipeds. Therefore, our first algorithms referred to parallelepipedic geometries. They were based on locally one-dimensional finite difference schemes ([4,5]) which allow for easy parallelization and time step control strategies.

Principally, more complex geometries can be handled very well in a finite element framework. Finite element grids allow for a precise approximation of the interfaces between the material domains of layered porous materials. Grid generation, however, is very time consuming. This is relevant here since many of the applications involve instationary boundaries and interfaces which would need very frequent re-meshing.

Fortunately, the newest developments in the area of level set-methods ([6–9]) opened the way towards the efficient solution of moving boundary problems on stationary, regular grids. We have adopted this approach for the time-dependent description of all internal and external boundaries of layered porous materials. Usually, we embed the general starting configuration into a parallelepipedic geometry according to the ideas of the fictitious domain method ([10,11]).

4 Examples

4.1 Hygienic Products

Regarding hygienic products such as diapers, sanitary napkins etc., the minimization of raw material use as well as an improved functionality are the most important issues in current product design. We have addressed these problems in several projects based on mathematical modelling and simulations. Here we will focus on the modelling of superabsorbers and their interactions with their surrounding media, and on some numerical problems related with instationary geometries typically encountered in hygienic product simulation.

Some mathematical models for diapers have already been introduced in [12,5]. In the presence of embedded swelling particles, however, these models must be modified. We will discuss here the modelling of porous media flow in the presence of superabsorbers (precisely, “Super Absorbent Polymers” (SAP), see [13]). Modern SAP materials can absorb a large amount of liquid and swell up to 2–3 orders of magnitude. Some attempts to model such processes can be found in [14,13] (see also the references therein). The most important aspects that have to be accounted for in modelling are the rate of water uptake from SAP particles, and the changes in the porous media characteristics due to the swelling. The absorption of the liquid by SAP materials is modeled by a sink term which is introduced into the mass balance equation for the liquid phase (i.e., in the Fokker–Planck equation (1)): $f_i = \frac{\partial \phi_{iw} \rho_w}{\partial t}$ where $\phi_{iw}$ is the volume fraction of the trapped liquid (the subscript $w$ stands for the liquid).
Analysis of Transport Processes for Layered Porous Materials 247

This treatment of the SAP interaction with the porous media flow is based on the fact that the Fokker–Planck equation accounts for the mass balance of the mobile liquid. At the same time, the liquid trapped by SAP becomes immobile. Furthermore, the question is how to model this sink term. It is clear that it depends on the number of SAP particles per unit volume, multiplied by the liquid trapping rate of a single SAP particle. A SAP particle can absorb some finite amount of liquid, and the rate of absorption depends on the amount of the liquid already absorbed. Thus, a first order model is suitable for modelling the rate of absorption. Assuming the SAP particle is a sphere with radius $r_{\text{SAP}}$, we have

$$\frac{\partial r_{\text{SAP}}}{\partial t} = \frac{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}}{\tau_{\text{SAP}}} \left[ S_w - 1 + \frac{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}}{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}^{\text{dry}}} \right]_+ \tag{3}$$

where $[\cdot]_+$ means that this expression is set to zero when it is less than zero. The last factor is introduced in order to account for SAP interaction with unsaturated flow: if the liquid saturation within a SAP particle is larger then the liquid saturation in its vicinity, no absorption takes place. The notation $\tau_{\text{SAP}}$ stands for a parameter of absorption, which depends on the coefficient of diffusivity of water in SAP.

Thus, the ratio of the mass of the entrapped (absorbed) liquid is

$$\frac{\partial \phi_{\text{tw}} \rho_w}{\partial t} = N_{\text{SAP}} \rho_w \frac{4\pi r_{\text{SAP}}^2}{1} \frac{4\pi r_{\text{SAP}}^2}{1} \frac{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}}{\tau_{\text{SAP}}} \left[ S_w - 1 + \frac{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}}{r_{\text{SAP}}^{\text{max}} - r_{\text{SAP}}^{\text{dry}}} \right]_+$$

Here, $N_{\text{SAP}}$ stands for the number of SAP granule per unit volume (we suppose it does not change in time), and $r_{\text{SAP}}$ is computed from the previous equation.

Next, let us discuss the effects of swelling itself. The swelling causes the solid skeleton of the porous materials to move. Its velocity can be computed using some elastic or viscoelastic model, or in a simplified case, using the mass balance of the solid phase. When the swelling is significant, it changes the porosity, density of the solid skeleton, etc., and in this way it influences all constitutive relations, such as the capillary pressure-saturation curve, conductivity, etc. The swelling of SAP particles can be completely compensated for by the change of the porosity, or in some cases it can also cause an expansion of the multi-layered porous material. We account for these effects in our models. One can observe the influence of SAP particles swelling on the propagation of the wetting front in Fig. 1. As the figure shows, a SAP layer can cause reduced liquid saturations in the neighboring layers, but it can also hinder or even stop the flow in the case of large swelling and an induced significant reduction of porosity.
Fig. 1. Saturations (abscissa) at different times as a function of the diaper width (ordinate). Left picture: no swelling, right picture: swelling changes the porosity. The layer containing SAP is located between 0.35 and 0.4 on the ordinate.

Fig. 2. Flow simulation in diapers: liquid distribution some time after the invasion of liquid through the top layer (the colors scaled between light blue/white refer to low/high liquid saturations).

In our simulations of hygienic products, geometries have been described by local level set functions [6–9]. These functions are also involved in the approximation of the equations based on finite differences and regular grids ([15,16]). Figure 2 shows an example where we have simulated a part of a diaper consisting of 3 layers (the liquid comes from the top layer, gravitation is acting from left to right).
4.2 Pressing of Technical Textiles

Figure 3 shows a situation where a multi-layered and partially fluid filled technical textile is being compressed by some pressing device. This kind of pressing processes arise in different industrial contexts (e.g., in order to squeeze out fluid from the textile). To optimize such a process, it is necessary to understand how fluid flow is generated in the various layers of the textile. This can be investigated in fluid flow models derived from our generic model described above.

**Fig. 3.** Pressing of a multi-layered technical textile: the textile moves from left to right with velocity $v$ and is being compressed by the pressing device in $[a, b]$.

**Fig. 4.** Fluid pressures initiated by a pressing device in a four-layer technical textile. The lines across the pressure surface indicate the layer boundaries.
Fig. 5. Comparison between the saturation distributions initiated by two rotating rolls in a four-layer technical textile for a small (left picture) and a high rolling velocity (right picture). For the high rolling velocity, the textile and roll surfaces separate at the downstream side of the rolls due to the viscoelastic behavior of the textile.

As an example, Fig. 4 shows a pressure distribution that was generated by some pressing device in a four-layer technical textile (the layer boundaries are indicated by the lines across the pressure surface). As can be seen, the pressure device (which was located approximately between $x = 170$ mm and $x = 230$ mm) initiates a strong pressure gradient across the layers of the textile.

All relevant quantities for process optimization can be obtained from the computations, including e.g. saturation distributions in the textile layers under different process conditions (Fig. 5).

The example in Fig. 5 also shows how the interaction of flow and deformation processes determines the performance of the multi-layered textile. As a result of the flow model computation, it is seen that a fully saturated zone develops between the rolls for high rolling velocities. The evolution of this fully saturated zone interacts with the results of the deformation model computations, which lead e.g. to a viscoelastic separation of the textile and roll surfaces at the downstream side of the rolls.

References


Modelling and Numerical Simulation of District Heating Networks with Time-Saving Solution Methods

Rolf Dieter Grigorieff and Ralf Köcher

Fachbereich Mathematik, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany
E-mail: grigo@math.tu-berlin.de, URL: http://www.math.TU-Berlin.de/numerik/

Abstract. The main aim of the project is the thermo-hydraulic calculation of district heating networks with an incompressible feed medium by time-saving numerical methods. The modelling of the hydraulic and thermal relations of the various components of the heating network are based on earlier studies on this topic. Graph-theoretical methods are used to describe the topology of the heating system, in which spanning trees play a crucial role. The solution concept is based on sparse-matrix techniques with special data structures, which have been fitted to the particular problem. The main interest of the district heating supplier is a long-time simulation of the heating network combined with an optimization of the heat distribution and the producer’s service. They are important for the power station to plan things like fuel supply and intervals of maintenance more efficiently. In the already existing computer codes an arising problem was the need to accelerate the simulation in order to obtain results in an acceptable amount of time on common computers.

The obtained simulation tool makes the calculation of the system states in acceptable short times possible. Furthermore, it approaches more realistically the behaviour of the heating system by considering a time-dependent consumer demand (heat load). Questions regarding the main operation and the flow time behaviour of the heat distribution of the system are also examined.

This project was supported by a grant from the german BMBF ministry and carried out in collaboration with the Bewag company in Berlin.

1 Formulation of the Problem

The aim of this project is the implementation of an effective simulation tool for the thermo-hydraulic calculation of district heating systems. Some problems concerning the performance characteristics will also be addressed.

There exists already a couple of codes for calculating stationary hydraulics; however there are only few that take the interaction of the thermal and hydraulic processes into consideration. One code for stationary hydraulic calculations is NESSY. This code performs calculations in an acceptable amount of time. BoFrIT represents the state of the development in the field of thermo-hydraulic simulation in regard to the complexity of questions to be solved and the range of applications. Different aspects of district heating
systems can be simulated with this code. The main problem with BoFiT is the amount of time that is consumed in actual calculations.

The main topics of the present project are:

- Development and computer implementation of a simulation tool for the thermo-hydraulic calculation of district heating systems with an incompressible feed fluid that fulfills the following requirements:
  - a simple fitting of the format of the data input, so that it can be easily used for different systems,
  - the ability of the code to be run on common computers,
  - code development based predominantly on public domain software,
  - the consideration of practical and relevant operation aspects (e.g. given consumer demand),
  - the use of mathematically reliable and quickly performing solution methods,
  - the development of a solution concept exploiting special features of the structure (sparse-matrix techniques),
  - validation by real data,
  - a meaningful representation of the simulation results,
- Examination of special properties of the system (flow time behaviour),
- Simulation of the system with integrated consumer models,
- Development of a solution method for the cost effective construction of networks,
- Transfer of the simulation tool into an optimization procedure for the planning of power stations.

1.1 Basic Principles

District heating systems are composed of segments (lines) and junctions (connections). Oriented graphs are used to describe such systems, in which all tracks and connections are shown as edges and nodes. The topology of the system is mapped by incident matrices. The tracks are represented by pre-defined data structures, in which the start and end node of the lines and their physical parameters are saved. The node-edge incident matrix $A$ is determined by the existing structure of the system, while the loop-edge incident matrix $B$ is defined with the help of the spanning tree of the system graph. The algorithm for finding the loops in the network is based on the breadth-first search. As a result always as small loops as possible are obtained. The algorithm is based on plausible strategies because exact ones are not easy to formulate. The use of incident matrices allows the simplest and most efficient description of the structure and the interaction of the components of the system. In analogy to Kirchoff’s laws for the electrical theory of systems, the equations

$$A \dot{\mathbf{m}} = 0 \quad \text{and} \quad B \Delta \mathbf{p} = 0$$

(1)
are valid for heating systems reflecting the mass and energy balance. The solution of (1) yields a directed mass flow vector $\tilde{\mathbf{m}}$. Based on the known $\tilde{\mathbf{m}}$ the calculation of the temperature distribution in the heating system is performed. For this purpose the convection equation with the inclusion of heat losses through the pipe insulation

$$\frac{\partial T}{\partial t} = -\frac{\dot{m}}{\rho A} \frac{\partial T}{\partial x} - \frac{kU}{\rho A c_p} (T - T_U)$$

(2)

is used in the pipes. Pipes form the main component of a district heating system.

1.2 Mathematical Modelling

Changes in pressure spread so quickly in the system that stationary hydraulic calculations are carried out. In contrast, the velocity of propagation of temperature is 1000 times smaller. Because of their strong interaction hydraulic and thermal calculations are coupled. If the calculation is limited to the hydraulic part only (as it is the case in some codes) loads that depend on the local temperature or the time-dependent behaviour of consumers can be considered only inaccurately.

There are simple mathematical models for the components pipe, valve and pump of the form $\Delta p \sim \dot{m}^2$ (see Fig. 1). For consumers no such convenient models are available. They are connected to the heating system via a flow control ($\dot{Q} \sim \dot{m}$) by heat transfer stations.

**Hydraulic Calculation:** A number of practical requirements must be taken into account in the simulation. The demands of the consumers are considered by the equation

$$A_R \tilde{\mathbf{m}}_0 = \mathbf{L}.$$  

(3)

The right hand side $\mathbf{L} \in \mathbb{R}^{(k-s)}$ ($k \equiv \#\text{nodes}$) denotes the given consumer load and $A_R \in \mathbb{R}^{(k-s) \times n}$ ($n \equiv \#\text{edges}$) is derived from the node-edge incident
matrix by deleting $s$ rows. Here $s$ is the number of nodes which are directly connected to a heating plant. The resulting system consists of $k-s$ equations and $n-r$ ($r \equiv \# \text{ of loops}$) actual unknowns, because the components of $\mathbf{\dot{m}}_0$ that belong to the $r$ connecting edges of the complete graph are set to zero. The solution of (3) gives the mass flow in the edges of the spanning tree. The system (3) is for $s = 1$ square, because no parallel edges (two different edges directly connected to the same nodes) are allowed and hence a spanning tree of the complete graph always consists of $k$ edges and $k-1 = n-r$ nodes.

The method of loop calculation is obtained from Kirchoff's 2nd law. The advantage of this method in comparison to the method of edge calculation (Kirchoff's 1st law) is that the dimension of the problem is significantly reduced (generally for heating systems $r \ll k$). With $\Delta \mathbf{p} = \mathbf{f}(\mathbf{\dot{m}})$ the resulting system of equations is

$$B \mathbf{f}(\mathbf{\dot{m}}) = 0,$$

where $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$ describes the pressure loss vector. The entries of $\mathbf{f}$ are specified by the type of edge (pump, valve, pipe). The loop flow vector is obtained as the collection of the mass flows through the connecting edges $\mathbf{\dot{m}}_M = \{\dot{m}_{m_1}, \ldots, \dot{m}_{m_r}\} \in \mathbb{R}^r$. The components of the loop flow vector $\dot{m}_{m_i}$ are the unknowns which have to be determined. The mass flow $\mathbf{\dot{m}} \in \mathbb{R}^n$ in the system is given by

$$\mathbf{\dot{m}} = \mathbf{\dot{m}}_0 + B^T \mathbf{\dot{m}}_M,$$

where $\mathbf{\dot{m}}_0$ is the solution of (3). Since $A_R \dot{m} = A_R (\mathbf{\dot{m}}_0 + B^T \mathbf{\dot{m}}_M) = A_R \mathbf{\dot{m}}_0 = \mathbf{L}$ (note that $A_R B^T = 0$), $\dot{m}$ satisfies the mass balance in the consumer load nodes. By substituting (5) into (4) the system of equations for the mass flow in the network is obtained.

The resulting algebraic system is solved with a damped Newton method. The iteration is obtained from the linearization of

$$F(\mathbf{\dot{m}}_M) := B \mathbf{f}(\mathbf{\dot{m}}_0 + B^T \mathbf{\dot{m}}_M), \quad F: \mathbb{R}^r \to \mathbb{R}^r,$$

and, including a damping, has the form

$$\mathbf{\dot{m}}_M^{k+1} = \mathbf{\dot{m}}_M^k - \lambda_k \mathbf{d}^k, \quad \text{where} \quad J^k \mathbf{d}^k = F(\hat{\mathbf{m}}^k).$$

$J^k \in \mathbb{R}^{r \times r}$ is the Jacobi-matrix with elements

$$J_{i,j} = \frac{\partial F_{i}}{\partial \dot{m}_{m_{ij}}}, \quad i, j \in [1, \ldots, r],$$

evaluated in $\mathbf{\dot{m}}_{M}^k$ and $\lambda_k \in \mathbb{R}^+$ denotes the damping factor. The determination of $\mathbf{d}^k$ in (6) may consume considerable calculation time, both for the matrix elements of $J$ and the solution of the linear system. The use of exact analytical derivatives is possible because all pressure losses are of the form

$$f_i(\dot{m}_i) = \Delta p_i(\dot{m}_i) = -a_1(i) + a_2(i) \dot{m}_i |\dot{m}_i|$$
with certain $a_1(i), a_2(i) \in \mathbb{R}^+$. For calculating $J$ first the vector $f' \in \mathbb{R}^n$ is determined by

$$f'_i = a_2(i)|\dot{m}_i| = \begin{cases} 2 R |\dot{m}_i| & : \text{if edge } i \text{ is a pressure loss} \\ 2 c |\dot{m}_i| & : \text{if edge } i \text{ is a pump} \end{cases}$$

Then

$$J = B \text{diag}(f') B^T. \quad (7)$$

Non-diagonal elements in $J$ appear only where loops have common lines. The matrix $J$ is symmetric and positive definite, but in general ill-conditioned. These properties have to be taken into account when selecting an algebraic solver.

**Thermal Calculation:** The main source for thermally changing processes in the heating system is the piping. The axial heat transport and the radial heat loss are described by Eq. (2), where $\kappa$ is the heat transfer coefficient, $U$ the pipe circumference, $A$ the cross-sectional area of the pipe, $c_p$ the specific heat of the fluid, $T_U$ the environmental temperature and $\dot{m}$ the mass flow through the pipe component. The following is assumed:

- ideal pipe flow (no back mixing),
- neglecting of the production of heat by mechanical energy,
- a constant heat transfer coefficient $\kappa$.

An exact calculation of the heat transfer is realistically speaking hardly possible because the type and state of the insulation is rarely known. Initial conditions for Eq. (2) are the temperature at the nodal inlet. The nodal temperatures are calculated with a balancing of the energy in every node according to the law of mixture

$$T_{\text{node}} = \frac{\sum_{\text{input}} \dot{m}_i T_{\text{edge}}}{\sum_{\text{input}} \dot{m}_i}. \quad (8)$$

It is assumed that the mixture is instant and complete. Of course, before performing the mixture at a given time all edge temperatures must have been already calculated. The remaining net components (valve, pumps) are thermally modelled by assuming a linear behaviour (see [4]).

The entire process of thermal calculation is described as follows:

- **Fluctuating part:** Calculation of the edge temperature with the aid of the convection equation (2) with a modified explicit upwind scheme of time differentiation that preserves also for large time steps the property of the differential equation to conserve the heat energy. Extremely short pipes are eliminated from the system by considering them as direct connections (which is a measure to allow for using longer time steps in the discretisation, see Sect. 2.2).

- **Stationary part:** Calculation of the nodal temperatures with the mixture equation (8).
2 Numerical Results

Successful calculations were made for two existing systems in the states of Berlin and Brandenburg, one in the city of Rheinsberg and one in the southern part of Berlin (run by Bewag). The Bewag system is a three pipe system composed of different pipe types, denoted by: 'HZG', 'RL', and 'KLB'. The results of the hydraulic and thermo-hydraulic calculations for the Bewag system are presented in the following sections. The given calculation times refer to a SGI Indy workstation with a R 5000 processor.

2.1 Hydraulic Calculation

As a first calculation we report the mass flow distribution of the HZG pipe system with almost 1600 pipe components (11 loops) driven by three pumps, where a specific consumer load is prescribed. The calculation time was well under one second. The obtained pressure diagram is shown in Fig. 2. The units of the axes are local geographical coordinates. Similar results were determined for the two other pipe systems (RL with 12 and KLB with 11 loops). A coupling of the three pipe system was imposed in node 'LI' (input

![Pressure diagram of the HZG pipe system of the Bewag system](image)
node), so that the complete Bewag system with about 4650 segments and 34 circuits was obtained. The calculation time was still under two seconds. In all cases the operating states were determined without any pre-information on the actual state, which is numerically speaking the most difficult situation. The number of necessary iterations could be kept at a minimum ($k < 11$) for all pipe systems due to enhanced graph-theoretical algorithms and efficient solution concepts. The calculation times for the HZG pipe system are shown in the following table. The criterium for terminating the iteration was $\|d^k\| < 10^{-2} \text{ kg/s}$ and $\|f(\dot{m})\| < 10^{-5} \text{ Pa}$.

<table>
<thead>
<tr>
<th>HZG pipe system</th>
<th># of iterations</th>
<th>calculation time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. with consumer load</td>
<td>10</td>
<td>0.3 s</td>
</tr>
<tr>
<td>2. without consumer load</td>
<td>6</td>
<td>0.2 s</td>
</tr>
<tr>
<td>3. data input</td>
<td></td>
<td>1.3 s</td>
</tr>
<tr>
<td>4. data output</td>
<td></td>
<td>0.3 s</td>
</tr>
</tbody>
</table>

The results of the calculation of the complete system are summarized in the following table. The criterium for termination was the same as above. Information about the expected state solution was, as before, not available to accelerate the iteration.

<table>
<thead>
<tr>
<th>complete Bewag network</th>
<th># of iterations</th>
<th>calculation time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. with consumer load</td>
<td>11</td>
<td>1.5 s</td>
</tr>
<tr>
<td>2. without consumer load</td>
<td>6</td>
<td>0.8 s</td>
</tr>
<tr>
<td>3. data input</td>
<td></td>
<td>6.5 s</td>
</tr>
<tr>
<td>4. data output</td>
<td></td>
<td>1.3 s</td>
</tr>
</tbody>
</table>

Validation of the results was done with already validated simulation results from another simulation code (NESSY). The maximum deviations did not exceed 0.03 $m^3/h$ and 0.17 bar for the KLB pipe system and 0.25 $m^3/h$ and 0.02 bar for the RL pipe system. A significant difference of 5 $m^3/h$ was only found for the HZG pipe system, that could be explained by an insufficient approximation of the characteristic curve of the pump. A smaller deviation of maximal 1 $m^3/h$ was already reached by exchanging one of the supporting points in the approximation of the pump characteristic. Altogether, the relative error for the pressure was $\varepsilon_{rel}(\Delta p) \leq 0.24\%$ and for the mass flow $\varepsilon_{rel}(\dot{m}) \leq 0.038\%$. These results indicate that the code seems to be well suited for a realistic simulation.

### 2.2 Thermal Calculation

The thermal state of the system can be determined if the mass flow is known as a function of time. The time discretisation of (2) was based on an explicit
finite difference method. Implicit methods did not sufficiently determine the temperature when the time step was large. The use of an explicit method imposes a restriction on the length of the maximal allowed time step for maintaining the stability of the scheme (so-called CFL condition). It took about two minutes to calculate the temperature in the HZG pipe system for a simulated period of 10 minutes. The relatively long calculation time is explained by the small maximal time step $\Delta t = 0.03\, s$ which comes from satisfying the CFL condition in short pipes (edges 1150, 1211, 1240) in which the mass flow is large. To accelerate the calculation of the temperature a significant improvement was obtained by replacing short pipes with a high flow velocity in the network by direct junctions. For the HZG pipe system it was sufficient to directly connect 14 of 1561 segments to increase the admissible time step to three seconds. In this way the calculation time was reduced from 107 s to about 7 s. The influence of the larger time step on the error was small (less than one percent). The maximum change in temperature was under 1 $K$.

The following table shows a comparison of the calculated temperatures in three selected nodes after 10 minutes of simulated time using two different time steps.

![Graph showing temperature distribution in the HZG pipe system of the Bewag after one hour starting from 20°C uniform net temperature](image-url)

**Fig. 3.** Temperature in the HZG pipe system of the Bewag after one hour starting from 20°C uniform net temperature
<table>
<thead>
<tr>
<th></th>
<th>T with $\Delta t = 0.03,s$</th>
<th>T with $\Delta t = 3,s$</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>node '384'</td>
<td>37.76</td>
<td>36.85</td>
<td>$\approx 0.9K$</td>
</tr>
<tr>
<td>node '1305'</td>
<td>51.14</td>
<td>50.52</td>
<td>$\approx 0.6K$</td>
</tr>
<tr>
<td>node '1307'</td>
<td>38.21</td>
<td>37.32</td>
<td>$\approx 0.8K$</td>
</tr>
</tbody>
</table>

For the calculation of the temperature in the HZG pipe system the complete system was initially filled with $20^\circ C$ and the environmental temperature was assumed to be $13^\circ C$. Fig. 3 shows the temperature distribution after one hour of simulated time under the assumption, that the power station feeds always $75^\circ C$ in the area of Lichterfelde of Berlin (node ‘LI’).

A validation of the thermal calculation with real data was not possible because of lack of measured data for the existing network. However, the results of the thermal calculation are plausible with respect to the flow time behaviour.

### 2.3 Determination of the Flow Times in the Network

District heating suppliers are interested in knowing how much time it takes that the producer injected heat reaches the consumer. Facts about the flow times in the system can be concluded from the instationary thermo-hydraulic simulation of the temperature distribution in the network. The results of a calculation for the original and the simplified pipe system (see Sect. 2.2) show after 35 minutes of simulated time a difference of 1.16% (corresponding to 18 consumer nodes).

![Graphs showing flow times comparison between HZG and KLB systems](image-url)

**Fig. 4.** Left HZG pipe system, right KLB pipe system
In Fig. 4 the flow times for the Bewag pipe systems HZG and KLB are presented. The significantly longer flow times for the KLB pipe system can be explained by its substantially smaller mass flows due to smaller consumer demands. According to Bewag, the shown curves reflect the real behaviour, which is a further hint to the reliability of the thermal calculation.

### 2.4 System Simulation with Integrated Consumer Models

Another study included the presence of consumers that are connected to the network. The time-dependent consumer behaviour was modelled by given profiles. Four characteristic classes of consumers were defined and connected to the system in given nodes. Assuming that the correlation between the environmental temperature and the heat demand is known the necessary mass flow for covering the consumer needs can be calculated in dependence of the forward temperature $T_{VL}$ of the system and the given return temperature ($T_{RL} \equiv 50°C$) via the following equation

$$\dot{m} = \frac{\dot{Q}}{c_p (T_{VL} - T_{RL})}.$$

The nodal consumer load $\dot{Q}$ corresponds to a mass flow to the consumer. The mass flows for one week are presented in Fig. 5. They depend on the given consumer profiles and the environmental temperature. The left figure shows

![Fig. 5. Demand during one week](image-url)
the four different consumer profiles and the right figure shows additionally how a decrease in the environmental temperature affects the mass flow.

Our results show that the time dependent behavior of the consumers can be integrated into the simulation without problems. The described method is practical for a simple representation of the system in which the consumers are summed up into a small number of typical groups. In this case, the individual behaviour of each consumer is negligible. The simulation of a small system consisting of 3 loops for one year was done in 26 minutes, where the change of the environmental temperature was taken into account at an hourly rate.

3 Outlook

By virtue of the reached short simulation times for calculating the state of district heating systems (in a few minutes) we hope to be able to solve optimization problems arising in this context for already existing systems as well as for the design of new networks. At the moment we are working on a procedure for the design of a new network at optimal costs, both for the construction and the net operation. The first results are promising.

It is planned to integrate our simulation tool into the code PLINST used by Bewag for the operation of power stations. The compound code will be employed to study the feasibility of newly designed power stations.

Finally, the graphical user interface has to be improved. Currently only a mask of input is implemented. A higher developed code surface is important to make the simulation tool attractive for possible users.

References

6. Oeljeklaus, G.: Thermohydraulische Berechnung vermaschter Fernwärmeheizwassernetze; Düsseldorf: VDI Verlag 1987
IV. Optics and Sensors

Adaptive Multigrid Methods for the Vectorial Maxwell Eigenvalue Problem for Optical Waveguide Design
    P. Deuflhard, F. Schmidt, T. Friese, L. Zschiedrich

Direct and Inverse Problems for Diffractive Structures – Optimization of Binary Gratings
    J. Elschner, R. Hinder, G. Schmidt

Computation of Electromagnetic Fields for a Humidity Sensor
    W. Hackbusch, S. Börm
Sensitivity and Robustness Analysis for Construction and Monitoring of Turbine-Generator Shafts

D. Prätzel-Wolters¹, P. Lang¹, A. Wirsen¹, and S. Kulig²

¹ Institut für Techno- und Wirtschaftsmathematik, Kaiserslautern, Germany
² Lehrstuhl für elektrische Maschinen, Antriebe und Leistungselektronik, University Dortmund, Germany

1 Problem Description

Sensitivity and robustness analysis are important tools in order to guarantee an efficient construction and error free running of a turbine generator shaft. It is very important to test different design parameters simultaneously before the machine is in use. Therefore, the knowledge of the characteristic frequencies of the system is crucial since they should be sufficiently different from the excitation frequency, e.g. the electrical main frequency. Often the important system parameters are inexact or hardly known and so the resulting model, which describes the dynamical behaviour of a turbine generator shaft, differs significantly from reality. Therefore one has to analyze how large the permissible errors in the critical design parameters are, so that a smooth running of the system is still assured. Besides estimating the influence of parameter variations, one also requires a suitable model adaption based on real measured dynamical system parameters. Such a well-adapted shaft model then is the basis for the monitoring of the machine during its life-time.

2 Turbine Generator Shaft

The shaft line of a typical turbine-generator shaft from the medium or large power class (800 MW) consists of a high pressure HP, a medium pressure MP, one or several low pressure turbines LP1, LP2, as well as a generator GEN and an excitator EXC. The structure of such a machine is shown in Fig. 1. The driving steam torques \( m_{HD}, m_{MD} \) and \( m_{ND} \) are compensated by the generator momentum \( m_{EL} \) and the torsional moments \( m_{\tau_i} \), which are proportional to the angular deflections \( \Phi_i \).

2.1 Finite Element Model

The shaft dynamics can be described by a partial differential equation [11] that is transformed by the finite element method into a \( n \)-dimensional system of second order ordinary differential equations. The finite element discretization of the shaft line hereby is prescribed through constructional restrictions,
i.e. it can not be chosen with respect to numerical aspects. Adding an output equation for computation of the torsional moments, we obtain:

\[ J\ddot{\phi}(t) + K\phi(t) = Bu(t) ; \quad y(t) = C\phi(t) \quad (1) \]

Here, \( B \in \mathbb{R}^{n \times m} \) denotes the input matrix for the external moments \( u(t) \) and \( C \in \mathbb{R}^{p \times n} \) transforms the angular deflections \( \phi(t) \) into torsional moments \( y(t) \) at the interesting shaft locations. The inertia matrix \( J \in \mathbb{R}^{n \times n} \) is symmetric and positive definite, whereas the stiffness matrix \( K \in \mathbb{R}^{n \times n} \) is symmetric and positive semidefinite. For typical turbine generator shafts the dimension will be \( n \approx 200 \).

It is well-known, that due to their special properties, the matrices \( K \) and \( J \) can simultaneously be diagonalized:

\[ V^T KV = \Lambda^2 ; \quad V^T JV = I_n \]

The diagonal matrix \( \Lambda \in \mathbb{R}^{n \times n} \) then contains the eigenfrequencies \( \lambda \), the columns \( v_i \) of the matrix \( V \) are the corresponding analytical mode shapes and \( I_n \) denotes the \( (n \times n) \)-identity matrix. Substituting \( \phi(t) = Vx(t) \) in equation (1), therefore leads to the following system of equations:

\[ \ddot{x}(t) + \Lambda x(t) = V^T Bu(t) ; \quad y(t) = CV x(t) \quad (2) \]

Damping, which was neglected in (1), is introduced in terms of modal damping coefficients. Augmenting (2) by a positive definite diagonal damping matrix \( \bar{D} \in \mathbb{R}^{n \times n} \) leads to:

\[ \ddot{x}(t) + \bar{D}\dot{x}(t) + \Lambda x(t) = V^T Bu(t) ; \quad y(t) = CV x(t) \quad (3) \]

### 2.2 Model Uncertainties

First, it should be noted that the introduction of a diagonal damping matrix \( \bar{D} \) in (3) refers to the assumption that the turbine generator shaft is classically
Sensitivity and Robustness Analysis

Damped. On one hand this assumption however is unrealistic for almost all real physical systems [1], [7], on the other hand it is the only way to incorporate damping since it is very difficult to obtain more detailed information about system damping in practice.

In general also the inertia and stiffness matrix are not given exactly, due to uncertainties in the physical system parameters as well as unmodelled dynamical effects. For example, bending oscillations in general are neglected when studying the torsional dynamics of turbine-generator shafts, but of course both kinds of oscillations interact in reality. Nevertheless, even if there are a lot of sources for uncertainty, the model is physical meaningful and it reflects the geometric structure of the system. Therefore, considering the class of matrices

$$\Sigma = \{ (\Delta J, \Delta D, \Delta K) : \Delta J^T = \Delta J \in \mathbb{R}^{n \times n}, J + \Delta J > 0, |\Delta J_{il}| \leq \alpha_{il} \}$$

with constants \( \alpha_{il}, \beta_{il}, \xi_{il} > 0 \), we assume that there exists one tuple of matrices \((\Delta J, \Delta D, \Delta K) \in \Sigma\), such that the real system can be described by \((J + \delta J, V^{-T} \dot{D} V^{-1} + \Delta D, K + \Delta K)\).

2.3 Measurements

Additionally to the physical and geometric data some eigenfrequencies \( \tilde{\lambda} \) and the corresponding mode shapes \( \tilde{v} \) of the system are known from measurements. More concretely, since the excitation of the turbine generator shaft typically is only possible in the low frequency domain, generally only the 5–7 lowest natural frequencies and the corresponding mode shapes are known from measurements. In the following the dimension of the measured eigen-system will be denoted by \( m \). On the other hand, our shaft model possess a \( n \)-dimensional eigensystem, where \( n \) is much larger than 100 in general.

Moreover, due to technical reasons it is only possible to measure the torsional oscillations at certain locations along the shaft line. Therefore, the dimension of the measured mode shapes, that will be denoted by \( r \), is much smaller compared to the analytical mode shape dimension. In practice it turns out however, that despite of the dimensionality problems it always makes sense to rely on this measured data for the model updating process.

2.4 Test Rig

To verify our results achieved in sensitivity analysis, robustness estimations, observer design and model updating by experiments a test rig was build up in the laboratory of our project partner Professor Kulig, University of Dortmund, in 1997.
Since the torque sensors for the test rig for comparison reasons had to be the same as for the real turbine generator shaft, a certain diameter range for the test rig shaft line was prescribed. Therefore, in order to obtain measurable torsional oscillations, rather big masses had to be added to the shaft [3]. Figure 2 shows a photo and the outline of the resulting test rig.

A FE discretization using 33 bars results in a tridiagonal inertia $J \in \mathbb{R}^{34 \times 34}$ and a tridiagonal stiffness matrix $K \in \mathbb{R}^{34 \times 34}$ which are shown in Fig. 3.

### Fig. 3. Inertia (left) and Stiffness matrix (right) of the test rig

### 3 Sensitivity and Robustness Considerations

In Sect. 2.2 we discussed the various sources of uncertainties for the turbine generator shaft model. Therefore the question arises in how far the behaviour of the model changes due to variations in certain model parameters. In the
following we first study the influence of parameter variations on the eigen-system of the shaft model. The second subsection is dedicated to estimates for the changes caused in the corresponding transfer function of the model.

3.1 Eigensystem of the Shaft Model

In order to get a first idea in how far variations of certain system parameters $\Theta_1, \ldots \Theta_L$ effect the eigensystem of the shaft model, the first partial derivatives of the eigenfrequencies and the eigenmodes with respect to these parameters are computed [5], [9]. If damping is neglected, the analytical expressions for the eigenfrequency derivatives are given by:

$$\frac{\partial \lambda_i}{\partial \Theta_l} = \frac{1}{8\pi^2 \lambda_i} v_i^T \left( \frac{\partial K(\Theta)}{\partial \Theta_l} - 4\pi^2 \lambda_i^2 \frac{\partial J(\Theta)}{\partial \Theta_l} \right) v_i$$ \hfill (4)

The eigenmode derivatives are given as solutions of the following system of equations:

$$\left( K(\Theta) - 4\pi^2 \lambda^2 J(\Theta) \right) \frac{\partial v_i}{\partial \Theta_l} = - \left[ \frac{\partial K(\Theta)}{\partial \Theta_l} - 8\pi^2 \lambda_i \frac{\partial \lambda_i}{\partial \Theta_l} J(\Theta) \right] v_i$$ \hfill (5)

where $(K(\Theta) - 4\pi^2 \lambda^2 J(\Theta))$ obviously is a singular matrix. Numerically, the eigenmode derivatives therefore are usually obtained via the computation of the pseudo-inverse.

Note, that up to now the uncertain parameter $\Theta_l$ could be any physical system parameter. Depending on the structure of the system matrices and the kind of uncertain parameters considered, the expressions for the eigensystem derivatives may become much simpler. In the case of torsional shaft oscillations, the system matrices $J$ and $K$ are tridiagonal, resulting from a bar model of the shaft line, where each bar is represented in a 2x2-block. Therefore, consideration of uncertainties in the structural system parameters lead to a decoupling in the expressions from above. Eigenfrequency derivatives with respect to uncertain stiffness parameters then simply are given by:

$$\frac{\partial \lambda_i}{\partial k_l} = \frac{1}{8\pi^2 \lambda_i} v_i^T \frac{\partial K(\Theta)}{\partial k_l} v_i = \frac{1}{8\pi^2 \lambda_i} (v_{i,l} - v_{i,l+1})^2$$

whereas the derivatives with respect to inertia parameters result in:

$$\frac{\partial \lambda_i}{\partial j_l} = -\lambda_i v_i^T \frac{\partial J(\Theta)}{\partial j_l} v_i = -\lambda_i \frac{v_i^2}{6} (v_{i,l} + v_{i,k} v_{i,l+1} + v_{i,l+1}^2)$$

These formulas directly reflect the typical behaviour of a torsional system. An increase in the stiffness leads to an increase in the system frequencies, whereas an increase in the inertia of the system tends to decrease the frequencies.
Example 1. Figure 4 shows the sensitivity of three eigenfrequencies versus the stiffness parameters for a typical turbine generator shaft. Obviously, the eigenfrequencies are sensitive to different sections along the shaft line.

Although the sensitivity analysis discussed so far gives us hints about possible changes in the eigensystem caused by certain parameter variations, it turns out that the first derivatives are not enough in order to quantify those changes given concrete parameter disturbances. Therefore, we also derived analytical expressions for the second order partial derivatives of the eigenfrequencies:

$$
\frac{\partial^2 \lambda_i}{\partial \theta_j \partial \theta_l} = -v_i^T \left[ \frac{1}{\lambda_i} \frac{\partial \lambda_i}{\partial \theta_j} J(\Theta) + \frac{\partial \lambda_i}{\partial \theta_j} \frac{\partial J(\Theta)}{\partial \Theta_l} + \frac{\partial \lambda_i}{\partial \theta_l} \frac{\partial J(\Theta)}{\partial \Theta_j} \right] v_i 
+ \frac{1}{8\pi^2 \omega_i} v_i^T \left[ \frac{\partial K(\Theta)}{\partial \theta_l} - 8\pi^2 \lambda_i \frac{\partial \lambda_i}{\partial \theta_l} J(\Theta) - 4\pi^2 \lambda_i^2 \frac{\partial J(\Theta)}{\partial \theta_l} \right] \frac{\partial v_i}{\partial \theta_j} 
+ \frac{1}{8\pi^2 \lambda_i} v_i^T \left[ \frac{\partial K(\Theta)}{\partial \theta_l} - 8\pi^2 \omega_i \frac{\partial \lambda_i}{\partial \theta_l} J(\Theta) - 4\pi^2 \omega_i^2 \frac{\partial J(\Theta)}{\partial \theta_l} \right] \frac{\partial v_i}{\partial \theta_l}
$$

(6)

Using these expressions in a second order Taylor expansion often leads to significant better estimates of the disturbed eigenfrequencies compared to the linear approach.

Example 2. In this example the original model and the updated model of the test rig are used to show the significance of second order Taylor expansion. Hereby $\lambda_i$ denotes the $i$-th eigenfrequencies of the nominal model, $\hat{\lambda}_i$ the $i$-th one of the updated model and $\check{\lambda}_i$ and $\dot{\lambda}_i$ are the results of the linear and of the second order Taylor expansion for the $i$-th eigenfrequency, respectively. For every eigenfrequency the second order Taylor expansion is an improvement compared to the linear case.
<table>
<thead>
<tr>
<th></th>
<th>(i = 2)</th>
<th>(i = 3)</th>
<th>(i = 4)</th>
<th>(i = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda_i)</td>
<td>26.8005</td>
<td>48.3267</td>
<td>69.3077</td>
<td>120.7436</td>
</tr>
<tr>
<td>(\hat{\lambda}_i)</td>
<td>26.8217</td>
<td>48.4304</td>
<td>70.0906</td>
<td>122.9776</td>
</tr>
<tr>
<td>(\bar{\lambda}_i)</td>
<td>26.8072</td>
<td>48.4368</td>
<td>70.1156</td>
<td>123.0270</td>
</tr>
<tr>
<td>(\check{\lambda}_i)</td>
<td>26.8090</td>
<td>48.4363</td>
<td>70.1163</td>
<td>123.0284</td>
</tr>
</tbody>
</table>

It should be noted however, that especially for high dimensional systems the computation of the higher order derivatives becomes very time expensive, since for each parameter \(\Theta_i\) the sensitivity of all the mode shapes have to be computed first via the systems of equations (5).

Note that all the considerations of this subsection in principal also can be performed for damped systems.

### 3.2 Transfer Function

Besides the pure statical consideration of the effects of parameter variations on the eigensystem of the shaft model, the resulting changes in the model dynamics are very important. The torsional behaviour of a proportional damped shaft model in the frequency domain is described through the following transfer function

\[
G(s) = \sum_{k=1}^{n} \frac{1}{s^2 + s d_k + \omega_k^2} C_D^k B_k. \tag{7}
\]

Here, \(\omega_k\) denotes the \(k\)-th circular frequency, \(d_k\) the \(k\)-th modal damping coefficient, \(B\) the modally transformed input matrix and \(C_D\) the modally transformed displacement output matrix of the shaft model. If one considers now the disturbed system parameters:

\[
\begin{align*}
W_k &= W_k + \Delta W_k, \\
d_k &= d_k + \Delta d_k, \\
C_D &= C_D + \Delta C_D, \\
B &= B + \Delta B,
\end{align*}
\]

then the following estimate for the error between the nominal and the disturbed transfer function holds [10]:

\[
\| G - \tilde{G} \|_\infty \leq \sum_{k=1}^{n} \frac{\| \Delta C_D^k \|_2 \| \tilde{B}_k \|_2 + \| C_D^k \|_2 \| \Delta B_k \|_2}{\alpha_k} \\
+ \sum_{k=1}^{n} \frac{1}{s^2 + s d_k + \omega_k^2} - \frac{1}{s^2 + s \tilde{d}_k + \tilde{\omega}_k^2} \| \infty \| C_D^k \|_2 \| B_k \|_2,
\]

where:

\[
\alpha_k := \begin{cases} \\
\tilde{d}_k (\tilde{\omega}_k^2 - 0.25 \tilde{d}_k^2)^{\frac{1}{2}} & \text{if } (\tilde{d}_k \leq \sqrt{2} \tilde{\omega}_k) \\
\tilde{\omega}_k^2 & \text{else}
\end{cases}
\]
Here $\| \cdot \|_{\infty}$ denotes the $H_{\infty}$-norm [16], which is defined for the class of real rational stable transfer functions by:

$$\| G \|_{\infty} = \sup_{\omega \in \mathbb{R}} \bar{\sigma}(G(i\omega)),$$

where $\bar{\sigma}$ stands for the largest singular value of the matrix $G$. The errors between the modal subsystems further can be estimated by:

$$\frac{1}{s^2 + s \bar{d}_k + \omega_k^2} - \frac{1}{s^2 + s \bar{d}_k + 6\bar{\omega}_k^2} \|_{\infty} \leq \max \left\{ \Theta_1, \Theta_2, \frac{6 \sqrt{(d^2 + d^2)(3\mu^2 \phi + \mu^2)}}{d \bar{d} (\theta(16\mu^2 + 9\theta(d^2 + d^2)))^{\frac{1}{2}}} \right\},$$

where:

$$\Theta_1 = \begin{cases} \frac{\sqrt{\mu}}{\bar{d}(\omega^2 - 0.25d^2)^{\frac{1}{2}}} & \text{if } (\nu > 0) \\ 0 & \text{else} \end{cases} \quad \text{and} \quad \Theta_2 = \begin{cases} \frac{\sqrt{\mu}}{\omega^2 \bar{\omega}} & \text{if } (\mu > 0) \\ 0 & \text{else} \end{cases}$$

If uncertainty intervals for the system disturbances are given, a central problem consists in determining an optimal nominal model. Under the assumption that only damping uncertainty intervals $[D_l, D_r]$ are given, we approximately solved the corresponding saddle point problem [10]:

$$\min_{d \in I_d} \max_{d \in I_d} \| \frac{1}{s^2 + s \bar{d} + \omega^2} - \frac{1}{s^2 + s \bar{d} + \omega^2} \|_{\infty},$$

resulting in the following nominal damping parameter:

$$d_{opt} = \frac{D_l \alpha + D_r \beta - \sqrt{\beta (2D_l D_r (\alpha - 2D_l) + \beta (D_l^2 + D_r^2)))}}{2D_l},$$

where:

$$\alpha = (\omega + D_r) \quad \text{and} \quad \beta = (\omega - D_r)$$

Simulations show that $d_{opt}$ coincides very well with the exact optimal solution of (8).

4 Model Updating

As already mentioned above the often inadequate knowledge of physical and structural parameters leads to significant differences between the model and the corresponding real system. One recognizes these errors through a
comparison of the measured natural frequencies and mode shapes with the corresponding analytical data. The aim of model updating now is to use the few known measurements to change the model parameters in such a way that a good correspondence between the model and the real system is achieved. The typical problems which have to be considered in this approach become clear by opposing the analytical and real data

Problem of model updating

\[
\begin{align*}
V & \in \mathbb{R}^{n \times n} , \\
A & \in \mathbb{R}^{n \times n} \\
J_{FE}, K_{FE} & \in \mathbb{R}^{n \times n} \\
\end{align*}
\]

The number of the FE knots \( n \) in general is considerably larger than the number of measurement locations \( r \). This means that the dimensions of the measured and the analytical mode shapes differ. Therefore in a first step of each model updating algorithm the dimensions have to be matched, i.e. the measured mode shapes have to be expanded.

Moreover, in general only \( m \) measured eigenfrequencies and mode shapes from the low frequency domain of the system are available and it is not always clear which measured modal data correspond to which modal FE-data. Therefore appropriate assignment procedures, like the MAC-Criteria (Modal Assurance Criterion), have to be applied.

### 4.1 Mode Shape Expansion

As explained above, the differences between the dimensions of the FE-mode shapes and the corresponding measured eigenmodes require an appropriate expansion of the latter ones. The standard approach for this expansion is the so-called Master-Slave Ansatz [5], [8], where the model itself and the measured eigenmode components are used in order to reconstruct the missing components. The obvious drawback of this approach is the influence of erroneous system components on the expansion process. An alternative approach that circumvents this problem is the mode shape expansion based on averaging splines. This method has been applied to the eigenmode data from the test rig leading to quite reasonable results. As expected, the performance of the spline expansion strongly depends on the distribution of the measurement points along the shaft line. Moreover, if the measurement data is corrupted by noise and if some information about the noise distribution is given, then based on the spline approach much better expansion results could be obtained compared to the Master-Slave Ansatz. For the future we plan to combine the advantages of both expansion methods in a mixed approach.
4.2 Model Updating Based on Taylor Approximations

In the context of model updating the linear Taylor expansion for the eigenfrequencies

\[ \tilde{\lambda} = \lambda + S^T \delta \Theta \]  

(9)

often is used for the iterative computation of the parameter updates \( \delta \Theta = (\Theta_{i+1} - \Theta_i) \). Here \( S = \left( \frac{\partial \lambda_k}{\partial \Theta_j} \right)_{i,j} \) denotes the so-called sensitivity matrix which is explicitly given by the expressions (4). The vector \( \tilde{\lambda} \in \mathbb{R}^r \) contains all measured eigenfrequencies, whereas the vector \( \lambda \in \mathbb{R}^r \) contains the corresponding analytical eigenfrequencies, which usually are assigned by the Modal Assurance Criterion (MAC). Based on the pseudo-inverse \( S^+ \) the update of the uncertain parameters is given by

\[ \Theta_{i+1} = \Theta_i + S^+ (\tilde{\lambda} - \lambda). \]  

(10)

This approach nicely shows the connection between sensitivity analysis and model updating. Note, however, that no informations from the measured mode shapes are used in this updating approach. Modifications of this algorithm incorporating the eigenmodes and additionally introducing frequency weights can be found in [5].

Based on the second order derivatives (6) we additionally studied the performance of updating algorithms related to the quadratic Taylor expansion. Let \( H_k = \frac{\partial^2 \lambda_k}{\partial \Theta_i \partial \Theta_j} \) denote the Hessian matrix with respect to the k-the eigenvalue. Adding the quadratic term one then has to solve the nonlinear optimization problem

\[ \min_{\delta \Theta} \sum_{k=1}^m \left( \tilde{\lambda}_k - \lambda_k \right) + S_k \delta \Theta + 0.5 \delta \Theta^T H_k \delta \Theta \), \]

which we did by applying the Levenberg-Marquardt approach. The quality of the model update strongly depends on the solution of this optimization problem. However, the analytical computation of the quadratic term within an iterative procedure especially in the case of high dimensional systems is very time expensive. Instead of a recalculation of the Hessian matrix after every iteration the BFGS method known from optimization literature [2], [4], [13] was used for the update of the Hessian matrix. This step lead to justifiable CPU times.

Comparing the quadratic with the linear approach one generally recognizes an improvement in the updating behaviour. In some cases the linear algorithm was even diverging while the quadratic approach still was converging.

As already mentioned above mode shapes are not considered in this approach. However, using the measured mode shapes very often leads to an improvement in the updating behaviour since they contain further information
of the real system. Therefore one has to expand the measured mode shapes first. It is obviously clear that the expansion is not exact and thus additionally to measurement errors expansion errors effect the process of model updating [8], [12], [15]. To reduce the influence of expansion errors we developed an alternating expansion-updating algorithm which is based on the theorem of Gerschgorin.

4.3 Model Updating Based on Gerschgorin’s Theorem

The Gerschgorin disk $D_i$ of a matrix pencil $(J, K)$ is given by

$$D_i = \{ \beta : |(K - \beta J)_{ii}| \leq \sum_{i \neq j} |(K - \beta J)_{ij}| \}. \quad (11)$$

Gerschgorin proved that the union of all these disks contains the eigenvalues of the matrix pencil. Moreover, if all disks are pairwise disjoint from each other then each disk contains exactly one eigenvalue [14].

For many FE-Models one easily can show that their Gerschgorin disks are not disjoint and therefore no useful information about the location of the eigenvalues is given. However, if we calculate the Gerschgorin disks for the modal transformed system i.e $(V^T KV, V^T JV)$ we get

$$\tilde{D}_i = \{ \beta : |v_i^T (K - \beta J) v_i| \leq \sum_{i \neq j} |v_i^T (K - \beta J) v_j| \} \quad (12)$$

Since the orthogonality relations $v_i^T K v_j = 0$ and $v_i^T J v_j = 0 \ \forall i, j$ and $i \neq j$ hold, the radius of each Gerschgorin disk of the modal transformed system $\tilde{D}_i$ is equal to zero and therefore every Gerschgorin disk contains only one single point, which obviously has to be an eigenvalue.

This properties can be used in a model updating algorithm since they also should be fulfilled for the updated system denoted by $(K(\Theta), J(\Theta))$ with respect to the measured eigendata.

Since the dimension of the measured mode shapes $\bar{v}$ generally is much smaller than the dimension of the corresponding analytical vectors $v$ as discussed in subsection 4.1 one has to expand the measured mode shapes appropriately. However, the use of an expansion algorithm introduces further errors which effect the model updating process. Let $\hat{v}_{i\text{exact}}$ be an exact but unknown n-dimensional mode shape of the system and $\hat{v}$ the corresponding expanded one. The aim of model updating using the Gerschgorin related conditions from above would be to find appropriate parameter changes such that

$$\hat{v}_{i\text{exact}}^T K(\Theta) \hat{v}_{j\text{exact}} = 0; \quad \hat{v}_{i\text{exact}}^T J(\Theta) \hat{v}_{j\text{exact}} = 0;$$

$$\hat{v}_{i\text{exact}}^T (K(\Theta) - \alpha_i J(\Theta)) \hat{v}_{i\text{exact}} = 0; \quad v(\Theta)^T_i (K(\Theta) - \alpha_i J(\Theta)) v(\Theta)_i = 0;$$
with $\alpha_i = 4\pi^2 \bar{\lambda}_i^2$. Since the exact mode shapes $\hat{v}_{\text{exact}}$ are unknown, however, we have to use the corresponding expanded ones. Doing so, we obtain:

\begin{align}
\hat{v}_i^T K(\Theta) \hat{v}_j &= 0 \\
\hat{v}_i^T J(\Theta) \hat{v}_j &= 0 \\
\hat{v}_i^T (K(\Theta) - \alpha_i J(\Theta)) \hat{v}_i &= 0
\end{align}

(13) \quad (14) \quad (15)

For every mode shape the expansion error can be expressed by a vector $\delta v_i$ with $\hat{v}_i = \hat{v}_{\text{exact}} + \delta v_i$. Now the effect of the expansion error in model updating can easily be shown by inserting $\hat{v}_{\text{exact}} + \delta v_i$ and reformulating equations (13)–(15):

\begin{align}
\hat{v}_{\text{exact}}^T K \hat{v}_{\text{exact}} &= -(\delta v_i^T K \delta v_j + \delta v_i^T K \delta v_j) \\
\hat{v}_{\text{exact}}^T J \hat{v}_{\text{exact}} &= -(\delta v_i^T J \delta v_j + \delta v_i^T J \delta v_j) \\
\hat{v}_{\text{exact}}^T (K - \alpha_i J) \hat{v}_{\text{exact}} &= -(\delta v_i^T (K - \alpha_i J) \delta v_i + \delta v_i^T (K - \alpha_i J) \delta v_i)
\end{align}

(16) \quad (17) \quad (18)

In order to reduce notational complexity we skipped the $\Theta$-dependency of $J$ and $K$. The expressions on the right hand side of these equations usually are not equal to zero and thus the goal of model updating is not achieved perfectly.

To reduce the influence of the expansion errors the conditions (13)–(15) should not be required in this form. Instead the formulas

\begin{align}
|\hat{v}_i^T K_{i+1} \hat{v}_j| &\leq (1 - \epsilon_{i,j}) |\hat{v}_i^T K_i \hat{v}_j| \quad i, j = 1 \ldots m \quad i \neq j \quad \epsilon_{i,j} > 0 \\
|\hat{v}_i^T J_{i+1} \hat{v}_j| &\leq (1 - \gamma_{i,j}) |\hat{v}_i^T J_i \hat{v}_j| \quad i, j = 1 \ldots m \quad i \neq j \quad \gamma_{i,j} > 0 \\
|\hat{v}_i^T (K_{i+1} - \alpha_i J_{i+1}) \hat{v}_i| &\leq (1 - \rho_i) \sum_{j=1}^{m} |\hat{v}_i^T (K_i - \alpha_i J_i) \hat{v}_i| \quad i = 1 \ldots m
\end{align}

(19) \quad (20) \quad (21)

with $K_{i+1} = K(\Theta)$ and $J_{i+1} = J(\Theta)$ should be used in an alternating expansion-updating algorithm. The design parameter $\epsilon_{i,j}$, $\gamma_{i,j}$ and $\rho_i$ are used to reduce the influence of the expansion errors.

**Example 3.** In this example the model updating algorithm based on Gershgorin’s theorem is used to update the structural parameters of the FE-model of the test rig. The first four measured eigenfrequencies and the corresponding mode shapes measured at 24 locations are given. The values of $\epsilon_{i,j}$, $\gamma_{i,j}$ and $\rho_i$ are set to 0.01. Figure 5 shows the eigenfrequencies of the starting FE-model (red), the eigenfrequencies of the updated model after every iteration (blue) and the measured eigenfrequencies (green). The parameter changes are within the prescribed tolerances and the algorithm is able to solve the updating problem quite satisfactorily.
Fig. 5. Iterative model updating based on Gerschgorin’s theorem

5 Outlook

All algorithms developed in this project are integrated in the already existing matlab tool designed in the proceeding BMBF-project. The goal is to obtain a complete tool concerning construction and monitoring of rotating machines. The functionality of this tool include

- modelling the torsional oscillations of rotating machinery
- sensitivity analysis and error estimations
- model updating
- model reduction
- Kalman and $H_\infty$ observer design including optimal sensor placement.

Concerning the mode shape expansion it is planned to study combinations of the Master-Slave Ansatz and the spline approximation approach. Moreover, it is well known that the undamped shaft model belongs to the class of so-called oscillating systems for which some structural properties of the eigensystem are known [6]. We therefore also want to integrate these informations into the expansion process.
In the context of model updating some further investigations will be done concerning the algorithm based on Gerschgorin’s Theorem. The quality of the update is influenced by the choice of the design parameter $\epsilon, \gamma$ and $\rho$ which depend on the size of the expansion errors. Therefore we will try to find appropriate bounds for the expansion errors and thus for the design parameters. The consideration of measurement errors and of damping will be further research topics.

References

3. Drubel, O., Lang, P.: Torsional Stress Analyser for Turbo Generator Shaft Systems with Contactless Torque Measurement, SDEMPED Gijon (Spain), 1999
8. Imregun,M., Ewins,D.J.: An investigation into Mode Shape Expansion Techniques, IMAC, Kismee(Florida) 168-175, 1993
Adaptive Multigrid Methods
for the Vectorial Maxwell Eigenvalue Problem
for Optical Waveguide Design

Peter Deuflhard\textsuperscript{1,2}, Frank Schmidt\textsuperscript{1}, Tilmann Friese, and Lin Zschiedrich\textsuperscript{1}

\textsuperscript{1} Konrad-Zuse-Zentrum für Informationstechnik Berlin, Germany
URL: http://www.zib.de
\textsuperscript{2} Freie Universität Berlin, Fachbereich Mathematik und Informatik, Germany

Abstract. This paper has been motivated by the need for a fast robust adaptive multigrid method to solve the vectorial Maxwell eigenvalue problem arising from the design of optical chips. Our nonlinear multigrid methods are based on a previous method for the scalar Helmholtz equation, which must be modified to cope with the null space of the Maxwell operator due to the divergence condition. We present two different approaches. First, we present a multigrid algorithm based on an edge element discretization of time-harmonic Maxwell's equations, including the divergence condition. Second, an explicit elimination of longitudinal magnetic components leads to a nodal discretization known to avoid discrete \textit{spurious modes} also and a vectorial eigenvalue problem, for which we present a multigrid solver. Numerical examples show that the edge element discretization clearly outperforms the nodal element approach.

AMS Subject Classification: 65N25, 65N30, 65N55

Keywords: Maxwell's equations, eigenvalue problem, edge elements, multigrid methods, waveguide, optical chip design

1 Eigenvalue Problems for Optical Waveguide Design

Integrated optical components like semiconductor lasers, optical switches, and filters are essential parts of modern fiber-optical networks, see [7], [8, ch. 2]. Figure 1 shows a mounted MQW-laser of the latest technological generation.

Each of them consists of various sub-components, which are connected by waveguides. Therefore, the design of optical waveguides is a central task. The analysis of optical waveguides is based on the knowledge of their guided modes and propagation constants. A schematic representation of an optical chip is given in Fig. 2. The optical beam propagates in z-direction. The geometry of the chip itself is regarded as invariable in this direction. Guided modes are modes that exhibit an intensity distribution invariant in z-direction and with finite lateral extension. In former work we had simplified the basic vectorial Maxwell's equations such that a scalar Helmholtz eigenvalue problem arose [7]. As industrial optical components get more and more complex [17],
including sharp and significant jumps in the permittivity of the waveguide materials, this approximation turns out to be too crude. Therefore we have to return to the exact Maxwell’s equations as a mathematical model. This leads again to an eigenvalue problem, which, however, is much more complex and is the topic of this paper.

Starting from Maxwell’s equations in a source and current free medium and assuming time-harmonic dependence of the electromagnetic field with angular frequency \( \omega \) the electric and magnetic fields

\[
\mathbf{E}(x, y, z, t) = \tilde{\mathbf{E}}(x, y, z) \cdot e^{i\omega t}, \quad \mathbf{H}(x, y, z, t) = \tilde{\mathbf{H}}(x, y, z) \cdot e^{i\omega t}
\]
must satisfy the time-harmonic Maxwell equations
\[
\begin{align*}
curl \mathbf{E} & = -i\omega \mu \mathbf{H}, \quad \text{div} \mathbf{E} = 0 \\
curl \mathbf{H} & = i\omega \varepsilon \mathbf{E}, \quad \text{div} \mathbf{H} = 0.
\end{align*}
\]

Herein \( \varepsilon = \varepsilon(x, y) \) denotes the permittivity and \( \mu \) the permeability of the material. For simplicity, we assume \( \mu \) to be constant, \( \varepsilon = \varepsilon_D - i\sigma/\omega \) complex, and drop the wiggles, so that \( \mathbf{E} \rightarrow \mathbf{E}, \mathbf{H} \rightarrow \mathbf{H}. \) From the equations above we then may derive (by direct substitution)
\[
\begin{align*}
curl \varepsilon^{-1} \mathbf{H} - \omega^2 \mu \mathbf{H} & = 0, \\
\text{div} \mu \mathbf{H} & = 0,
\end{align*}
\]
where only the magnetic field is involved. Motivated by the \( z \)-invariance of our geometry, we seek solutions of equation (1), which depend harmonically on \( z \), i.e.
\[
\mathbf{H}(x, y, z) = \hat{\mathbf{H}}(x, y) \cdot e^{-ik_z z}.
\]

Here \( k_z \) is the propagation constant, which is the eigenvalue of interest. In the following, we again drop the hat, so that \( \mathbf{H} \rightarrow \mathbf{H}. \) Let us introduce a reference permittivity \( \varepsilon_0 \), the relative permittivity \( \varepsilon_r = \varepsilon/\varepsilon_0 \) and the corresponding reference wave number \( k_0^2 = \varepsilon_0 \mu \omega^2 \), which is assumed to be given. Upon splitting the magnetic field into a transversal part \( \mathbf{H}_\perp(x, y) \) and a longitudinal part \( \mathbf{H}_z(x, y) \),
\[
\mathbf{H}(x, y) = \mathbf{H}_\perp(x, y) + \mathbf{H}_z(x, y) \cdot e_z,
\]
equation (1) is equivalent to the eigenvalue problem
\[
\begin{align*}
\nabla_\perp \times \varepsilon_r^{-1} \nabla_\perp \times \mathbf{H}_\perp - k_0^2 \mathbf{H}_\perp & = -k_z^2 \varepsilon_r^{-1} \mathbf{H}_\perp + i k_z \varepsilon_r^{-1} \nabla_\perp \mathbf{H}_z \\
- \nabla_\perp \cdot \varepsilon_r^{-1} \nabla_\perp \mathbf{H}_z - k_0^2 \mathbf{H}_z & = i k_z \nabla_\perp \cdot \varepsilon_r^{-1} \mathbf{H}_\perp \\
\nabla_\perp \cdot \mathbf{H}_\perp & = i k_z \mathbf{H}_z.
\end{align*}
\]

In principle, Maxwell eigenvalue problems divide into two classes. In the so-called resonance problem
\[
\begin{align*}
\begin{bmatrix}
\nabla_\perp \times \varepsilon_r^{-1} \nabla_\perp \times + k_z^2 \varepsilon_r^{-1} & -ik_z \varepsilon_r^{-1} \nabla_\perp \\
-ik_z \nabla_\perp \cdot \varepsilon_r^{-1} & -\nabla_\perp \cdot \varepsilon_r^{-1} \nabla_\perp
\end{bmatrix}
\begin{bmatrix}
\mathbf{H}_\perp \\
\mathbf{H}_z
\end{bmatrix}
& = k_0^2 \cdot \begin{bmatrix}
\mathbf{H}_\perp \\
\mathbf{H}_z
\end{bmatrix}
\end{align*}
\]
we ask for an eigenvalue \( k_0 \) or \( \omega \), respectively. The structure of this problem is rather simple – the left hand side of equation (5) is a selfadjoint operator in the case of loss-free media and the right hand side consists of a positive definite mass term. However, this is not the appropriate problem in integrated optics. There the task is to determine the propagation constant \( k_z \), which
appears implicitly in equation (5). By introducing \( u_z = k_z / i \cdot H_z \) this so called waveguide problem also allows an explicit eigenvalue problem formulation,

\[
\begin{bmatrix}
\nabla \times \epsilon_r^{-1} \nabla \times -k_0^2 & \epsilon_r^{-1} \nabla \times -k_0^2 & 0 \\
0 & -\nabla \times \epsilon_r^{-1} \nabla \times -k_0^2 & 0 \\
0 & 0 & -k_z^2 \\
\end{bmatrix}
\begin{bmatrix}
H_z \\
\epsilon_r^{-1} \cdot H_z \\
u_z \\
\end{bmatrix}
= 0
\]

(6)

with a non-selfadjoint operator on the left and a singular operator on the right hand side. In view of a numerical approximation we have to choose a finite domain of discretization and hence boundary conditions must be prescribed. This is a rather complex issue in the case of optical waveguides, because in many problems boundary conditions are not explicitly given on finite domains. If only guided modes are sought, the magnetic field decays exponentially fast to zero outside a finite domain, so we may prescribe zero boundary conditions on a sufficiently large domain.

2 Variational Formulations and Discretizations

We present two different variational formulations of the system (2)–(4). These two approaches differ by the incorporation of the divergence condition (4) and the treatment of the \( H_z \)-component. In the first approach we discretize the transversal components \( H_z \) with linear edge elements, \( H_z \) with nodal elements and set up a discrete analog of Maxwell’s equations. In the second approach we eliminate the \( H_z \)-component by the divergence condition and choose a nodal element discretization for \( H_z \). For a discussion of these approaches see [10]. In order to avoid confusions with the standard notation for Sobolev spaces we write \( \mathbf{u}_\perp \) instead of \( H_z \) and introduce \( u_z = 1 / i \cdot H_z \).

Recall that \( \epsilon \) is in general not smooth. So the computational domain may be split up according to

\[ \Omega = \Omega_1 \cup \ldots \cup \Omega_N, \]

where \( \epsilon \) is now smooth on each subdomain \( \Omega_i \). The boundary between \( \Omega_i \) and \( \Omega_j \) is denoted by \( \Gamma_{ij} \). Furthermore we introduce the inner products by

\[ (v_\perp, u_\perp) = \int_{\Omega} v_\perp \cdot u_\perp \, d\Omega \]

and also \( (v_z, u_z) = \int_{\Omega} v_z u_z \, d\Omega \).

**Edge element discretization.** In this approach we set up a direct discrete analog to problem (5) and additional to the divergence condition (4). The weak form of the waveguide eigenvalue problem is to find \( k_z \), \( u_\perp \), \( u_z \), such that for any \( v_\perp \in H_0(\text{curl}, \Omega) \) and \( v_z \in H_0^1(\Omega) \)

\[
\begin{align}
(\nabla \times v_\perp, \epsilon_r^{-1} \nabla \times u_\perp) + k_z \cdot (v_\perp, \epsilon_r^{-1} \nabla \perp u_z) + k_z^2 \cdot (v_\perp, \epsilon_r^{-1} u_\perp) &= +k_0^2 \cdot (v_\perp, u_\perp) \\
(\nabla \perp v_z, \epsilon_r^{-1} \nabla \perp u_z) + k_z \cdot (\nabla \perp v_z, \epsilon_r^{-1} u_\perp) &= k_0^2 \cdot (v_z, u_z) \\
(\nabla \perp v_z, u_\perp) &= k_z \cdot (v_z, u_z).
\end{align}
\]

(7) (8) (9)
In order to derive these equations, we use the continuity of $E_z = \varepsilon_r^{-1} \nabla \times \mathbf{u}_\perp$, so that all line integrals over $\Gamma_{ij}$ and $\Gamma_{ji}$ cancel in the interior of $\Omega$.

It is an important fact that a solution $(\mathbf{u}_\perp, u_z) \in H_0(\text{curl}, \Omega) \times H_0^1(\Omega)$ of the equations (7), (8) with finite $k_z$ and $k_\sigma$ also satisfies the divergence condition (9). This can be seen by the special choice $\mathbf{v}_\perp = -1/k_z \cdot \nabla \times \mathbf{v}_z \in H_0(\text{curl}, \Omega)$ and inserting equation (8) into (7).

Given a regular triangulation of $\Omega$, the structure of (7)-(9) can be passed on to a discrete version by using edge elements ([16],[4], [1]) for the transversal components and nodal elements for $u_z$. Let $V_\perp \subset H_0(\text{curl}, \Omega)$, $V_z \subset H_0^1(\Omega)$ be the corresponding linear finite element spaces with bases $\psi_1 \cdots \psi_m$ and $\phi_1 \cdots \phi_p$. Here $m$ is the number of interior edges and $p$ is the number of inner points of the triangulation. We introduce the system matrices

$$
(A_\perp)_{jk} = (\nabla \times \psi_j, \varepsilon_r^{-1} \nabla \times \psi_k),
(A_z)_{jk} = (\nabla \phi_j, \varepsilon_r^{-1} \nabla \phi_k),
(B_\perp)_{jk} = (\psi_j, \varepsilon_r^{-1} \psi_k),
(M_z)_{jk} = (\phi_j, \phi_k)
$$

(10)

The weak gradient of $\phi_j$ is an element of $V_\perp$, so we may define $G$ to be the matrix representation (for the above bases) of the linear map

$$
\nabla_\perp : V_z \rightarrow V_\perp.
$$

In this way we arrive at the discretized version of (7)-(8)

$$
\begin{bmatrix}
A_\perp + k_z^2 \cdot B_\perp & k_z \cdot B_\perp G \\
k_z \cdot G^* B_\perp & A_z
\end{bmatrix}
\begin{bmatrix}
\mathbf{u}_\perp \\
\mathbf{u}_z
\end{bmatrix} = k_0^2 \cdot
\begin{bmatrix}
M_\perp & 0 \\
0 & M_z
\end{bmatrix}
\begin{bmatrix}
\mathbf{u}_\perp \\
\mathbf{u}_z
\end{bmatrix},
$$

(11)

and the discrete divergence condition

$$
\begin{bmatrix}
G \\
-I
\end{bmatrix}^* 
\begin{bmatrix}
M_\perp & 0 \\
0 & M_z
\end{bmatrix} 
\begin{bmatrix}
\mathbf{u}_\perp \\
\mathbf{u}_z
\end{bmatrix} = 0.
$$

(12)

As in the continuous system, the divergence condition (12) automatically holds for a solution $(\mathbf{u}_\perp, u_z)$ of (13) with finite $k_z$ and $k_0$. Problem (11) is a standard eigenvalue problem $A_{\text{res}} u = \lambda M_{\text{res}} u$ for $\lambda = k_0^2$ with a selfadjoint matrix $A_{\text{res}}$ and a positive definite mass matrix $M_{\text{res}}$. Here the unknown propagation constant $k_z$ appears implicitly. As above for the continuous problem, we may rearrange equation (11) by substituting $\mathbf{u}_z = k_z \cdot \mathbf{u}_z$ we arrive at an explicit eigenvalue problem

$$
\begin{bmatrix}
A_\perp & B_\perp G \\
0 & A_z
\end{bmatrix}
\begin{bmatrix}
\mathbf{u}_\perp \\
\mathbf{u}_z
\end{bmatrix} = -k_0^2 \cdot
\begin{bmatrix}
M_\perp & 0 \\
0 & M_z
\end{bmatrix}
\begin{bmatrix}
\mathbf{u}_\perp \\
\mathbf{u}_z
\end{bmatrix},
$$

(13)

for $k_z$. Since the above matrix $B$ is singular, this formulation is not well suited for the construction of a multigrid method. That is why, in the following, we will focus on a multigrid algorithm for equation (11), which we will solve for $k_z$, subject to the divergence condition (12).
Nodal element discretization. We use the divergence condition (4) to substitute \( H_z \) in (2), which gives the modified transversal equation [10]

\[
\nabla \times \epsilon_r^{-1} \nabla \times \mathbf{u}_\perp - \epsilon_r^{-1} \nabla \cdot (\nabla \perp \cdot \mathbf{u}_\perp) - k_0^2 \mathbf{u}_\perp = -k_z^2 \epsilon_r^{-1} \mathbf{u}_\perp,
\]

where only the transverse field \( \mathbf{u}_\perp \) is involved. The corresponding weak problem now reads

\[
(\nabla \times \mathbf{v}_\perp, \epsilon_r^{-1} \nabla \times \mathbf{u}_\perp) - k_0^2 (\mathbf{v}_\perp, \mathbf{u}_\perp) +
\sum_{i=1}^N \int_{\Omega_i} (\nabla \cdot \epsilon_r^{-1} \mathbf{v}_\perp) (\nabla \cdot \mathbf{u}_\perp) d\Omega_i +
- \sum_{i,j=1}^N \int_{\Gamma_{ij}} (\mathbf{n} \cdot \mathbf{v}_\perp) \epsilon_r^{-1} (\nabla \cdot \mathbf{u}_\perp) d\Gamma_{ij} =
- k_z^2 (\mathbf{v}_\perp, \epsilon_r^{-1} \cdot \mathbf{u}_\perp)
\]

for all \( \mathbf{v}_\perp \in H_0^1(\Omega) \times H_0^1(\Omega) \).

As above all line integrals involving \( E_z = \epsilon_r^{-1} \nabla \times \mathbf{u}_\perp \) vanish in the interior of \( \Omega \). This is not the case for the second sum in (15), because \( \nabla \cdot \mathbf{u}_\perp = ik_z H_z \) is continuous and hence \( \epsilon_r^{-1} \nabla \cdot \mathbf{u}_\perp \) may jump across \( \Gamma_{ij} \). In (15) not only the curl operator but also the div operator act on \( \mathbf{u}_\perp \), which inhibits the use of linear edge elements for the transversal field. A finite element discretization based on the linear nodal elements space \( V_\perp \subset H_0^1(\Omega) \times H_0^1(\Omega) \) is yielding straightforward the algebraic system

\[
A \mathbf{u}_\perp = -k_z^2 \cdot \mathbf{M} \mathbf{u}_\perp,
\]

with a non-symmetric matrix \( A \) and a canonical mass matrix \( \mathbf{M} \).

3 Multigrid Algorithms

In an adaptive finite element discretization of the above problems we have a set of sequentially refined triangulations \( \{\mathcal{T}_h\} \) of \( \Omega \) with corresponding finite element spaces \( V_h \subset H_0(\text{curl}, \Omega) \times H_0^1(\Omega) \) resp. \( V_h \subset H_0^1(\Omega) \times H_0^1(\Omega) \). In each case this yields an algebraic eigenvalue problem \( A_h \mathbf{u} = \lambda \cdot \mathbf{B}_h \mathbf{u} \). As in [12][13] we generalize this problem for the ability to calculate simultaneously a certain number \( q \) of clustered or degenerate eigenvalues with smallest real part. Hence we seek a \( q \) - dimensional invariant subspace \( U_h \), in particular

\[
A_h U_h = B_h U_h T_h.
\]

As mentioned above the structure of equation (17) depends on the chosen discretization. The edge element discretization for the waveguide problem (13) leads to a singular matrix \( B_h \), whereas a nodal basis discretization of the modified transverseal equation (14) and the resonance problem (11) give a canonical mass matrix \( B_h \). There exist different multigrid solvers for the above
problem with a positive definite $B$ \cite{14, 5, 6}. Here we present the method developed in \cite{12} for the Helmholtz eigenvalue problem. The formulation of the waveguide problem based on the modified transversal equation fits perfectly into this multigrid concept. Therefore we can extend this algorithm directly to the vectorial case. Unfortunately, this is not true for magnetic or lossy materials. In these cases the mass matrix $B_h$ is no longer positive definite.

### 3.1 General Concept

In the following we suppress the subindex $h$ and assume $B$ to be a positive definite operator. The backbone of our method is a pcg-like iterative eigenvalue solver for problem (17), see \cite{11}. The main advantage of this method is that it allows the handling of subspaces whose $B$-orthonormality will not be destroyed by the algorithm. As in \cite{12} we use this method as a smoother. This method reduces the high-frequency error on each grid very effectively and gets inefficient after a few iterations. How does the smoother works?

Recall that the above eigenvalue problem admits a Schur decomposition

$$AQ = BQK, \quad Q'BQ = \text{Id}.$$ 

Herein $K = \text{diag}(K_i)$ is a block diagonal matrix, with upper triangular blocks $K_i$. To each $K_i$ corresponds an invariant subspace $E_i$. The subspaces $E_i$ are chosen so that they possess no non-trivial invariant subspace. The matrices $K_i$ depend on the chosen $B$-orthonormal basis, while trace $K_i$ does not depend on it. Any $q$-dimensional invariant subspace $Y$ is the sum of particular $E_i$, say

$$Y = E_{i_1} + \cdots + E_{i_\ell}.$$ 

To $Y$ corresponds the upper triangular matrix $Z = \text{diag}_{i=1\ldots\ell} (K_i)$. Hence we may define $\text{trace } Y = \text{trace } Z$. We characterize the sought $q$-dimensional invariant subspace $U$ with corresponding upper triangular matrix $T$ by

$$\text{trace } U = \min \{ \text{trace } Y \mid Y \text{ is } q\text{-dimensional invariant subspace} \}.$$ 

Details of this algorithm are given in Algorithm 1.

Given an initial guess $U^{(0)}$, $T^{(0)}$ for the sought $q$-dimensional invariant subspace on the finest grid we construct a correction space $P$ in a pcg-like manner. We use for example a Jacobi iteration step as the preconditioning matrix $C^{-1}$. Now, we correct $U^{(0)}$ by solving a small projected eigenvalue problem (Ritz step). Assuming that the $q$ smallest eigenvalues are sufficiently well approximated such that there is a spectral gap between the $q$ first eigenvalues and the remaining ones of the projected system, the Schur decomposition in Algorithm 1 supplies upper triangular matrices $T^{(k)} \in \mathbb{C}^{q \times q}$, $T_S \in \mathbb{C}^{q \times q}$ with

$$\text{Re}(T_{11}^{(k)}) \leq \cdots \leq \text{Re}(T_{qq}^{(k)}) \leq \text{Re}(T_{S,11}) \leq \cdots \leq \text{Re}(T_{S,qq}).$$

Algorithm 1 Döhler pcg as smoother with $\nu$ iteration steps

Require: $U^{(0)}$, $T^{(0)}$ {initial guess}

\[ G = C^{-1} \left( A U^{(0)} - B U^{(0)} T^{(0)} \right) \]

\[ P = G \{ \text{initial correction space} \} \]

\textbf{for} $k = 0 \text{ to } \nu \textbf{ do} \]

\[ \tilde{A} = \left[ U^{(k)} P \right] A \left[ U^{(k)} P \right] \{ \text{projected problem} \} \]

\[ \tilde{B} = \left[ U^{(k)} P \right] B \left[ U^{(k)} P \right] \]

\[ \left[ \tilde{U} \tilde{S} \right] ' \tilde{A} \left[ \tilde{U} \tilde{S} \right] = \begin{pmatrix} T^{(k)} & 0 \\ 0 & T_S \end{pmatrix} \{ \text{Schur decomposition} \} \]

\[ \left[ \tilde{U} \tilde{S} \right] ' \tilde{B} \left[ \tilde{U} \tilde{S} \right] = \text{Id} \]

\[ U^{(k+1)} = \left[ U^{(k)} P \right] \tilde{U} \{ \text{update of } U \} \]

\[ S = \left[ U^{(k)} P \right] \tilde{S} \]

\[ G = C^{-1} \left( A U^{(k)} - B U^{(k)} T^{(k)} \right) \]

\textbf{solve for} \( X : T_S X - X T^{(k)} = -P' \left( A G - B G T^{(k)} \right) \) \{Sylvester equation\}

\[ P = G + SX \{ \text{new correction space} \} \]
\textbf{end for}

This correction procedure is motivated by the minimal principle above. After the correction of $U$ we have to construct a new correction space $P$ which is done similar to the pcg-method again.

The proposed correction space $P$ in Algorithm 1 is generated by a multiplication of the current $U$ with the discrete second order "differential operator" $(A [.] - B [.] T)$ and hence high-frequency errors are overstressed in $P$. But fortunately, by the multigrid structure we can force low-frequency corrections to appear in $P$. So, given the prolongation matrix $I^h_H$ from any coarse grid to the current fine grid, we restrict the eigenvalue problem to the subspace $[U I^h_H]$, especially

\[ \begin{pmatrix} U I^h_H \\ U I^h_H \end{pmatrix} ' \tilde{A} \begin{pmatrix} U I^h_H \\ U I^h_H \end{pmatrix} \tilde{U} = \begin{pmatrix} U I^h_H \\ U I^h_H \end{pmatrix} ' \tilde{B} \begin{pmatrix} U I^h_H \\ U I^h_H \end{pmatrix} \tilde{U} T \]

(18)

and carry out Algorithm 1 for this restricted problem. Alternatively, this procedure may be interpreted in the sense that we just use the coarse grid basis $I^h_H$ to construct low-frequency correction spaces $P$ (see Algorithm 1).

On the coarsest level we may use an exact solver or the iterative method (1) with a fixed number of iterations as well.
3.2 Edge Element Method for Waveguide Problem

Algorithm 1 is based on the Schur decomposition of a small restricted eigenvalue problem which can only be done if \( B_h \) is a positive definite matrix. But this is not the case in our variational formulation (13). Therefore we go one step back to equation (5) and its discrete analog (11) in which the desired value \( k_z \) appears implicitly within the selfadjoint eigenvalue problem \( A_{\text{res}}(k_z)u = k_0^2M_{\text{res}}u \). The idea now is to solve this equation for \( k_z \) by a Newton-like iteration. To keep the notation simple, we outline this algorithm for the case of a single non-degenerate eigenvalue \( k_z \) in the self-adjoint case \((q = 1)\). Assume that we can solve \( A_{\text{res}}(k_z)u = (k_0^2 + \delta) \cdot M_{\text{res}}u \) in a neighborhood of the exact value for \( k_z \) and that the “disturbed” normalized eigenvector \( u = u(k_z) \) depends smoothly on \( k_z \). The disturbed resonance wave number is given by the Rayleigh quotient

\[
(k_0^2 + \delta) = \frac{u^T A_{\text{res}}(k_z)u}{u^T M_{\text{res}}u}
\]

and \( k_z \) is determined via the condition \( \delta(k_z) = 0 \). On the basis of

\[
\delta'(k_z) = u^T(k_z)A'_{\text{res}}(k_z)u(k_z) + \underbrace{u^T(k_z)A_{\text{res}}(k_z)u'(k_z)}_{O(\delta)},
\]

we may construct a Newton-like iteration dropping the \( O(\delta) \)-thus arriving at the iteration

\[
k_z^{(i+1)} = k_z^{(i)} - \frac{\delta(k_z^{(i)})}{u^T(k_z^{(i)})A'_{\text{res}}(k_z^{(i)})u(k_z^{(i)})},
\]

The convergence properties of such an iteration are roughly the same as for a simplified Newton iteration [9, chapter 2]. Numerical tests also show, that this Newton-like iteration converges very fast, if we use the \( k_z \) obtained from the coarser grid as the initial guess. In order to get an algorithm of multigrid complexity we need a multigrid solver for the resonance problem \( A_{\text{res}}(k_z)u = (k_0^2 + \delta) \cdot M_{\text{res}}u \).

3.3 Edge Element Method for Resonance Problem

Even the selfadjoint resonance problem (11) fits well into our multigrid concept, some difficulties may arise due to the null space of the operator \( A_{\text{res}} \) and to the fact that we are only interested in positive eigenvalues close to \( k_0^2 \). Therefore a method which minimizes the Rayleigh quotient will converge to this null space. As can be seen in equation (5), the null space consists in the continuous case of 3D-curl-free vector fields

\[
\begin{bmatrix}
\nabla_\perp \varphi \\
-ik_z \varphi
\end{bmatrix}.
\]
This null space is closely tied to the divergence condition (4), see [3]. In fact, by solving the Poisson problem

\[-\Delta \varphi + k_z^2 \varphi = (ik_z H_z - \nabla \cdot H_\perp),\]

one can split the magnetic field into 3D-div-free and curl-free parts (Helmholtz decomposition)

\[
\begin{bmatrix}
H_\perp \\
H_z
\end{bmatrix} = \begin{bmatrix}
\nabla \cdot \varphi \\
-ik_z \varphi
\end{bmatrix}_{\text{curl-free}} + \begin{bmatrix}
H_\perp \\
H_z
\end{bmatrix}_{\text{div-free}}.
\]

The curl-free part is non-physical and violates the divergence condition. Following [15], [2, p. 122] we remove that part throughout the multigrid algorithm, whenever it arises (projection to the div-free subspace).

4 Numerical Examples

The above two algorithms have been implemented in our fully adaptive software package ModeLab. In order to compare the nodal with the edge element discretization on the same hierarchy of grids we restrict ourselves to uniform mesh refinements for the following problems. In each problem choose \(k_0 = 2\pi/1.55\) and the smallest eigenvalue is computed \((q = 1)\). Recall, that the refractive of an material is defined via \(n^2 = \epsilon_r\). In case of the edge element discretization we solve for \(k_z\) by a Newton iteration. In each Newton step we solve a resonance problem by our multigrid algorithm. In the Tables 1–3 we give the required cycles per Newton step to reduce the error to a relative residual of \(10^{-5}\).

Rectangular core Waveguide. The geometry consists of a rectangular core of relative size 1:2 and \(n = 0.7\) embedded in a medium with \(n_{\text{medium}} = 0.5\). As can be seen in Table 1, the approximation of the eigenvalue \(k_z\) is much better in the edge element discretization. In this example the Maxwell solution differs significantly from the Helmholtz approximation, which provides

<table>
<thead>
<tr>
<th>Level</th>
<th>edge cycles</th>
<th>(k_z)</th>
<th>nodal cycles</th>
<th>(k_z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>2.5558</td>
<td>–</td>
<td>2.5362</td>
</tr>
<tr>
<td>2</td>
<td>3+1</td>
<td>2.5555</td>
<td>3</td>
<td>2.5474</td>
</tr>
<tr>
<td>3</td>
<td>3+1</td>
<td>2.5554</td>
<td>3</td>
<td>2.5518</td>
</tr>
<tr>
<td>4</td>
<td>2+0</td>
<td>2.5554</td>
<td>3</td>
<td>2.5537</td>
</tr>
</tbody>
</table>

Table 1. Rectangular core waveguide. Required cycles of our multigrid-method and approximated eigenvalue on each level and Newton iteration. The relative residual error is reduced to \(10^{-5}\)
Fig. 3. Rectangular core waveguide. Isolines of the $H_z$-components computed with the edge element discretization. The rectangular core is plotted in grey

$k_z = 2.5724$. In Fig. 3 the $H_z$-components of the two orthogonal eigenfunctions of smallest eigenvalue are plotted. In the Helmholtz approximation these components are assumed to be zero.

Rib waveguide. The geometry is sketched in Fig. 2. Outside the waveguide we have a medium of $n = 1$. The relative permittivity of the horizontal stripe with a width of 0.2 is $n = 3.38$. This strip is embedded in a material of permittivity $n = 3.17$ at distance 0.2 to the medium. The rectangular rib has a size of $2.4 \times 1$. Again, the edge element discretization approximates the eigenvalues far better than the nodal one (Table 2). Furthermore the number of cycles diminishes in each step for the edge element discretization. The magnetic field strength is plotted in Fig. 4. You can see a singularity-like $H_z$-distribution at the corner of the ribs.

<table>
<thead>
<tr>
<th>Level</th>
<th>edge cycles $k_z$</th>
<th>nodal cycles $k_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>– 12.9685</td>
<td>– 12.9572</td>
</tr>
<tr>
<td>2</td>
<td>19+2 12.9672</td>
<td>6 12.9624</td>
</tr>
<tr>
<td>3</td>
<td>9+0 12.9669</td>
<td>12 12.9647</td>
</tr>
<tr>
<td>4</td>
<td>6+0 12.9668</td>
<td>10 12.9658</td>
</tr>
</tbody>
</table>

Table 2. Rib waveguide (compare Table 1)

Circular optical fiber. The optical fiber consists of a circular core with diameter 2 and permittivity $n = 1.55$ embedded in a medium of permittivity $n_{\text{medium}} = 1.5$. Due to rotational symmetry the lowest eigenvalue is twice
degenerated. As can be seen in Table 3 both variants converge after a small number of cycles per level. As in the above two examples the edge element discretization approximates the eigenvalue better. In Fig. 5 we plot the $H_z$-components of two orthogonal eigenfunctions.

![Fig. 4. Rib waveguide. Isolines of $H^2_z$ (left) and $H^2_\perp$ (right)](image)

<table>
<thead>
<tr>
<th>Level</th>
<th>edge cycles</th>
<th>$k_z$</th>
<th>nodal cycles</th>
<th>$k_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>6.1308</td>
<td>--</td>
<td>6.1295</td>
</tr>
<tr>
<td>2</td>
<td>3+1</td>
<td>6.1317</td>
<td>2</td>
<td>6.1307</td>
</tr>
<tr>
<td>3</td>
<td>3+0</td>
<td>6.1313</td>
<td>2</td>
<td>6.1310</td>
</tr>
<tr>
<td>4</td>
<td>2+0</td>
<td>6.1313</td>
<td>2</td>
<td>6.1311</td>
</tr>
</tbody>
</table>

**Table 3.** Circular optical fiber (compare Table 1)

5 Conclusions

The multigrid concept developed earlier for the scalar Helmholtz equation has been extended herein to vectorial time-harmonic Maxwell’s equations in non-magnetic materials. The algorithm depends on the chosen finite element discretization of the magnetic field. Using the divergence condition
Fig. 5. Circular optical fiber. Isolines of $H_z^2$ of two orthogonal eigenfunctions

one can eliminate the $H_z$-component, which leads to a modified transversal Maxwell equation, which may be discretized by nodal elements. Alternatively, we have directly discretized Maxwell's equations by linear edge elements for the transversal components and nodal elements for the $H_z$-component thus setting up a discrete analog of the continuous Maxwell equations. For both variants a multigrid algorithm has been presented. It is shown experimentally that the edge element discretization approximates the eigenvalues clearly better already on rather coarse grids. All our codes are collected in the software package ModeLab, which also includes adaptive mesh refinements.

References

Direct and Inverse Problems for Diffractive Structures – Optimization of Binary Gratings

Johannes Elschner, Rainer Hinder, and Gunther Schmidt

Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, 10117 Berlin, Germany, E-mail: elchner@wias-berlin.de,
URL: http://www.wias-berlin.de

Abstract. The goal of the project is to provide flexible analytical and numerical tools for the optimal design of binary and multilevel gratings occurring in many applications in micro-optics. The direct modelling of these diffractive elements has to rely on rigorous grating theory, which is based on Maxwell's equations. We developed efficient and accurate direct solvers using a variational approach together with a generalized finite element method which appears to be well adapted to rather general diffractive structures as well as complex materials. The optimal design problem is solved by minimization algorithms based on gradient descent and the exact calculation of gradients with respect to the geometry parameters of the grating.

1 Introduction

Diffractive optics is a modern technology in which optical devices are micromachined with complicated structural features on the order of the length of light waves. Exploiting diffraction effects, these devices can perform functions unattainable with conventional optics and have great advantages in terms of size and weight. The current applications in micro-optics are far-reaching, including high-power laser beam shaping and splitting, solar cell design, image processing and optical document security. Therefore the optimal design of microoptical devices has received considerable attention in the engineering community and has stimulated several mathematical investigations.

One of the most common geometrical configurations for diffractive optical structures is a periodic pattern embedded into a thin-film layer system, such as the multilevel diffraction grating shown in Fig. 1. The pattern is usually created using tools from semiconductor industry. In most applications the grating is illuminated by an incoming time-harmonic plane electromagnetic wave whose length is comparable to the period of the grating. In this situation geometrical optics approximations to the underlying electromagnetic field equations are not accurate, hence, the mathematical modelling has to rely on Maxwell’s equations or related partial differential equations.

The electromagnetic theory of gratings has been studied extensively since Rayleigh’s time. In particular, during the last decade significant progress has been made concerning the direct diffraction problem, i.e. the calculation of the reflection and transmission coefficients of the propagating wave components
of the diffracted field. Several approaches and numerical methods have been proposed for obtaining rigorous solutions to the problem, including modal expansion, differential and integral methods, analytical continuation, and variational methods. The latter approach turned out to be sufficiently flexible to overcome the difficulties associated with non-smooth grating profiles and the highly oscillatory nature of waves and interfaces. The variational method also leads to effective formulas for the gradient of cost functionals arising in optimal design problems, so that gradient-based minimization algorithms can be used to find gratings with specified optical functions.

2 Mathematical Model

Consider a periodic diffraction grating formed by a periodic pattern of non-magnetic materials (of permeability \( \mu \)) with different dielectric constants \( \epsilon \); see Fig. 1. If the coordinate system is chosen such that the grating structure is periodic in \( x_1 \)-direction and invariant in \( x_3 \)-direction, then the diffraction problem is determined by the function \( \epsilon(x_1, x_2) \) which is say \( d \)-periodic in \( x_1 \). This function is assumed to be piecewise constant and complex valued with \( 0 \leq \arg \epsilon < \pi \). The material above and below the grating is assumed to be homogeneous with dielectric constants \( \epsilon = \epsilon^+ > 0 \) and \( \epsilon^- \), respectively. The grating is illuminated by an incoming plane electromagnetic wave

\[
E^i = p e^{i \alpha x_1 - i \beta x_2 + i \gamma x_3} e^{-i \omega t}, \quad H^i = q e^{i \alpha x_1 - i \beta x_2 + i \gamma x_3} e^{-i \omega t}
\]

from the top with the angles of incidence \( \theta, \phi \in (-\pi/2, \pi/2) \). In our applications the wavelength \( \lambda = 2 \pi c/\omega \), \( c \) denoting the speed of light, is comparable to the period \( d \). For notational convenience we will change the length scale by a factor of \( 2 \pi / d \) so that the grating becomes \( 2 \pi \)-periodic: \( \epsilon(x_1 + 2 \pi, x_2) = \epsilon(x_1, x_2) \). Note that this is equivalent to multiplying the frequency \( \omega \) by \( d/2 \pi \).
Then the wave vector of the incident field is expressed in terms of the angles of incidence as
\[ k = (\alpha, -\beta, \gamma) = k^+ (\sin \theta \cos \phi, - \cos \theta \cos \phi, \sin \phi) \quad \text{with } k^+ = \omega(\mu \varepsilon^+)^{1/2}. \]

Note that \((E^i, H^i)\) satisfy the time-harmonic Maxwell equations if the constant amplitude vectors \(p, q\) fulfil the relations \(p \cdot k = 0\) and \(q = (\omega \mu)^{-1} k \times p\). Thus the incoming field is determined by two of their components, for example, \(p_3\) and \(q_3\). The total fields then also satisfy Maxwell’s equations, together with transmission conditions for their tangential components at the interfaces and a radiation condition at infinity.

In the following we mainly restrict ourselves to the case \(\gamma = 0\), i.e. the so-called classical diffraction problem, where \(\phi = 0\) so that the wave vector \(k\) lies in the \(x_1-x_2\) plane. In that case the resulting electromagnetic field can be split into the cases of TE and TM polarization, where either the electric field or the magnetic field is parallel to the \(x_3\)-axis. In both cases Maxwell’s equations can be reduced to transmission problems for a scalar Helmholtz equation
\[ \triangle v + k^2 v = 0 \]
in \(\mathbb{R}^2\), where \(k = \sqrt{\omega^2 \varepsilon \mu}\) is the refractive index and the function \(v\) stands for the \(x_3\)-component of the total electric or magnetic field, and is \(\alpha\) quasi-periodic in \(x_1\): \(v(x_1 + 2\pi, x_2) = \exp(2\pi i \alpha) v(x_1, x_2)\). For TE polarization the solution and its normal derivative \(\partial_n v\) have to cross the interfaces continuously, whereas in TM polarization \(\varepsilon^{-1} \partial_n v\) has to be continuous. Moreover, the diffracted field can be expanded as an infinite sum of plane waves,
\[ v = \sum_{n \in \mathbb{Z}} A_n^\pm \exp(i(n + \alpha)x_1 + i\beta_n^\pm x_2), \quad x_2 \to \pm \infty, \]
with the unknown Rayleigh amplitudes \(A_n^\pm\). Here we have used the notation
\[ \beta_n^\pm = \sqrt{(k^\pm)^2 - (n + \alpha)^2}, \quad n \in \mathbb{Z}, \]
where \(k^-\) denotes the refractive index of the homogeneous medium below the grating structure. Since \(\beta_n^\pm\) is real for at most a finite number of indices \(n\), we see that only a finite number of plane waves in the sum propagate into the far field, with the remaining evanescent modes decaying exponentially as \(x_2 \to \pm \infty\). The number of propagating modes and the direction of propagation for each mode is determined by the frequency of the incident wave, the refractive index of the material, and the period of the structure. The Rayleigh coefficients \(A_n^+\) (resp. \(A_n^-\)) corresponding to these propagating modes are called the reflection (resp. transmission) coefficients. From an engineering point of view, these coefficients are the key feature of any grating since they indicate the energy and phase shift of the propagating modes. In particular, the ratio of the energy of a given propagating mode to the energy
of the incoming wave is called the efficiency of the mode. The reflected and transmitted efficiencies in the TE case are given by

\[ e_n^{TE,\pm} = (\beta_n^\pm / \beta)|A_n^\pm|^2 \]

and in the TM case by

\[ e_n^{TM,+} = (\beta_n^+/\beta)|A_n^+|^2, \quad e_n^{TM,-} = (k^+/k^-)^2(\beta_n^-/\beta)|A_n^-|^2. \]

The exact computation of these quantities is the main goal of the direct diffraction problem. More details can be found in [8], [3], whereas the general case of the so-called conical diffraction problem \((\gamma \neq 0)\) has been studied in [7]. In that case the invariance of the diffractive structure in the \(x_3\)-direction allows us to reduce Maxwell’s equations to a system of two-dimensional Helmholtz equations, which are coupled via transmission conditions at the interfaces.

3 Optimal Design of Binary Gratings

A major part of the motivating applications in diffractive optics is associated with the inverse problems of optimal interface shape design or profile reconstruction from scattered fields. There have been a number of papers from the engineering community that are concerned with the optimal design of periodic gratings; see [10]. By far the greatest activity has been in optimization for ray-tracing and phase-reconstruction techniques which are valid within the domain of Fourier optics. A few of these papers are devoted to optimization problems using rigorous diffraction theory. However, the optimization procedures are usually only based on the values of certain cost functionals, i.e., they require the solution of a large number of direct problems and are therefore computationally expensive. Sometimes the approximation of gradients by simple difference quotients is used, which is, however, very inefficient for a large number of parameters. More advanced methods to find optimal solutions utilize, besides the values of cost functionals, also its gradients or even properties of higher order differentials. The simplest example are descent-type algorithms, which are computationally efficient if explicit gradient formulas are available.

Let us consider the model problem of designing a binary grating on top of a multilayer stack in such a way that the propagating modes have a specified intensity or phase pattern for a chosen range of wavelengths or incidence angles. Assume that the period of the grating and the number of transition points and of thin-film layers are fixed (cf. Fig. 2). Typical minimization problems involving the diffraction efficiencies or Rayleigh coefficients are the following.
To realize prescribed values \( c_{n}^{TE,\pm}, c_{n}^{TM,\pm} \) of certain reflection and transmission efficiencies, the functional

\[
\sum \left( |\epsilon_{n}^{TE,\pm} - \epsilon_{n}^{TE,\pm}|^2 + |\epsilon_{n}^{TM,\pm} - \epsilon_{n}^{TM,\pm}|^2 \right) \\
+ \sum \left( |\epsilon_{n}^{TE,\pm} - \epsilon_{n}^{TE,\pm}|^2 + |\epsilon_{n}^{TM,\pm} - \epsilon_{n}^{TM,\pm}|^2 \right) \to \min
\]

(1)
can be used.

The optimal design of a grating providing a given phase shift \( \varphi \) between the \( n \)th reflected TE and TM mode can be performed using the functional

\[-\epsilon_{n}^{TE,\pm} - \epsilon_{n}^{TM,\pm} + |A_{n}^{TE,\pm} - \exp(i\varphi)A_{n}^{TM,\pm}|^2 \to \min \]

(2)

Note that the efficiencies are functions of the transition points, the thicknesses of layers and the height of the grating, so that the minimum has to be taken over some compact set in the (finite dimensional) parameter space reflecting, e.g., technological constraints on the design of the grating and the thin-film layers. Obviously many other functionals are possible and have been investigated, especially if a corresponding optimization over a range of wavelengths or incidence angles is required.

To find local minima of these cost functionals via gradient descent methods, we must calculate the gradient of Rayleigh coefficients. Explicit gradient formulas based on the solution of the direct problem and its adjoint will be outlined in the next section.

4 Analysis

The direct diffraction problems admit variational formulations in a bounded periodic cell, enforcing implicitly the transmission and radiation conditions. If we introduce two artificial boundaries \( I^\pm = \{x_2 = \pm b\} \) lying above resp. below the grating structure, denote by \( \Omega \) the rectangle \( (0, 2\pi) \times (-b, b) \) (cf.
Fig. 2), and define the $2\pi$-periodic function $u = v \exp(-i\alpha x_1)$, then the diffraction problem for TE polarization can be transformed to a variational problem for $u$ in the rectangle $\Omega$ (cf. [2], [3]):

\[
B_{TE}(u, \varphi) := \int_{\Omega} \nabla_\alpha u \cdot \nabla_\alpha \varphi - \int_{\Gamma} k^2 u \varphi + \int_{\Gamma^+} (T_\alpha^+ u) \varphi + \int_{\Gamma^-} (T_\alpha^- u) \varphi
\]

\[
= - \int_{\Gamma^+} 2i\beta p_3 \exp(-i\beta b) \varphi \quad \forall \varphi \in H^1_p(\Omega)
\]

Here $\nabla_\alpha = \nabla + i(\alpha, 0)$, and $H^1_p(\Omega)$ denotes the Sobolev space of functions which are $2\pi$-periodic in $x_1$. The non-local operators on the artificial boundaries are defined by

\[
(T^\pm_\alpha u)(x_1, \pm b) = - \sum_{n \in \mathbb{Z}} i\beta_n^\pm \hat{u}_n^\pm \exp(inx_1),
\]

where $\hat{u}_n^\pm$ are the Fourier coefficients of $u$ on $\Gamma^\pm$. Similarly, the TM diffraction problem can be formulated as

\[
B_{TM}(u, \varphi) := \int_{\Omega} \frac{1}{k^2} \nabla_\alpha u \cdot \nabla_\alpha \varphi - \int_{\Gamma} u \varphi + \int_{\Gamma^+} (T_\alpha^+ u) \varphi
\]

\[
+ \frac{1}{(k^+)^2} \int_{\Gamma^-} (T_\alpha^- u) \varphi = - \frac{1}{(k^+)^2} \int_{\Gamma^+} 2i\beta q_3 \exp(-i\beta b) \varphi \quad \forall \varphi \in H^1_p(\Omega)
\]

The sesquilinear forms $B_{TE}$ and $B_{TM}$ are strongly elliptic, i.e., coercive modulo compact operators on $H^1_p(\Omega)$. This leads to existence, uniqueness and regularity results for the variational equations in all cases of physical interest; see [3]. In particular, the TE and TM diffraction problems are uniquely solvable for all but a sequence of countable frequencies $\omega_j$, $\omega_j \to \infty$, and the solution is unique for all frequencies if one of the materials is absorbing. While the solution to the TE problem is always sufficiently smooth ($u \in H^2_p(\Omega)$), the TM solution may have singularities at the corner points of the grating. More precisely, near corners one has $u = r^\lambda f + g$, where $r$ denotes the distance to the corner point, the exponent $\lambda$ with $0 < \text{Re} \ \lambda < 1$ is determined by the refractive index of the grating material and $f, g$ are some smoother functions. In particular, if two materials with optical indices $k_1$ and $k_2$, respectively, meet at some corner, then $\lambda$ is the solution with minimal positive real part of the equation

\[
\left( \frac{\sin(\pi\lambda/2)}{\sin(\pi \lambda)} \right)^2 = \left( \frac{k_1^2 + k_2^2}{k_1^2 - k_2^2} \right)^2.
\]

Hence, the partial derivatives of $u$ are not square integrable, in general.
A detailed solvability and regularity theory of the conical diffraction problem, which is also based on a variational formulation, can be found in [7].

The variational approach leads to effective formulas for the gradient of cost functionals arising in the optimal design of binary and multilevel gratings. As an example, we present a formula for the partial derivatives of the Rayleigh coefficients with respect to the transition points $t_j$ of a binary profile (cf. Fig. 2) in the TM case:

$$D_j A_n^\pm = \frac{(-1)^{j-1}}{2\pi} \exp(-i\beta_n^\pm b)((k_0)^2 - (k^\pm)^2) \int_{\Sigma_j} \nabla_j (u) \cdot \nabla (w_\pm) \, dx_2,$$

where $\Sigma_j$ denotes the vertical segment at $t_j$, $u$ is the solution of the direct TM problem, $w_\pm$ solves the adjoint problem

$$B_{TM}(\varphi, w_\pm) = \int_{\Gamma^\pm} \varphi \exp(-inx_1) \, dx_1, \quad \forall \varphi,$$

and

$$\nabla_j (u) = \frac{1}{k+k_0} \left( \frac{k_0}{k^+} \partial_{x_1} u_+ + \partial_{x_2} u_+ \right),$$

where the plus sign denotes the one-sided limit as the interface is approached from the region above. Similar formulas are valid for the derivatives with respect to the height and the layer thicknesses. Note that the above gradient formula is only well-defined if the TM solution has mild singularities at the corners of the grating profile; see [5] for an approach in case of arbitrary singularities and more general non-smooth material interfaces, which also extends to conical diffraction [6].

5 Numerical Methods

The direct and adjoint diffraction problems have the form: Find $u \in H^1_p(\Omega)$ satisfying the equation

$$a(u, \varphi) = (f, \varphi), \quad \text{for all } \varphi \in H^1_p(\Omega),$$

where $a(u, \varphi)$ is a strongly elliptic sesquilinear form, and $(f, \varphi)$ stands for a linear and continuous functional on the function space $H^1_p(\Omega)$. The strong ellipticity implies that finite element approximations for all invertible problems under consideration lead to a uniquely solvable linear system of equations if the meshsize is sufficiently small. Moreover, the approximate solutions converge to the corresponding exact solution in the norm of the function space with optimal order.
Due to the rectangular geometry of binary gratings it is quite natural to choose piecewise bilinear functions as finite elements on a uniform rectangular partition of $\Omega = (0, 2\pi) \times (-b, b)$. This leads to a linear system with a block-tridiagonal matrix. The nonlocal boundary terms in the sesquilinear forms imply that the first and the last block of the main diagonal are fully occupied matrices, whereas the remaining blocks are sparse.

The computation of the nonlocal terms in the sesquilinear forms can be performed very efficiently with an accuracy comparable with the computer precision. Since the traces of the finite element functions on $\Gamma^\pm$ are piecewise linear periodic functions with uniformly distributed break points, it is possible to use recurrence relations for the Fourier coefficients of spline functions and convergence acceleration methods.

If e.g. the artificial boundary $\Gamma^+$ is divided into $m$ subintervals of equal length and the basis of hat functions $\{\varphi_j\}$ is used, then the form

$$\int_{\Gamma^+} (T^+_{\alpha\beta} \varphi_p) \overline{\varphi_q} \, dx_1 , \quad p, q = 0, \ldots, m - 1$$

corresponds to an $m \times m$ circulant matrix with the eigenvalues

$$\tau_0 = -2i\pi \beta , \quad \tau_p = -4i\pi \left(\frac{\sin(\pi p/m)}{\pi}\right)^4 \sum_{r=-\infty}^{\infty} \frac{\beta^+_{rm+p}}{m(r + p/m)^4} .$$

Thus one only has to expand

$$\frac{\beta^+_{rm+p}}{m} = \sqrt{\left(\frac{k^+}{m}\right)^2 - \left(\frac{\alpha}{m} + r + \frac{p}{m}\right)^2}$$

with respect to powers of $|r + p/m|$ and to use fast computation of the generalized Zeta function.

Usual FE approximations of the Helmholtz equation involve besides the approximation error also the so-called pollution error which increases together with the wavenumbers and enlarging domains. Roughly speaking, the pollution error is caused by the well-known fact that the discretization of the Helmholtz equation with the wave number $k$ results in an approximate solution possessing a different wave number $k_h$. In one-dimensional problems, for example, the usual piecewise linear FE solution of the equation $u'' + k^2 u = 0$ on a uniform grid has the discrete wave number

$$k_h = \frac{1}{h} \arccos \frac{2(3 - (kh)^2)}{6 + (kh)^2} = \frac{k^3h^2}{24} + O(k^5h^4) .$$

It turns out that this "phaselag" leads to suboptimal error estimates.

In the one-dimensional case it is possible to construct a generalized FEM without pollution by modifying the evaluation of the sesquilinear form. However, in higher dimensions it is not possible to eliminate the pollution in
the FE error by any modification of the evaluation of the sesquilinear form. Therefore, we extended an approach by Babuška et al [1] to design a so-called GFEM with minimal “phase lag” for piecewise uniform rectangular partitions; see [4] for the details.

The method was used to evaluate the reflection and transmission efficiencies of binary gratings on multilayer systems of different geometries and materials and it turned out to be robust and reliable in both the TE and TM case. Compared with the usual FEM the obtained results were accurate already for rather poor discretizations. In Fig. 3 we compare the numerical values of some reflection and transmission efficiencies versus the square root \( n \) of total number of grid points computed with the usual FEM and the GFEM on quadratic meshes for a binary grating with one groove per period situated on a layer. In each case the GFEM results differ already for \( n = 40 \) only by 2% from the corresponding values for \( n = 200 \), whereas the FEM results converge rather slowly to these values.

![Fig. 3. Comparison of efficiencies computed with FEM and GFEM](image)

The sparse structure of the matrix can be used to apply efficient direct or iteration methods for solving linear systems. We used a block version of the so-called sweep method, which utilizes the block-tridiagonal structure of the matrix and additionally the circulant properties of the dense blocks. Since the matrices of the discretized variational problems are nonsymmetric, we also applied preconditioned GMRES-type and BiCGstab methods as iterative solvers. The corresponding equations with averaged wave numbers \( k \) are good candidates for preconditioners, which can be inverted very efficiently using FFT.
In the case that the grating is situated on top of a multilayer stack (cf. Fig. 2), one can reduce the integration domain $\Omega$ used in the FE solution by taking into account that the solution is smooth within the layers. We introduce a new artificial boundary, $\tilde{\Gamma}^-$ say, into the first layer and new nonlocal boundary operators which model the layer system below $\tilde{\Gamma}^-$ together with the radiation condition for $x_2 < -b$. Combining the GFEM in the reduced domain with Rayleigh series expansions in the layer system then leads to a considerable reduction of the computational complexity; see [4] for a detailed presentation.

After having solved the linear system corresponding to the GFE discretization of the variational equations, the diffraction efficiencies are determined from the Fourier coefficients of the solution on the artificial boundaries. A stable recursive algorithm is used for the computation of the transmission efficiencies and the solutions on the layer interfaces, which appear in the gradient formulas.

6 Some Optimization Results

The GFEM and the gradient formulas were integrated into a computer program to find the optimal design of binary gratings with desired phase or intensity pattern for a given range of incidence angles or wavelengths. The optimal design problems were treated as nonlinear optimization problems with linear constraints, and we implemented a projected gradient algorithm and an interior point method for their numerical solution. Several numerical examples including polarisation gratings, high reflection mirrors and beam splitters successfully demonstrated the efficiency of the algorithm.

As a first example, we provide the optimization results for some beam splitters. The illuminating unpolarized wave with $\lambda = 0.633\mu m$ is normally incident from a dielectric medium with optical index $\nu = 1.5315$. Recall that the optical index of a material with permittivity $\epsilon$ is defined by $\nu = (\epsilon/\epsilon_0)^{1/2}$, where $\epsilon_0$ denotes the permittivity of the vacuum. Choosing the period $d = 1.266\mu m$, three diffraction orders propagate with angles $0^\circ$ and $\pm30^\circ$. Such beam splitters with large diffraction angles are useful in, e.g., optical clock signal distribution. The goal is

a) to maximize the efficiencies of the orders $\pm1$

b) to obtain maximal and equal efficiencies of all three orders

by optimizing the height $H$ and the fill factor $f$ of the grating with one groove per period. Using the cost functional (1), the following results have been obtained (cf. Fig. 4):

a) $H = 0.734\mu m$, $f = 0.72$,  

b) $H = 0.43\mu m$, $f = 0.58$.

Further applications in laser design are discussed in [9].

The next example concerns the design of a zero-order copper grating ($\nu = 12.7 + 51.1i$) as circular polarizer for a CO$_2$ laser with $\lambda = 10.6\mu m$ such that
in the range of incident angles $\theta \in (29^\circ, 31^\circ)$ the efficiencies of the reflected TE and TM polarized wave are maximal and the phase difference between them is close to $\pi/2$. Here one has to minimize the functional (2) extended over the range of incident angles, which possesses many local minima. One of the reasonable geometries is $d = 3.0\,\text{mm}$, $H = 1.65\,\text{mm}$, and the fill factor is 0.24. Table 1 contains the computed values.

$$
\begin{array}{c|ccc}
\theta & \text{TE} & \text{TM} & \text{phase} \\
29.0 & 97.50 & 95.72 & 90.72 \\
29.2 & 97.50 & 95.72 & 90.58 \\
29.4 & 97.51 & 95.72 & 90.45 \\
29.6 & 97.51 & 95.72 & 90.32 \\
29.8 & 97.52 & 95.72 & 90.18 \\
30.0 & 97.52 & 95.72 & 90.04 \\
30.2 & 97.53 & 95.72 & 89.91 \\
30.4 & 97.53 & 95.72 & 89.77 \\
30.6 & 97.54 & 95.72 & 89.63 \\
30.8 & 97.54 & 95.72 & 89.49 \\
31.0 & 97.55 & 95.72 & 89.35 \\
\end{array}
$$

Table 1. Zero order efficiencies and phase difference for circular polarizer

7 Conclusion

We focused on optimal design problems for binary and multilevel gratings, using exact formulas for the gradients of cost functionals and a fast and reliable method for the numerical solution of direct problems. The method is based on a variational formulation and combines a finite element method in the grating structure with Rayleigh series expansions in the layer system below the grating. The approach is not restricted to rectangular profiles,
but allows the numerical treatment of rather general diffraction structures, together with a rigorous convergence analysis.

We proposed a generalized finite element method (GFEM) with minimal pollution, which provides highly accurate numerical results in the computation of diffraction efficiencies. So far the extension of this method to more general (e.g., polygonal) grating profiles remains an open problem.

To solve optimal design problems for binary gratings by gradient descent, explicit formulas for the gradients with respect to the parameters of the grating profile and the thicknesses of layers have been developed. These formulas involve the solutions of direct and adjoint diffraction problems and reduce considerably the computational costs compared to simple difference approximations of the gradients.

We expect that this approach is also applicable to the inverse problem of profile reconstruction from far field data. Another challenging direction of future research is the efficient solution of direct and inverse problems for non-periodic and three-dimensional diffractive structures.

References

Computation of Electromagnetic Fields for a Humidity Sensor

Wolfgang Hackbusch\textsuperscript{1} and Steffen Börn\textsuperscript{2}

\textsuperscript{1} Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany
\textsuperscript{2} Universität Kiel, Germany

\textbf{Abstract.} We simulate an electromagnetic humidity sensor in order to determine its eigenfrequencies. The underlying model is given by Maxwell’s field equations, they are transformed into two decoupled two-dimensional problems by eliminating the magnetic field and changing to cylinder coordinates.

A time-harmonic approach leads to two generalized eigenproblems that are discretized using appropriate finite elements. The usual multigrid solvers are modified in order to cope with the anisotropic structure of the resulting matrices. These modifications can be used in order to build robust eigenproblem solvers of optimal asymptotic complexity.

Numerical experiments show that the resulting method is able to capture those effects that are relevant for the sensor under investigation.

\section{Introduction}

In many applications, it is desirable to determine the humidity of a given material without affecting it, i.e., without drilling holes or taking samples. One example might be an installation for baking bread that needs to know whether the bread is already done or needs more time in the oven.

The humidity of a material corresponds to its dielectricity, so a possible approach is to devise a way of measuring this property. The idea is to measure the electromagnetic eigenfrequencies of a system in order to derive some information on the dielectricity in the material.

Since the material should not be affected by the sensor, we attach a resonator and an antenna. The antenna is used to induce a time-harmonic electromagnetic field of changing frequency, the effects of this field can be measured in the resonator.

If the amplitude of the field is high, the frequency currently applied by the antenna is likely to be close to an eigenfrequency, so by trying different frequencies in a selected range, the part of the spectrum corresponding to this range can be approximately determined.

Of course, the eigenfrequencies alone would not allow us to recover information on the dielectricity. To do so, we need reference values corresponding to known dielectricities. Obviously, there are two way to get these: By performing extensive experiments for different dielectricities and geometries or by simulation. The simulation approach has a number of advantages: It is
much more cost-effective than the experimental approach, it allows us to perform a large number of simulated experiments in a short time and it gives us more information than any experiment could by computing the entire field, not just its parameters in a number of reading points.

Our partner, the Sensortech GmbH & Co KG of Neumünster, has done a number of experiments to verify the usability of the sensor and asked us to provide them with a tool that simulates the system in order to determine optimal parameters and geometries.

In the following, we will describe the basic model for the system’s physics, the transformations necessary to get two-dimensional variational eigenproblems, the modified multigrid techniques used to solve them. We conclude by presenting numerical results for simple domains.

2 Maxwell’s Equations

The starting point for our simulation are Maxwell’s equations. By eliminating the magnetic field and assuming isotropic linear material laws, they take the form

$$- \frac{\partial}{\partial t} j_{\text{exc}} = \nabla \times \left( \frac{1}{\mu} \nabla \times E \right) + \sigma \frac{\partial}{\partial t} E + \varepsilon \frac{\partial^2}{\partial t^2} E.$$

We use the time-harmonic approach

$$E(x,t) = E_0(x) e^{i\omega t},$$
$$j_{\text{exc}}(x,t) = J_0(x) e^{i\omega t}$$

to translate this to

$$-i\omega J_0 = \nabla \times \left( \frac{1}{\mu} \nabla \times E_0 \right) + (\sigma \omega - \varepsilon \omega^2) E_0.$$

$\omega \in \mathbb{R}$ is an eigenfrequency whenever the right-hand operator is singular, i.e., when there is a field $E_0 \neq 0$ satisfying

$$(\varepsilon \omega^2 - \sigma \omega i) E_0 = \nabla \times \left( \frac{1}{\mu} \nabla \times E_0 \right).$$

Since the right-hand side is real, the imaginary part of the left-hand side must vanish, i.e., $\sigma \omega E_0 = 0$. In our case, the conductivity $\sigma$ is very small compared to the other quantities, so we can assume $\sigma = 0$ turning the above equation into a real eigenproblem.

To apply finite element techniques, the problem has to be translated into a variational form on a domain $\Omega$. Introducing a test function $v$ and performing partial integration assuming perfectly conducting boundaries leads to

$$\omega^2 \int_{\Omega} \epsilon(x) \langle E_0(x), v(x) \rangle dx = \int_{\Omega} \frac{1}{\mu} \langle \nabla \times E_0(x), \nabla \times v(x) \rangle dx,$$

a symmetric positive definite eigenproblem.
In our case, the domain $\Omega$ is axisymmetric, so by transforming the above eigenproblem to cylinder coordinates, it takes the form

$$2\pi\omega^2 \int_{\Omega} r \epsilon(E_{r,z}, v_{r,z}) d(r, z) = 2\pi \int_{\Omega} \frac{r}{\mu} (\text{curl}_2 E_{r,z})(\text{curl}_2 v_{r,z}) d(r, z),$$

$$2\pi\omega^2 \int_{\Omega} r \epsilon E_\phi v_\phi d(r, z) = 2\pi \int_{\Omega} \frac{r}{\mu} \left( \begin{array}{cc} 1 & 0 \\ 0 & 1/r^2 \end{array} \right) \nabla E_\phi, \nabla v_\phi d(r, z).$$

These eigenproblems are independent. The first corresponds to field components in the $(r, z)$-plane, while the second corresponds to angular components.

It is well known that the so-called edge elements introduced by Nédélec [11] are especially well suited for the discretization of problems involving the curl operator, so the obvious approach to solving the planar problem is to discretize it using a two-dimensional variant of these elements. We use first order elements on axiparallel rectangles and thus reduce the infinite-dimensional to a finite-dimensional eigenproblem.

The angular component is a scalar anisotropic problem where the anisotropy is aligned with the coordinate axes, so we use bilinear Lagrange elements on axiparallel rectangles for its discretization and get another finite-dimensional eigenproblem.

3 Eigenproblem Solvers

The eigenproblems derived in the preceding section can be treated directly using standard methods like QZ-, Lanczos- or Jacobi-Davidson-iterations (cf. [6]), but these usually do not have optimal complexity, i.e., the amount of work grows more than linearly with the number of unknowns. Since we have to be able to deal with large numbers of unknowns (in the range of several millions) in order to reduce the approximation error, these techniques would simply be too slow.

Among the methods with optimal complexity, multigrid algorithms are the most common. We have investigated two variants of this type: The eigenvalue multigrid method by Hackbusch (cf. [7]) and the preconditioned subspace iteration by Knyazev et al. (cf. [5]). The first algorithm allows us to compute single eigenvectors and eigenvalues, while the second focusses on blocks of eigenpairs corresponding to the smallest eigenvalues.

Since both of our eigenproblems do not correspond to “harmless” $H^1$-elliptic bilinear forms, we cannot expect these algorithms to work without modifications. So we first studied the performance of multigrid solvers for linear equations in order to find out how they can be adapted to our type of problem.

3.1 Multigrid Method for Linear Equations

The bilinear form of the angular problem corresponds to an anisotropic problem with unbounded coefficients. The coefficients depend only on the
r-coordinate and, since the anisotropy matrix is diagonal, the anisotropic effects are aligned with the coordinate axes.

To construct a robust smoother for this type of problem, we combined a row-block smoother with a semicoarsened grid hierarchy. This approach is well known from the field of computational fluid dynamics and has been applied to scalar anisotropic equations before, e.g., by Pflaum (cf. [12]).

For rectangular domains, it is possible to prove optimal robustness of this modified method (cf. [3]): The convergence bound does not depend on the behaviour of the parameters, as long as they are positive almost everywhere.

The planar problem is related to Maxwell’s equation in two dimensions, but, too, has unbounded coefficients and is therefore not $H$(curl)-elliptic. In addition, another problem has to be dealt with: The treatment of Maxwell’s equation with multigrid techniques requires special smoothing techniques to take care of a subspace corresponding to the kernel of the curl operator (cf. [9]).

By once more combining row-block smoothing techniques with a semicoarsened grid hierarchy, we get a robust multigrid method. The robustness proofs in [4,1,2] demonstrate that we get the same bound for the convergence rate as in the scalar case for a similar class of problems.

### 3.2 Eigenvalue Multigrid Iteration

The eigenvalue multigrid iteration (cf. [7,8]) is based on a multigrid solver for indefinite problems that result from a spectral shift of the operator under consideration.

The main advantage of this method is its ability to compute single eigenpairs, the main disadvantage in our context is the fact that, in some cases, the shift leads to singular diagonal blocks, so the block smoothing procedure is no longer stable.

It is possible to stabilize the block smoother by using inexact shifts, this leads to relatively fast convergence, but the selection of these shifts is rather difficult and up to experiments.

### 3.3 Preconditioned Subspace Iteration

The problems of the eigenvalue multigrid solver are due to the indefinite matrices that have to be treated in its course. So it is obvious to look for eigenvalue solvers that do not rely on the (approximate) solution of indefinite problems. The preconditioned eigenvalue solvers satisfy this requirement. The simplest variant of these solvers is the preconditioned subspace iteration of Knyazev (cf. [5,10]).

The original form of this iteration computes a fixed number of eigenpairs corresponding to the lowest eigenvalues and requires only a suitable preconditioner for the original problem.
In the case of the planar problem, we are not interested in the lowest eigenvalues, but in the lowest non-zero eigenvalues, so we modify the iteration by introducing an approximate projection filtering out the kernel components. This leads to a fast and very robust solver. One additional advantage of this approach is that the algorithm can be parallelized in a straightforward way, so we are able to make use of the SMP abilities of our workstations in order to speed up the computations.

There is a price to be paid for the robust convergence of the subspace iteration: All of the lowest non-zero eigenvalues have to be computed simultaneously, so a large amount of memory is required.

4 Numerical Results

In a first simulated experiment, we examine a cylinder that is composed of a resonator part with dielectricity $\varepsilon_r = 80$ and of a sample part with a dielectricity $\varepsilon_s \in [1, 8]$. This parameter interval is split into 40 segments and for each segment the eight lowest non-zero eigenvalues are computed using the preconditioned subspace iteration. Using a discretization with first-order edge elements on a regular grid, we have to compute about half a million degrees of freedom for each eigenvector and each segment. After running for about 17 hours on a SUN Ultra 80 machine with four processors, the Fig. 1 is produced.

Apparently, all the eigenvalues decrease monotonous with growing dielectricity, as is to be expected. The interesting point is that some eigenvalues

![Graph](image_url)

Fig. 1. Dependence between eigenfrequencies and sample dielectricity in a cylinder
Fig. 2. Dependence between eigenfrequencies and sample dielectricity in a domain composed of two concentric cylinders

appear to decrease much faster than others. If the sensor is built to probe the corresponding frequency range, the reconstruction of the dielectricity can benefit from this behaviour.

Of course, not only cylindrical domains are of interest. We can for example also simulate an axisymmetric domain with L-shaped cross-section. We set the dielectricity to $\varepsilon_r = 80$ in the upper part of the domain and once more consider dielectricities between 1 and 8 in the lower part. The result is given in Fig. 2 and exhibits the same behaviour as in the case of the cylinder: The eigenvalues decrease monotonous and some eigenvalues decrease much faster than others.

In this second experiment, the results are not as good as in the first one: There are some rather large gaps at the points where the distance between eigenvalues is small. We performed experiments with higher grid resolutions. In these experiments, the size of the gap was significantly reduced, so it seems to be a consequence of discretization errors introduced by singularities at the non-convex edge.

5 Conclusion and Extensions

To simulate the humidity sensor, we have to combine modern tools like

- edge elements for the discretization of Maxwell-type equations,
- hybrid smoothers for treating kernel components of the iteration error

and
preconditioned eigensolvers for the computation of eigenvalues and eigenvectors without the necessity to solve indefinite systems with classical techniques like

- multigrid preconditioners for optimal-order preconditioners and
- block smoothers for treating the unbounded coefficients

while introducing some problem-specific modifications like

- the approximate projection included in the eigensolver to remove zero eigenvalues and
- the reduced implicit block smoother that improves the efficiency of the multigrid solver.

By using this combination, it is possible to determine frequency ranges that allow for a stable reconstruction of the dielectricity and therefore of the humidity of the sample under consideration.

There are a number of possible extensions to our work: An adaptive grid refinement strategy (possibly by some local defect correction or domain decomposition method in order to keep the data structures simple) could reduce the approximation error for non-convex domains. Faster preconditioned eigensolvers (like the optimal block conjugate gradient method suggested by Knyazev in [10]) could reduce the number of necessary iteration steps.

It might even be possible to work on infinite samples (e.g., by coupling the finite element approach with a boundary integral formulation of the exterior domain problem).

References

V. Crystal Growth, Semiconductors

Simulation of Industrial Crystal Growth by the Vertical Bridgman Method
  G. Dziuk, S. Boschert, A. Schmidt, K. G. Siebert, E. Bänsch, K.-W. Benz, T. Kaiser

Numerical Simulation and Control of Industrial Crystal Growth Processes
  K.-H. Hoffmann, A. Voigt, M. Metzger

Optimal Control of Sublimation Growth of SiC Crystals
  J. Sprekels, O. Klein, P. Philip, K. Wilmański

Mathematical Modelling and Numerical Simulation of Semiconductor Detectors
  H. Gajewski, H. Chr. Kaiser, H. Langmach, R. Nürnberg, R. H. Richter

Optimal Design of High Power Electronic Devices by Topology Optimization
  R. H. W. Hoppe, P. Böhm, G. Mazurkevitch, S. Petrova, G. Wachutka, E. Wolfgang

Modelling and Simulation of Strained Quantum Wells in Semiconductor Lasers
Simulation of Industrial Crystal Growth 
by the Vertical Bridgman Method

Gerhard Dziuk\textsuperscript{2}, Stefan Boschert\textsuperscript{1}, Alfred Schmidt\textsuperscript{2}, Kunibert G. Siebert\textsuperscript{2}, Eberhard Bänsch\textsuperscript{3}, Klaus-Werner Benz\textsuperscript{1}, and Thomas Kaiser\textsuperscript{1}

\begin{itemize}
\item \textsuperscript{1} Kristallographisches Institut, Universität Freiburg, Hebelstr. 25, 79104 Freiburg, Germany
\item \textsuperscript{2} Institut für Angewandte Mathematik, Universität Freiburg, Hermann-Herder-Str. 10, 79104 Freiburg, Germany
\item \textsuperscript{3} Zentrum für Technomathematik, Universität Bremen, Postfach 330440, 28334 Bremen, Germany
\end{itemize}

http://www.krist.uni-freiburg.de

http://www.mathematik.uni-freiburg.de/IAM

http://www.math.uni-bremen.de/zetem

Abstract. Single crystals of Cadmium-Zinc-Telluride are used as a substrate material for the production of infrared detectors and are usually grown by the vertical Bridgman method. We present a simulation of the whole growth process in two steps: In the first step, the (stationary) heat transport in the furnace is modeled and calculated for different positions of the ampoule. This provides information about the most important parameter during this process: the temperature distribution in furnace and ampoule. The obtained temperatures are then used in the second step as boundary conditions for the (time dependent) simulation of temperature and convection in the ampoule. Only the use of adaptive finite element methods allows an efficient numerical simulation of the moving phase boundary, the convection in the melt and the temperature distribution in melt and crystal. Numerical results are presented for both furnace and ampoule simulations.

1 Introduction

Single crystals of the semiconductor Cadmium-Zinc-Telluride (Cd,Zn)Te are excellent substrate materials for growth of thin Mercury-Cadmium-Telluride (Hg,Cd)Te layers, which are used as detector material for infrared radiation, see [7]. Such (Cd,Zn)Te crystals are usually grown by the vertical Bridgman method, where an ampoule with material melt is moved in a temperature field such that the material slowly crystallizes. Fig. 1 shows the schematics of the furnace with ampoule and heating distribution.

The material, which is placed in a sealed quartz-glass ampoule, is molten in the hot area of the crystal growth furnace. To start the crystallization process, a relative motion between the furnace and the ampoule is started resulting in a temperature reduction at the lower end of the growth ampoule. When the temperature at the bottom falls below the melting temperature the crystallization process starts. With further movement more and more
material solidifies until finally the whole crystal is grown. The most important parameter during this process is the temperature distribution in furnace and ampoule.

In the industrial production of infrared detectors, a significant dependence of the efficiency of the manufactured detectors on the quality of the substrate material is observed. Therefore an investigation of the growth conditions for the substrate material becomes necessary, to optimize the production process. Key parameters which determine the quality of the grown crystals are the shape of the phase boundary and the temperature distribution in its vicinity. Experimental determination of such values during the growth experiment is almost impossible. To determine these temperatures, numerical simulations are a feasible way to overcome the technical problems in the direct measurement.

The simulations are performed in a two step process [5]. The first step is a global simulation of the heat distribution in the growth furnace. Here, the (stationary) heat transport in the furnace is calculated for different positions of the ampoule. To obtain the temperature distribution in the furnace, numerical simulations of the heat transport in the whole assembly have been performed similarly to [4] by a 2-D axially symmetric finite element model, using the commercial code FIDAP [11].

The second step is a local simulation of the (time dependent) phase transition and convection in the ampoule. The objective here is the investigation of the influence of convection on the moving phase boundary in the ampoule. The temperature distributions obtained from step one, the global furnace simulation, are used as boundary conditions.

The model for the local simulation consists of a Stefan problem with convection in both solid and liquid phases and the incompressible Navier–Stokes equations in the melt (compare Sect. 3). Effects of varying concentration are neglected in the simulations presented here. Solvers for the Stefan problem
and the Navier–Stokes equations on time dependent domains are combined in an adaptive finite element algorithm for the solution of the coupled system and are implemented in the adaptive finite element toolbox ALBERT [18,19]. The underlying finite element meshes are adapted to the solution in each time step using information from a posteriori error estimators. Only the use of adaptive finite element methods allows an efficient numerical simulation of the moving phase boundary, the convection in the melt, and the temperature distribution in melt and crystal.

2 Modelling of the Crystal Growth Furnace

In this section we discuss the modelling of the crystal growth furnace and the influence of the growth ampoule on the temperature distribution. Then the calculation of the heat transport and relevant results from the simulation are presented.

2.1 Furnace and Crystal Growth

The crystal growth experiments are performed in a multi-zone resistance furnace. This furnace consists of a hot zone and a cold zone, subdivided into ten sub-zones each, and separated by an adiabatic zone. The upper hot heating sub-zones are all kept at a fixed temperature (1140°C) above the melting point of (Cd,Zn)Te (1092°C). The lower cold sub-zones are adjusted to 840°C to realize a temperature gradient in the area of the adiabatic zone. For the solidification process the furnace is slowly moved upwards (30 mm/day) to generate the temperature decrease in the ampoule.

The diameter of the growth ampoule is 65 mm, and the length of the grown crystals is typically around 125 mm, resulting in crystals of about 2 kg weight. This makes it necessary to have a furnace with an inner diameter of ≈ 95 mm and a height of more than 1 m to ensure a uniform temperature distribution. Due to the fact that the temperatures in the hot area of the furnace are well above 1100°C, radiation plays an important part in the heat transport. Furthermore, convection (of the gas between furnace wall and ampoule and of the melt) as well as conduction contribute to the overall heat transport. In particular, the heat conduction in the furnace insulation is another key parameter for the heat transport in the assembly [9].

2.2 Modelling of the Heat Transport

For the calculation of the heat transport several assumptions and simplifications have to be made. Thus the convection in the furnace, either of the gas between ampoule and furnace wall and in the melt, is modeled by an effective conductivity consisting of the pure conductivity and an additional part from convection. In the furnace insulation pure conductivity is assumed.
The influence of convection (in the porous medium) is neglected as well as inhomogeneities caused by power supply, furnace control etc.

Thus the heat equation (in dimensional form) has to be solved for the temperature \( \vartheta \):

\[
\rho c \partial_t \vartheta = \nabla (\kappa \nabla \vartheta) + H
\]

where \( \rho \) is the density, \( c \) the specific heat, \( \kappa \) the conductivity and \( H \) gives the contribution of an volumetric heat source. On a boundary segment the general heat transfer condition for the normal heat flux \( q \) is

\[
q = -\mathbf{n} \cdot (\kappa \nabla \vartheta) = q_c + q_r
\]

which consists of a conductive part \( q_c \) and a radiative part \( q_r \). The latter one is modeled by an additional nonlinear equation for the radiative heat exchange between all surfaces which includes mutual visibility and shadowing, see [5] for details.

In the furnace simulation the phase transition in the ampoule is not computed. But using a temperature dependent heat conductivity (different values for temperatures below or above the melting point), the distinct behavior of the solid and liquid phase is considered. The simulations are performed for several positions of the ampoule in the furnace. These values can be used as boundary conditions for the calculation of the convection/solidification problem discussed in more detail in Sect. 3. But also some valuable information for the crystal grower concerning the ampoule-design and crystal growth process can be derived, which will be presented in the following.

During industrial production, the temperature at the bottom of the crystal is monitored. Temperatures from our simulations are compared with such measurements, performed by M. Bruder (AIM, AEG Infrarot Module GmbH). A good agreement between the experimental values and the simulation is observed.

### 2.3 Results of the Furnace Simulation

A typical temperature distribution in the whole crystal growth assembly is shown in Fig. 2. The temperature difference between the neighbouring isotherms is 200°C. The shape of the ampoule and the melt is indicated as well. It can be seen that the ampoule is completely in the hot area of the furnace. Later, during the growth process, the ampoule is shifted downwards to start the crystallization.

An important parameter for the crystal grower is the temperature distribution in the melt and crystal at different ampoule locations in the furnace. Of special importance is the position and shape of the melting-point isotherm, which corresponds to the phase boundary in the simplified model used here.

In Fig. 3 the temperature distribution for different ampoule positions is given. From left to right a larger portion of material crystallizes. The shape
**Fig. 2.** Calculated temperature distribution in furnace and ampoule. The shape of the ampoule and several isotherms are given.

**Fig. 3.** Temperature distribution in crystal and melt at different ampoule positions in the furnace (ampoule position measured from bottom of furnace in mm). The temperature difference between the isotherms is 5°C. The position of the melting isotherm (θ = 1092°C) is also marked (dashed line).
of the melting-point isotherm changes from slightly concave in the beginning (unfavorable for crystal growth) to slightly convex in the end. The absolute position of the phase boundary also moves downwards (during the whole experiment \(\approx 20\) mm). This means that the growth velocity is lower than the velocity of the ampoule movement. As a consequence the crystal growth process needs more time to finish than expected. Such information is important for the crystal grower because it gives proper means to plan the experiment – with the pulling rate of 30 mm/day there is nearly one day difference between the expected and the calculated end of the growth process.

![Diagram showing quartz-tube and quartz-rod with nucleation points]

**Fig. 4.** Influence of the ampoule mounting on the nucleation process. The change from a quartz-glass tube (on the left) to a solid quartz rod (on the right) improves the nucleation process.

The thermal simulation can also be used to improve the design of the growth ampoule. In the original configuration the ampoule is placed upon a quartz tube during the experiment. From the simulation can be seen that this tube influences the thermal field in the beginning of the solidification process. The coldest place is not the tip of the ampoule but rather the place where ampoule and quartz tube are connected. This leads to an undesired nucleation at that place. To overcome this problem a different type of ampoule mounting has been tested: the quartz tube is replaced by a solid quartz rod. From Fig. 4 it can be seen that after this modification the position of the first nucleation is shifted to the tip of the ampoule. The undesired nucleation at the side can thus be avoided. The first crystals produced by AIM with such a modified ampoule mounting show an improved quality of the grown crystal.
3 Local Simulation in the Ampoule

In this section we investigate the influence of the convection on the moving phase boundary in the ampoule, where the results from the global furnace simulations are used as boundary conditions for the temperature at the ampoule boundary in the Stefan problem.

Once the pulling rate of the crystal growth process is given, there is a one-to-one relation between an ampoule position \( z \) in the furnace and the corresponding time \( t \). Temperature values at the ampoule boundary are computed for a discrete number of ampoule positions by the global simulation. The piecewise linear interpolant (in time) of these discrete temperature values is used for the calculation of enthalpy boundary values \( u_\theta \) in (1c).

Using the Boussinesq approximation, the moving phase boundary and convection in the ampoule \( \Omega \) are modeled by system (1)-(2) for the temperature \( \vartheta \), energy density (enthalpy) \( u \), velocity \( \mathbf{v} \), and pressure \( p \). The time dependent liquid sub-domain is defined via the melting temperature \( \vartheta_m \) by

\[
\Omega_l(t) := \{ x \in \Omega : \vartheta(x, t) > \vartheta_m \}.
\]

Let \( k \) denote the heat conductivity, \( \rho \) the density, \( c \) the specific heat, \( L \) the latent heat, \( \chi \) the characteristic function of the liquid phase \( \Omega_l \), \( \eta \) the kinematic viscosity, \( \beta_\theta \) the thermal expansion coefficient, and \( \mathbf{g} \) the vector of gravity.

The Stefan problem with convection applies in both liquid and solid phases \( (\Omega = \Omega_l(t) \cup \Omega_s(t)) \),

\[
\begin{align*}
\partial_t u + (\chi \mathbf{v}) \cdot \nabla u - \nabla \cdot (k \nabla \vartheta) &= 0 & \text{in } \Omega, t > 0, \\
u &= \rho(c \vartheta + \chi L) & \text{in } \Omega, t > 0, \\
u &= u_\theta & \text{on } \partial \Omega, t > 0, \\
u(\cdot, 0) &= u_0 & \text{in } \Omega,
\end{align*}
\]

and the incompressible Navier–Stokes equations hold in the liquid phase \( \Omega_l(t) \),

\[
\begin{align*}
\partial_t \mathbf{v} - \eta \Delta \mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{v} + \frac{1}{\rho} \nabla p &= (1 - \beta_\theta (\vartheta - \vartheta_m)) \mathbf{g} & \text{in } \Omega_l, t > 0, \\
\text{div } \mathbf{v} &= 0 & \text{in } \Omega_l, t > 0, \\
\mathbf{v} &= \mathbf{0} & \text{on } \partial \Omega_l, t > 0, \\
\mathbf{v}(\cdot, 0) &= \mathbf{v}_0 & \text{in } \Omega_l, t = 0.
\end{align*}
\]

The system is coupled by the convection term \((\chi \mathbf{v}) \cdot u\) in (1a) and the Boussinesq forcing term in (2a). The geometry for the local simulation with solid and liquid phases is shown in Fig. 5.

The solver for the ampoule problem is based on existing solvers for the classical Stefan problem and the incompressible Navier–Stokes equations,
implemented using the finite element toolbox ALBERT [18,19]. The adaptive method combines finite elements in space on an underlying triangulation consisting of triangles (in 2-D) and tetrahedra (in 3-D) with an appropriate discretization in time. Extensions of the standard solvers were needed to handle convection in the Stefan problem and a time dependent domain for the Navier–Stokes equations.

3.1 Adaptive Finite Element Methods for Time Dependent Problems

Given some tolerance for the error between the discrete and true solution, the aim of an adaptive method is the efficient approximation of the solution within this prescribed tolerance. For efficiency the underlying grid should be as coarse as possible but fine in regions where a high resolution is needed to keep the error below the given tolerance. Usually, these regions of high resolution move in time and thus, the underlying triangulation is adapted in each time step. Additionally, the time step size is adjusted, i.e. it is reduced if the solution changes more rapidly in time and is enlarged if the problem becomes more stationary.

Since the true solution is not known, information about the error between discrete and true solution has to be obtained by computable quantities. This goal can be achieved by a posteriori error estimators, which involve only information about the discrete solution and data of the problem and are thus computable [21]. The a posteriori error estimator is given by local indicators, contributions on single mesh elements, and by a time discretization estimator.

The adaptive method uses such estimators for the adaptation of grids and time step sizes. Meshes and time step sizes are adjusted by local refinement and coarsening of mesh elements (see Fig. 6) and reduction or enlargement of the time step size for equidistribution of local contributions over mesh elements and time steps. The aim is the reduction of computational work while the error is below the given tolerance.

Starting with an initial grid which is adapted to initial data, we construct a sequence of (variable) time step sizes and conforming triangulations. In each time step, the adaptive algorithm iteratively solves the discrete problem
Fig. 6. Atomic refinement and coarsening operation in 2-D and 3-D

on the current mesh, estimates the error, and adapts the mesh and time step size until the given tolerance is reached. We start with the mesh and time step size from the last time step. The mesh adaptation is performed by local refinement and coarsening. All elements with large local indicator are marked for refinement and those elements with local indicators much smaller than the local tolerance are marked for coarsening. Coarsening is an important ingredient for time dependent problems since regions where a high resolution is needed move in time. As a consequence, the highly refined regions have to move in time, too, and parts of the previously refined regions have to be coarsened in order to keep the mesh as coarse as possible.

Additionally, the time step size is adjusted to the temporal behavior of the solution. If the time step size is reduced or the grid is refined, the discrete problem is solved again, and the adaptive process is iterated. Usually, only a small number of adaptive iterations are needed in each time step. If the changes of the solution in time are very small, an explicit adaptive strategy is also sufficient, where the problem is solved only once in each time step.

The refinement and coarsening routines construct a sequence of nested meshes with a hierarchical structure. During mesh modifications, degrees of freedom (DOFs) are created or deleted and finite element functions must be transformed to the new finite element spaces. For example, in time dependent problems we have to transform the solution from the old time step during these modifications. Usually, these transformations can be realized by a sequence of local operations during the atomic refinement/coarsening operation, shown in Fig. 6.

ALBERT provides all tools for the local mesh modifications. It supports different finite element spaces on the same triangulation, administrates all used DOFs and all finite element data is automatically transformed during mesh modifications. For the ampoule simulation, different ansatz spaces for temperature, pressure, and velocity are used, see Sect. 3.4.

3.2 Navier–Stokes Equations on a Time Dependent Domain

For transient flow problems on fixed domains a solver for the incompressible Navier–Stokes equations is already implemented in ALBERT. The finite element discretization is based on the Taylor–Hood element using piecewise polynomials of degree $p$ for the velocity and of degree $p - 1$ for the pressure approximation. For the time discretization the so called fractional $\theta$-scheme is
used which was introduced in [6]. We use this scheme in a variant as operator splitting, which decouples the two fundamental difficulties in the numerical treatment of the incompressible Navier–Stokes equations: the solenoidal condition and the nonlinearity. Each time step is split into three fractional steps. In the first and third step we compute a divergence free velocity field with corresponding pressure by solving a linear saddle point problem and handling the nonlinearity explicitly. In the second step we disregard the solenoidal condition and solve a non-linear elliptic problem for the velocity. This approach works well in the case of moderate Reynolds number [1,2]. The stability and convergence properties of this scheme for the time discretization of the Navier–Stokes equations are analyzed in [12,14,15].

One main difficulty for the discretization of system (2) is the fact that the domain is changing in time. The domain for the Navier–Stokes equations will shrink from the full ampoule at the beginning of the ampoule simulation to the empty set at the end. We have analyzed and implemented the following approaches for solving parabolic problems on time dependent domains.

**Penalty Approach.** The liquid phase is a sub-domain of the complete ampoule for all times, i.e. \( \Omega_t(t) \subset \Omega \) for all \( t \geq 0 \). Extend given data of the differential equation to \( \Omega \) and solve the equation in \( \Omega \) and treat the boundary condition

\[
v = 0 \quad \text{on} \quad \Gamma(t) := \partial \Omega_t(t) \setminus \partial \Omega
\]

by penalizing the solution for being non zero on \( \Gamma(t) \). We prove for a linear model problem [5] that this penalty approach leads to a stable discretization.

The existing solver was extended to time dependent domains by adding the penalty term. This term reduces to a boundary integral over \( \Gamma(t) \) that has to be assembled into the system matrix for all parts of the fractional \( \theta \)-scheme. On the one hand, the penalization part makes the resulting scheme more stable, because we add a positive semi-definite term. On the other hand, for stronger penalization this term becomes dominant and makes the system more stiff. Thus, a special preconditioning for an efficient solution of the discrete system is needed. For more details about this approach see [5].

**Remark.** The *fictitious domain method* was introduced in [13] for the solution of the Navier–Stokes equations on a fixed domain \( \omega \) with a complicated boundary \( \Gamma = \partial \omega \). The equations are solved on a bigger domain \( \Omega \supset \omega \) which can be discretized more easily, and the boundary condition on \( \Gamma \) is treated as a constraint, similar to the solenoidal condition \( \text{div} \, v = 0 \). The objective in this approach is to avoid a complicated meshing of the domain \( \omega \) but the method can directly be extended to a time dependent domain. Due to the additional constraint and the discretization of the corresponding Lagrange multiplier (jump of the normal derivatives of \( v \) across \( \Gamma(t) \)) this method turns out to be unstable in practice.
**Sub-domain Approach.** From a triangulation of the whole ampoule, we construct a triangulation for the discretization of the Navier–Stokes equation by collecting all mesh elements which belong completely to the liquid phase, compare [3]. The original fractional $\theta$-scheme is then applied on finite element spaces corresponding to this discrete sub-domain. No-slip boundary values for the velocity are prescribed explicitly on the boundary of this sub-domain. In contrast to the penalty approach, the resulting discrete linear systems do not become more stiff and can thus be solved more efficiently. Additionally, the dimensions of the finite element spaces get smaller when the liquid sub-domain shrinks.

### 3.3 Stefan Problem with Convection

After a Kirchhoff transformation and an appropriate scaling of variables, the equation for enthalpy $u$ and temperature $\vartheta = \beta(u)$ reads

$$
\begin{align*}
\partial_t u + v \cdot \nabla u - \Delta \beta(u) &= 0 \quad \text{in } \Omega, \ t > 0 \\
u &= u_\partial \quad \text{on } \partial \Omega, \ t > 0 \\
u(\cdot, 0) &= u_0 \quad \text{in } \Omega
\end{align*}
$$

(3a) \quad (3b) \quad (3c)

Assuming piecewise constant physical coefficients in the phases, the nonlinear function $\beta$ connecting temperature and enthalpy is piecewise linear, monotone, and vanishes identically for $u \in [0, 1]$, resulting in a degenerate parabolic equation for the energy density. The full discretization of the Stefan problem combines piecewise linear finite elements in space and a time discretization including the convection that is based on the method of characteristics [10]. Adaptive finite element methods based on a posteriori error estimates for such equations were derived in [8,16,17]. Such error indicators are used to adapt the meshes and time step sizes for the coupled ampoule problem.

### 3.4 Numerical Method and Results for the Ampoule Problem

We present results of two- and three-dimensional simulations for model problems. While using non physical coefficients and parameters, they show that our numerical method is able to efficiently solve the problems of phase transition and convection in the melt. Results for physical parameters as well as axisymmetric simulations are work in progress and will be published in a forthcoming paper.

The simulations are performed for an ampoule with 65 mm in diameter and 127 mm long. A fixed temperature profile, piecewise linear in $z$ direction, is shifted vertically with a pulling rate of 0.5 mm/sec to define the time dependent boundary values for the enthalpy. We use a Reynolds Number $Re = 1000$, corresponding to a fluid which is about twice as viscous as $(Cd,Zn)Te$, and a dimensionless heat conductivity $k = 1.0$ in both phases. All computations were performed with the adaptive finite element toolbox.
ALBERT [18] using the cubic/quadratic Taylor–Hood element for velocity and pressure; figures were prepared using the GRAPE visualization library [20].

We use the following semi-implicit discretization of system (2),(3). In the explicit adaptive strategy for the ampoule simulation, the sequence of calculations for time step \( n + 1 \) is:

- Solve the Stefan problem for new temperature \( \Theta^{n+1} \) and enthalpy \( U^{n+1} \) with the given velocity \( V^n \) from the last time step.
- Define the new discrete liquid phase \( \Omega_l^{n+1} \).
- Solve one time step of the fractional \( \theta \)-scheme for the Navier–Stokes equations for new velocity \( V^{n+1} \) and pressure \( P^{n+1} \) with the given temperature \( \Theta^{n+1} \) and liquid sub-domain \( \Omega_l^{n+1} \).
- Calculate error estimators for the Stefan problem with convection, adapt mesh and time step size.

**Two-Dimensional Simulations.** Figure 7 depicts interface positions and velocities at three different times from a two-dimensional simulation. While the crystal grows upward, the convection gets stronger in the beginning. This behavior continues until the crystal grows up to about 3/4 of the total ampoule height. Later, convection decreases until finally the whole domain is filled with solid material. Corresponding meshes are shown in Fig. 8. The adaptive method automatically generates highly refined meshes near the moving interface, while the meshes in the solid part are very coarse, except near the corners of the domain boundary. In order to resolve the convection, the mesh in the liquid part of the domain is refined more. The reason for the high mesh refinement at the interface can be clearly seen in Fig. 9: the enthalpy \( U \), shown in the upper graph, varies from 0 to 1 in a very narrow region, while it distributes almost uniformly inside the solid and liquid phases. The velocity is important only in the liquid, the graph of the modulus \( |\chi V| \) is shown in the middle and the adapted mesh in the lower part of the figure. Mesh and graphs are shown for the same time as in the left part of Figs. 7 and 8.

The influence of the convection on the temperature distribution is shown in Fig. 10. It depicts the temperature gradients along the axis of symmetry of the ampoule for a simulation with and without convection. According to the Stefan condition

\[
[\nabla \vartheta]_s^l = -V_T[u]_s^l,
\]

the temperature gradient jumps across the interface; the size of the jump is given by the product of interface velocity \( V_T \) and latent heat \( [u] = L \). The curve for the simulation with convection corresponds to the result shown in the middle part of Fig. 7. Especially in the important region around the solid-liquid interface (around \( z = 0.058 \)), the gradients of the temperature are larger. On the other hand, the gradients are smaller in the upper part of
Fig. 7. 2-D simulation. Interface and convection in the melt at different times; arrow length corresponds to $|V|$.

Fig. 8. 2-D simulation. Adaptive grids for same times as in Fig. 7.

the melt. The influence of the convection on the temperature gradients is not seen in the furnace simulation and underlines the necessity of the ampoule simulation. This influence is also reflected by the concave parts of the interface in the middle part of Fig. 7.
**Fig. 9.** 2-D simulation. Graph of enthalpy $U$ (top), convection $|\chi V|$ in the melt (middle) and adaptive grid (bottom).

**Fig. 10.** 2-D simulation. Modulus of temperature gradients along the axis of symmetry for simulation with and without convection. Gradients jump across the interface which is located around $z = 0.058$. 
**Three-Dimensional Simulation.** In three space dimensions, simulations are much more costly than in two dimensions. ALBERT allows program development in two dimensions and, with only few program modifications, simulations in three dimensions. Figure 11 presents a triangulation of the three-dimensional ampoule consisting of tetrahedra and the solid-liquid interface together with the trace of the velocity on the center plane at two different time steps from the 3-D simulation.

![Fig. 11. 3-D simulation. Geometry and grid, interface and trace of convection on center plane at two different times](image)

**Acknowledgements.** We would like to thank Dr. M. Bruder (AIM, Heilbronn, Germany) for his collaboration and crystal growth experiments.

**References**

7. Bruder, M., Schwarz, H.-J., Schmitt, R., Maier, H., Möller, M.: Vertical Bridgman growth of Cd1-yZnyTe and characterization of substrates for use in Hg1-
Numerical Simulation and Control of Industrial Crystal Growth Processes

Karl-Heinz Hoffmann¹, Axel Voigt¹, and Michael Metzger²

¹ nanotechnology – crystal growth, research center caesar, Bonn, Germany
² Department of Material Science, Institute VI, University Erlangen-Nürnberg, Germany

Abstract. This project is carried out jointly by the Crystal Growth Laboratory at the Chair of Material Science VI at the University of Erlangen (Prof. G. Müller) and the Chair of Applied Mathematics at the Technical University of Munich (Prof. K.-H. Hoffmann) with the support of the companies Korth Kristalle GmbH and Freiberger Compound Materials GmbH. The goal of the project is the development of proper numerical models for optimisation and control of industrial crystal growth processes. To this end the research group concentrates on the numerical treatment, simulation and optimal control of techniques for growing high quality single and compound semiconductor crystals. The result of this project is the flexible and efficient computer code CrysVUn that is today successfully used to optimise industrial growth processes.

1 Introduction

The growth of high quality bulk single crystals (Si, GaAs, InP) is one of the key processes that determine the yield and the profitability in semiconductor device manufacturing. The electronic industry, which is based on that materials is today the largest industry in the world. In the 1999–2002 international technology roadmap for semiconductors the Semiconductor Industry Association (SIA) predict an annual growth of the industry of approximately 20% and more than 200 billion in sales by the year 2000. The growth is being driven by the surge of wired and wireless information appliances, as well as Internet infrastructure products. Wafer suitability for microelectronic devices is determined by several different measurements that include surface uniformity, lattice perfection, impurity concentration and crystal electrical properties such as resistivity. Increasing device size increases the requirements on the quality of the crystal. For a given concentration of defects, device failures are greater if fewer, more expensive devices are fabricated on the wafer. Crystal quality will therefore be the utmost importance in the coming years as these trends continue to decrease the tolerances on wafer specifications. It is widely accepted that today without an effective attendant process modelling no appropriate quality of the crystal can be achieved. Besides the help in understanding the complex processes in crystal growth the model can also be used in predicting optimal parameters to grow appropriate crystals. There are mainly two techniques used to grow high quality bulk crystals. The main
Fig. 1. Cross sectional view of CZ furnace

Fig. 2. Silicon single crystal being pulled

In the field of compound semiconductors, such as GaAs or InP, the Vertical Gradient Freeze (VGF) method is the favorable one for growing bulk single crystals with low defect density. The set-up of such a VGF-furnace for the growth of GaAs crystals with 3" diameter is depicted in Fig. 3. The semiconductor material GaAs is filled in a crucible. In order to avoid loss of arsenic the crystal material is covered by a layer of liquid boric oxide. Loss of arsenic is undesired because of the danger of pollution and a lack of arsenic in the crystal may generate dislocations that can induce polymorphic
growth – see Amon [1]. To avoid this loss is also the reason for filling the cavity with argon at a pressure above the dissociation pressure of GaAs. To bring in the required power the model-furnace is supplied with eight heating elements, each of them is controlled autonomously by a thermocouple beside the heater on the ceramic pipe. At the beginning, the raw material is molten in this furnace, with the exception of the seed, which is placed in the small cylindrical part of the crucible. Then in a distinct thermal process the temperature field with certain gradients on the crystal is shifted into the axial direction. This leads to a directional growth of the crystal. The quality of the material strongly depends on the growth conditions – i.e. on the resulting thermal stress.

![Diagram](image.png)

**Fig. 3.** Sketch of a Vertical Gradient Freeze (VGF) or set-up for the growth of GaAs crystals with 3” diameter with a furnace consisting of 8 heaters

This article concentrates on the numerical modelling and optimal control of annealing and growth processes of Si and GaAs, which include a large number of physical transport mechanisms, such as convection, mass transport, segregation, interface dynamics, defect formation and radiative heat transfer, which interact in complex ways.
2 Mathematical Models

The governing equations can be derived from the basic principles of conservation of mass, momentum and energy. In the melt we have the Navier–Stokes equation

\[
\frac{\partial \mathbf{u}_m}{\partial t} + (\mathbf{u}_m \cdot \nabla) \mathbf{u}_m - \nu \Delta \mathbf{u}_m + \nabla p = \beta (T_m - T_0) \mathbf{g},
\]

\[
\nabla \cdot \mathbf{u}_m = 0
\]

with \( \mathbf{u}_m \) velocity, \( p \) pressure and \( T_m \) temperature, where we assume the melt being Newtonian and variations of melt density are negligible, except in the calculation of the body force induced by thermal buoyancy, which is the onset of Grashof convection. This simplification is known as Boussinesq approximation.

Heat transfer in the Czochralski apparatus occurs by conduction, convection and radiation. The equations for the temperature are

\[
\rho_{i_0} c_p \left( \frac{\partial T_i}{\partial t} + \mathbf{u}_i \cdot \nabla T_i \right) - \nabla \cdot \left( k_i \nabla T_i \right) = \rho_{i_0} P_i,
\]

where \( \rho_{i_0} \) is the density of the component, \( c_p \) is the heat capacity, \( \mathbf{u}_i \) is the velocity, \( k_i \) is the thermal conductivity and \( P_i \) is a heat source term. \( i \) indicates the different components of the apparatus, like melt, crystal, crucible, heaters, heat shields and insulators. The heat sources \( P_i \) are nonzero only in the heaters. We neglect the influence of a gas flow in our model. Further we assume that almost no radiation is emitted or absorbed by the gas and that radiation is only diffuse. Moreover, we consider that emission, absorption and reflection of radiating waves occur only at the surfaces. The heat balance at the surface relating the heat flux caused by conduction to the surface \( q \), the radiosity and the irradiation then becomes \( q = G(\sigma T^4) \), where \( G = (I - \epsilon K(1 - (1 - \epsilon)K)^{-1}) \epsilon \) is the infinite dimensional equivalent of the so called Gebhart factor. \( \epsilon \) is the emissivity, the operator \( K \) is defined as

\[
K \lambda(r) = \int_{\Gamma} \omega(r, s) \Xi(r, s) \lambda(s) ds, \forall u \in \Gamma,
\]

where \( \Xi(r, s) \) is a visibility factor. This is mainly based on Tiihonen [10]. Analytic results concerning the heat equation with non-local radiation terms can be found in Tiihonen [11] and Metzger [8]. Along the free melt–crystal interface we have mass conservation and a no-slip boundary condition. Due to the rotation of the crystal this no-slip condition is the onset of forced convection by crystal rotation. Because the melt–crystal interface is a thermal free boundary the thermal conditions need to specify the temperature and the position and shape of the boundary, we set \( T_m = T_c = T_{eq} \) at that boundary, where \( T_{eq} \) is the equilibrium melting temperature. The second thermal condition takes account of the heat flux balance between the two phases, the heat
flux in the crystal is greater than in the melt by the amount of latent heat released during solidification, i.e. $k_m \nabla T_m \cdot \mathbf{n} - k_c \nabla T_c \cdot \mathbf{n} = \rho_c o_l (\mathbf{u}_c - \mathbf{u}_f) \cdot \mathbf{n}$, with $l$ the latent heat, $\mathbf{n}$ the normal vector and $\mathbf{u}_f$ the speed of the phase-boundary. The meniscus separating the melt from the ambient gas is another free interface, whose shape is part of the problem unknowns. Here we again make a Boussinesq-type approximation, where we assume that the surface-tension coefficient $\gamma$ equals a constant $\gamma_0$, except in the boundary condition for the tangential stress. The temperature differences at the interface have an influence on the transport of momentum and heat near the interface. The surface-tension gradient resulting from this differences acts like a shear stress on the melt–gas interface and thereby generates a surface flow. This phenomenon is known as Marangoni convection. We assume that the melt and in the VGF-process also the crystal perfectly adheres to the crucible wall and specify a no-slip condition for the velocity which is the onset of forced convection due to rotation of the crucible. At the remaining boundaries we only need to specify a condition for the temperature. At the surfaces of the melt, the crystal, the crucible, the heaters, heat-shields, insulators and the inner surface of the enclosure we take
\[
k_i \nabla T_i \cdot \mathbf{n} = -G(\sigma T_i^4),
\]
where $i$ stands for the different components. We need to describe initial conditions for the temperature $T_i$ in all components of the apparatus, conditions for the velocity $\mathbf{u}_m$ in the melt and an initial domain for melt and crystal.

If the power to the heaters are given, the temperature can be calculated by solving these equations with suitable initial and boundary conditions. This is usually called the forward problem. But in most applications the temperature field for given sources is of minor interest. One typical task of crystal growth is the adjustment of the heating power to achieve a given temperature distribution. Thus, due to the desire for optimisation of the growth process the temperature has to fulfil several constraints, e.g. position and shape of the crystal–melt interface and the temperature difference between seed and top of the melt. From the mathematical point of view it is much more difficult to solve this problem. In general $N$ vertices with indices $i_1, i_2 \ldots i_N$ can be chosen in which the temperatures are desired to fulfil the conditions
\[
T(\mathbf{x}_{i_n}) = \vartheta_n \quad \forall 1 \leq n \leq N.
\]
This is called an inverse problem. We solve (5) by a repetitive scheme based on the weaker formulation
\[
\frac{1}{2} \sum_{n=1}^{N} |T(\mathbf{x}_{i_n}) - \vartheta_n|^2 + \frac{\mu}{2} \sum_{m=1}^{M} |P_m|^2 = \min.
\]
Thereby $\mu$ is the parameter for the standard Tikhonov regularisation and $P_m$ is the power in $M$ heaters. Often one wants to solve (6) such that some
bounds $C_l$ for the temperature gradients can be guaranteed or one has to take into account that there are maximal and minimal admissible heater powers $\Pi_m$ and $\pi_m$ respectively, i.e. for some indices $k_1, k_2 \ldots k_L$ one has to keep the conditions:

$$\|\nabla T(x_{k_1})\|_2 \leq C_l \quad \forall 1 \leq l \leq L,$$

$$\pi_m \leq P_m \leq \Pi_m \quad \forall 1 \leq m \leq M.$$  

Minimising (6) with respect to the constraints (7) and (8) means solving an optimal control problem.

3 Numerical Methods

Often in industrial crystal growth one can assume an axisymmetric furnace, so that we formulate our models in cylindrical coordinates and are able to reduce the complexity of the problem to 2D. Because of the low growth velocities in industrial crystal growth, it is in some cases also possible to make a quasistationary approximation and to drop the time derivatives in the mathematical models, this implies a great reduction of computing time. But there are applications, which have to be treated fully time depended. Our simulation of Czochralski growth of Silicon concentrates on the melt-crystal element and models the full time-dependent problem with free capillary and phase boundaries. The capillary surface is calculated by minimizing the total energy, which include surface, coating and gravitational energy with the constraint of volume conservation.

$$\min_{S \text{ feasible}} E(S) \quad \text{s.t.} \quad V(S) = A,$$

with the surface $S$, the total energy $E$ and the volume $V$. This is done by the Newton-method by making use of the special structure of the iteration matrix [9]. The Navier–Stokes equation is discretized by the finite element ansatz of Bernardi and Rauge [3] with divergence free elements. For the discretization of the heat equation piecewise linear finite elements are used. A front-tracking method is applied in order to solve for the Stefan boundary and the time diskretisation is based on the method of characteristics. The algorithm for one timestep reads as follows

1. Solve the Laplace–Young equation by minimizing the total energy and find a starting domain.
2. Find an initial solution of the transport equations in that domain.
3. From the heat flux difference at the phase boundary calculate the growth speed and a new Stefan boundary.
4. Calculate the crystallized volume $\delta V$.
5. With the new volume $V - \delta V$ calculate a new capillary boundary.
6. Solve the transport equations on the displaced domain.
7. goto 3.
For the simulation of GaAs-growth with the Vertical Gradient Freeze Method the influence of convection in the melt is small and can be approximated by efficient temperature dependent heat conductivities [5]. The discretization of the differential equations of anisotropic heat conduction with temperature dependent heat conductivities, grey body radiation in cavities and liberation of latent heat at the crystal–melt interface leads in the quasistationary case to the system

\[
F(T, T_d) - \sum_{m=1}^{M} h_m P_m = 0, \tag{10}
\]

that we call the furnace equation and have to solve for the temperature vector \( T \). The vector \( T_d \) denotes the Dirichlet values. It is important to notice that calculations here are done on a fixed grid and no front tracking has to be used in order to follow the free phase boundary. This is done in such a way, that the non-differentiability due to the liberation of latent heat is (artificially) smoothed out. The function \( F \) depends on the geometry, on the numerical grid and on the applied discretization. The materials data enter \( F \) as parameters. The entries of the vectors \( h_m \) result from the discretization of the source terms. \( h_m \) is a characteristic vector describing the distribution of the heating power \( P_m \), \( 1 \leq n \leq M \), of the \( m^{th} \) heater into the numerical mesh. \( N \) is the number of heaters which are considered. The discretization is done by the finite volume technique on unstructured triangular grids. The first step is to recognize that due to the furnace equation (10) the temperature field may be seen as a function of the heating powers \( P_m \)

\[
T \rightarrow T(P_1, \ldots, P_M). \tag{11}
\]

Then the solution of (6) can iteratively be obtained with standard Gauß–Newton iterations, i.e. we improve an initial guess for all \( P_m \) by a correction \( \delta P_m \) obtained from the linearized problem

\[
\frac{1}{2} \sum_{n=1}^{N} \left( T_{n} + \sum_{m=1}^{M} \frac{\partial T_{in}}{\partial P_m} \delta P_m - \psi_n \right)^2 + \frac{\mu}{2} \sum_{m=1}^{M} |P_m + \delta P_m|^2 \equiv \min. \tag{12}
\]

The missing derivatives \( \frac{\partial T_{in}}{\partial P_m} \) in (12) can be obtained by implicit differentiation of the furnace equation (10) with respect to \( P_m \). In the case of state constraints of the form (7), we use the method of sequential quadratic programming together with BFGS like updates for the hessian to minimise (6). This algorithm is introduced in the numerical software system \texttt{CrysVUn} [6], [7], which was developed at the Crystal Growth Laboratory in Erlangen. \texttt{CrysVUn} has already successfully been used for the direct modelling of the global heat transfer and temperature distributions in various crystal growth configurations like Vertical Gradient Freeze [2] with good accordance between simulation and experimental results.
4 Results

For the Cz-crystal growth of Silicon we use a control strategy in order to grow crystals with constant diameter. This is done by observing the triple point and adjusting the pulling velocity by

$$v_s^{n+1} = v_s^n + \alpha (t^n - t^{n+1}) v_d^n,$$

(13)

with $v_s^n$ pulling velocity during the n-th time step, $v_d^n$ velocity of the triple point during the n-th time step in the radial direction and $\alpha$ a process parameter. The influence of the convection on the temperature distribution and the shape of the phase boundary is analyzed. Figures 4–7 display the stream function and the temperature distribution in the melt-crystal element.

At $t = 0$ the stationary solution of the transport equations in the initial domain is shown, the other results are the solutions at different time steps. This calculations indicate that there is a strong dependence of the shape of the phase boundary and the temperature distribution in the crystal on the flow field in the melt. The adjusted pulling velocity in order to grow crystals with a constant diameter is shown in Fig. 8. The velocity varies between 1 cm/h and 7 cm/h and is smaller if convection is taken into account. The control strategy used here is similar to the one used in industrial CZ growth processes, where one also observes an oscillation in the pull-velocity.

CrystVuUn can be used very well in order to find an optimised thermal process for the VGF growth of GaAs crystals with the set-up shown in Fig. 3. We are using above mentioned methods to maintain controlling conditions during the VGF growth run. The most important requirements for a good VGF process are nearly constant axial temperature gradients, a flat face.

![Fig. 6. $t = 0, 40$ min, 80 min](image1)

![Fig. 7. $t = 0, 40$ min, 80 min](image2)
boundary and the GaAs melt must not exceed a given temperature limit. In the simulation of the seeding and the final state, the growth velocity was set to $v_{\text{growth}} = 0$. In all other states $v_{\text{growth}} = 2.5 \text{ mm h}^{-1}$ was applied. So the liberation of latent heat is taken into account in the quasi steady state approximation. The applied temperature conditions are given in Fig. 9.

**Fig. 8.** pulling velocity as function of time

**Fig. 9.** Controlling conditions (fixed temperatures $\vartheta_1$–$\vartheta_4$) for an optimized VGF growth process of GaAs in the set-up shown in Fig. 3. The condition for a planar growth interface is simulated by two fixed temperatures $\vartheta_1 = \vartheta_3 = 1511\text{ K}$ (the melting temperature of GaAs). The nodes $x_{i_1}$ and $x_{i_3}$ belonging to $\vartheta_1$ and $\vartheta_3$ are on the same height in the centre and the interior boundary of the crucible. The condition for an axial temperature gradient of $2 \text{ K cm}^{-1}$ in the melt at the growth interface is obtained by a fixed temperature $\vartheta_2 = \vartheta_1 + 2\text{ K}$ in node $x_{i_2}$ located 1 cm above $x_{i_1}$. Finally we choose a condition to prevent the overheating of the GaAs melt. For that purpose the temperature $\vartheta_4$ in node $x_{i_4}$ at the top of the melt is also conditioned (see text). The nodes belonging to $\vartheta_1$–$\vartheta_3$ are shifted in vertical direction in order to find the heating powers for several positions of the interface. For clarity only the heaters $H_2$–$H_5$ are shown, compare Fig. 3. In the simulation all 8 heaters are considered.
For the resulting optimal heater powers for 15 investigated time steps of the growth run see Fig. 10: This time steps correspond to certain equidistant positions of the growth interface on the furnace axis as the growth rate \(v_{\text{growth}} = 2.5 \text{ mm/h}\) is constant. The different graphs are representing the growth states which belong to the corresponding set of the 8 heater powers \(P_m, \ 1 \leq m \leq 8\), represented in Fig. 10 by one set of the 8 symbols for one fixed position of the solid–liquid interface at the furnace axis of symmetry. Only for the position of the seeding \((z = 0.665 \text{ m})\) and for the tail end of the crystal \((z = 0.805 \text{ m})\) we have set the growth velocity to \(v_{\text{growth}} = 0\).

Next, we analyze the shapes of the crystal–melt interfaces. They are plotted in Fig. 11 again for the different considered time steps corresponding to the set of heating powers depicted in Fig. 10. Most of them have a planar shape, but deviations in the vicinity of the periphery of the crystal become apparent. This deviation can be explained by the strongly anisotropic conductivity \(\lambda\) of the pBN crucible.

GaAs bulk crystals are typically annealed before wafering in order to increase the homogeneity of dopant and defect distribution. The crystals are heated to some annealing temperature. Due to economical considerations the heating of the crystal should be as fast as possible. But during heating thermoelastic stress is induced in the crystal and may generate additional dislocations or slip-lines [4], [12]. At high temperature the thermal conductivity of GaAs is much lower than at lower temperatures. Therefore, the crystal may be heated much faster at low temperature than at high temperature assuming...
Especially if we search for an optimal control of the heating problem, it is advantageous to reduce the complexity of the equations. For this purpose we assume that the crystal is quite long compared with its radius and so we may neglect the axial variation of the temperature profile in the crystal. Secondly we neglect the surrounding ceramic pipe, because we set the temperature at the inner wall of the pipe to be known. Together with the axisymmetric assumption we end up with a 1D nonlinear heat transfer equation to be solved in the crystal with radiation boundary conditions. For the state constraints we skip Cauchy’s equations and use the radial temperature differences as an indicator for thermoelastic stress. This strong simplifications are used to get an (sub)optimal heating strategy, the final forward calculations are done with the full model. In Fig. 13 the maximal von Mises stress $S_{\text{max}}$ during the heating of different processes with constant heating rate and the optimal process are shown. While the time needed to reach the annealing temperature with the optimised process corresponds to a heating rate of 1.5 K/u.a. the maximum von Mises Stress $S_{\text{max}}$ is still lower than even reached with the lowest heating rate of 1 K/u.a.
References


Optimal Control of Sublimation Growth of SiC Crystals

Jürgen Sprekels, Olaf Klein, Peter Philip, and Krzysztof Wilmański

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany, URL: http://www.wias-berlin.de,
E-mail: sprekels@wias-berlin.de

Abstract. The project aims at providing numerical tools to control and optimize sublimation growth of SiC bulk single crystals via the Modified Lely Method. It is in cooperation with the experimental group of Dr. Dietmar Siche at the Institute of Crystal Growth in Berlin. In the course of the project the Modified Lely Method is mathematically modeled and numerically simulated. We present a transient model which for the gas phase consists of balance equations for mass, momentum and energy, and reaction-diffusion equations. The model for the solid components takes into account heat transfer via conduction inside the solid materials and via radiation between solid surfaces of cavities. Results of transient numerical simulations of the temperature evolution inside the growth apparatus are depicted, illustrating the paramount influence of radiation at growth temperature.

1 Introduction

During recent years, the interest in silicon carbide (SiC) single crystal growth has been increasing continuously. This is due to the wide range of industrial applications SiC single crystals have found, e.g. in electronic and optoelectronic devices, such as the blue diode, the blue laser and UV-sensors. The robustness of SiC is especially advantageous if devices have to work in a high temperature or intensive radiation environment.

In view of the demands of applications, the growth process has to be optimized with respect to at least three aspects: Minimization of defects in the single crystal structure, such as micropipes, maximization of the size of the crystal, most importantly of its diameter, and maximization of the growth rate, in order to reduce production costs. SiC single crystal growth has been the subject of numerous articles by both experimental and theoretical scientists, confer e.g. [BMH+93], [Kan93], [Lil93], [HHW+95], [Kon95], [Nis95], [PBD+96], [BKP+99], [PAC+99], [RMD+99], [KPSW00].

One of the most successful growth techniques of recent years has been sublimation growth via the Modified Lely Method. Polycrystalline SiC source powder is evaporated inside an intensely heated graphite crucible at temperatures of some 3000 K, thereby creating a gas mixture consisting predominantly of Ar (which is pumped into the crucible as inert gas), Si, Si$_2$C and SiC$_2$. The SiC single crystal is growing from a cooled seed via sublimation due to chemical reactions (cf. Fig. 1).
Owing to the high temperatures inside the growth apparatus, it is virtually impossible to conduct any direct measurements inside the reactor. Hence, it is necessary to develop reliable mathematical models as a foundation for numerical experiments, which help to get a better understanding of the processes inside the growth chamber.

The inner part of the actual reactor exhibits a quite complicated structure which consists of cavities and of several materials, such as different sorts of graphite, SiC in various modifications, and others. The optimal inner structure is not yet known, and it is one goal of the numerical simulations to determine it.

Induction heating induces eddy currents in the outer part of the reactor. The Joule effect then heats the crucible. The heat is transported to the center of the reactor, i.e. to the source powder and to the crystal seed, via radiation through the cavities, and via conduction through the different materials.

One of the most complicated issues in the process is the occurrence of chemical reactions between the different gas species in the gas phase itself, on the source, on the graphite walls, and on the surface of the growing crystal. Clearly, one has to model these chemical reactions in order to simulate the crystal growth, and here time is playing an important role. The stoichiometry
of the gas species is changing with the course of time from the following reasons: Si is leaking out of the porous graphite reaction chamber, C is closing the pores of the walls and covering the source, and the source is sintering and changing its evaporation rate. Thus, the crystal surface, the surface of the source powder and the walls of the reaction chamber constitute several free boundaries.

The distance between the source and the crystal is of great importance. Since it changes permanently during the whole procedure, this is another crucial point where time dependence is involved. In addition, as the temperature field is not constant in the reaction chamber but very significant for the crystal growth, one ought to aim at optimizing the temperature field during the whole process corresponding to the growing crystal.

Furthermore, experiments show that the quality of the crystal strongly depends on the initial conditions. It is therefore necessary to simulate the initial phase of the process, i.e. the time period when the reactor is heated up from room temperature to actual growth conditions. To account for the numerous time-dependent effects one has to carry out transient numerical simulations.

The numerical experiments which we present in Sect. 3 are taken from [KPSW00] and they are based on the transient model in [BKP+99] which is described in Sect. 2.

2 Modelling

2.1 Gas Phase

The following equations for the gas phase have been derived in [BKP+99] using continuous mixture theory, simplifications and material laws for ideal gases. The resulting partial differential equations consist of balance laws for mass, momentum, and energy, and reaction-diffusion equations (cf. [BKP+99, (2.18a–2.18d)]):

\[
\begin{align*}
\text{mass:} & \quad \frac{d\rho_{\text{gas}}}{dt} + \rho_{\text{gas}} \text{div } \mathbf{v}_{\text{gas}} = 0, \quad (1a) \\
\text{momentum:} & \quad \rho_{\text{gas}} \frac{d\mathbf{v}_{\text{gas}}}{dt} + \text{grad } p_{\text{gas}} = \rho_{\text{gas}} \mathbf{g}, \quad (1b) \\
\text{energy:} & \quad \frac{dU_{\text{gas}}}{dt} + \frac{1}{\rho_{\text{gas}}} \text{div } \mathbf{q}_{\text{gas}} + \frac{1}{\rho_{\text{gas}}} p_{\text{gas}} \text{div } \mathbf{v}_{\text{gas}} = r_{\text{gas}}, \quad (1c) \\
\end{align*}
\]

reaction-diffusion equations (one for each gas species \(\alpha\)):

\[
\begin{align*}
\frac{dc^{(\alpha)}}{dt} & - \frac{1}{\rho_{\text{gas}}} \text{div } \left( \rho_{\text{gas}} c^{(\alpha)} \left( D^{(\alpha)} \right)^{-1} \left( \text{grad } p^{(\alpha)} - c^{(\alpha)} \text{grad } p_{\text{gas}} \right) \right) \\
& = \frac{1}{\rho_{\text{gas}}} \rho^{*(\alpha)}, \quad (1d)
\end{align*}
\]
where
\[
\frac{d}{dt} := \frac{\partial}{\partial t} + \mathbf{v}_{\text{gas}} \cdot \nabla, \quad (2a)
\]
\[
p_{\text{gas}} = \sum_{\alpha} p^{(\alpha)}, \quad U_{\text{gas}} = \sum_{\alpha} c^{(\alpha)} U^{(\alpha)}, \quad (2b)
\]
\[
q_{\text{gas}} = -\kappa_{\text{gas}} \nabla T - \sum_{\alpha} \left( \rho_{\text{gas}} c^{(\alpha)} U^{(\alpha)} + p^{(\alpha)} \right) \left( D^{(\alpha)} \right)^{-1} \left( \nabla p^{(\alpha)} - c^{(\alpha)} \nabla p_{\text{gas}} \right). \quad (2c)
\]

In the above equations, the subscript "gas" is used for quantities in the gas mixture, while superscripts \( (\alpha) \) indicate quantities in the gas species \( \alpha \). The meaning of the symbols is as follows:

- \( \rho_{\text{gas}} \) – mass density,
- \( \mathbf{v}_{\text{gas}} \) – local mean velocity of gas molecules,
- \( p_{\text{gas}} \) – total pressure,
- \( p^{(\alpha)} \) – partial pressure,
- \( g \) – gravity,
- \( U_{\text{gas}} \) – total internal energy,
- \( U^{(\alpha)} \) – partial internal energy,
- \( c^{(\alpha)} \) – mass concentration,
- \( q_{\text{gas}} \) – heat flux,
- \( r_{\text{gas}} \) – radiation,
- \( \rho^{*(\alpha)} \) – partial mass source (chemical reactions, phase transitions),
- \( \kappa_{\text{gas}} \) – thermal conductivity,
- \( T \) – absolute temperature.

The theory of ideal gases provides material laws that constitute an additional coupling of equations (1a)–(1d). The material laws read (cf. [BKP+99, (2.14*)]):

\[
p^{(\alpha)} = \rho_{\text{gas}} c^{(\alpha)} \frac{R}{M^{(\alpha)}} T, \quad U^{(\alpha)} = z^{(\alpha)} \frac{R}{M^{(\alpha)}} T, \quad (3a)
\]

\[
\rho^{*(\alpha)} = \sum_{i=1}^{N} \gamma_{i}^{(\alpha)} M^{(\alpha)} \mu_{H} A^{(i)}, \quad (3b)
\]

where \( R \) is the universal gas constant, \( M^{(\alpha)} \) is the molecular weight, \( z^{(\alpha)} = \frac{3}{2} \) for single-, \( z^{(\alpha)} = \frac{5}{2} \) for double-, and \( z^{(\alpha)} = 3 \) for multi-atomic gas molecules, \( N \) is the number of chemical reactions and phase transitions, \( \gamma_{i}^{(\alpha)} \) are the stoichiometric coefficients, \( \mu_{H} \) is the hydrogen molecular weight, and \( A^{(i)} \) are rates of chemical reactions or phase transitions, respectively.

**2.2 Solid Components**

Apart from interface and boundary conditions, the model in [BKP+99] is completed by writing the energy balances for the different solid components.
of the growth apparatus (cf. [BKP+99, (2.20)]):

\[ \rho^{[\beta]} c_s^{[\beta]} \frac{\partial T}{\partial t} + \text{div} \mathbf{q}^{[\beta]} = \rho^{[\beta]} \mu^{[\beta]}, \] (4)

where superscripts \([\beta]\) refer to quantities in the solid component \(\beta\), \(\rho^{[\beta]}\) denotes mass density, \(c_s^{[\beta]}\) denotes specific heat (as measured in physical experiments, in contrast to [BKP+99, (2.20)]), \(\mathbf{q}^{[\beta]}\) denotes heat flux and \(\mu^{[\beta]}\) is a heat source corresponding to induction heating. The law of the heat flux is given by

\[ \mathbf{q}^{[\beta]} = -\kappa^{[\beta]} \text{grad} \, T, \] (5)

\(\kappa^{[\beta]}\) denoting thermal conductivity. Suitable interface conditions couple equations (4) to the corresponding equations of the gas phase.

### 2.3 Radiation

Radiation is modeled by introducing additional terms into the interface conditions between solid and gas, while any direct interaction between radiation and gas is neglected. All solid parts of the growth apparatus except the SiC single crystal are supposed to be opaque. The SiC single crystal is treated as semi-transparent using the energy-band model. According to this model, the spectrum of wavelengths decomposes into two bands such that the crystal is opaque for radiation from the one band, while it does not interact with radiation from the other band. We will restrict ourselves to a description of the opaque case and refer to [DNR+90] and [KPSW00] for more details on the semi-transparent case. Moreover, it is assumed that reflection and emission depend neither on the wavelength nor on the angle of incidence, i.e. radiation is supposed to be diffuse-grey.

In order to determine the radiation related terms in the gas-solid interface conditions, we start by computing the areal power density \(S(x)\) of the total outgoing radiation at a point \(x\) on the surface of a cavity. It is the sum of the respective areal power densities \(E(x)\) and \(R(x)\) of the emitted and reflected radiation at \(x\),

\[ S(x) = E(x) + R(x). \] (6)

\(E(x)\) and \(R(x)\) can be calculated using the Stefan-Boltzmann law and Kirchhoff's law, resulting in

\[ E(x) = \sigma \varepsilon (x, T(x)) T^4(x), \] (7a)

\[ R(x) = (1 - \varepsilon(x, T(x))) J(S)(x), \] (7b)

where \(\sigma = 5.670 \cdot 10^{-8} \frac{W}{m^2 K^4}\) denotes the Boltzmann radiation constant, \(\varepsilon\) is the emissivity of the surface and \(J\) denotes the areal density of the power of the incoming radiation.
The hypothesis of diffuseness enables us to compute $\mathcal{J}$, using the integral operator defined by

$$\mathcal{J}(S)(x) = \int_{\Gamma} A(x, y) \omega(x, y) S(y) \, dy,$$

where $\Gamma$ denotes the surface of the considered cavity (e.g. consisting of $\Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ for the gas region in Fig. 1), $A$ is called visibility factor and $\omega$ is called view factor. $\Lambda(x, y)$ equals 1 or 0 depending on the surface points $x$ and $y$ being mutually visible or not. $\omega$ is given by the formula

$$\omega(x, y) = \frac{(n_{\text{gas}}(y) \cdot (x - y)) \cdot (n_{\text{gas}}(x) \cdot (y - x))}{\pi ((y - x) \cdot (y - x))^2},$$

where $n_{\text{gas}}$ is the unit normal vector on $\Gamma$ pointing from gas to solid.

Using the preceding equations of this section, one finds that $S$ can be determined from the non-local equation

$$S(x) - (1 - \varepsilon(x, T(x))) \mathcal{J}(S)(x) = \sigma \varepsilon(x, T(x)) T^4(x).$$

Finally, the interface condition including radiation between the solid $\beta$ and the gas phase reads

$$q_{\text{gas}} \cdot n_{\text{gas}} - S + \mathcal{J}(S) = q^{[\beta]} \cdot n_{\text{gas}},$$

assuming that the respective locations of all solid components of the growth apparatus do not change with time.

### 3 Numerical Simulation of the Heat Transport

#### 3.1 Setting

The current stage of our work does not comprehend a solution of either mass, momentum or reaction-diffusion equations, the quantities $c^{(a)}$, $v_{\text{gas}}$ and $\rho_{\text{gas}}$ are treated as being given. While for lower temperatures the gas phase consists solely of inert gas, for temperatures above 2500 K a substantial part of the gas mixture is made up of Si, Si$_2$C and SiC$_2$ (cf. [ABEP98, Fig. 10]). However, the concentration data in [ABEP98] are global and cannot be used to provide local, temperature dependent concentrations. This can only be achieved with sufficient accuracy by solving the reaction-diffusion equations. Since at temperatures greater than 2500 K, radiation is the dominant mechanism of heat transport, we think that using $c^{(\text{Ar})} = 1$ is justified as long as the concentrations are treated as given. In consequence, (2c) is simplified to

$$q_{\text{gas}} = -\kappa_{\text{gas}} \text{grad} T.$$ 

All components of the growth apparatus, as well as all distributions of relevant physical quantities, are supposed to be cylindrically symmetric. As
in [BKP+99], we assume that \( \rho_{\text{gas}} \) merely depends on the vertical coordinate \( z \), and for simplicity we restrict to the case without convection, i.e. \( \mathbf{v}_{\text{gas}} = 0 \). Confer [KPSW00] for transient simulations including convection. As stated in Sect. 2.3, no direct interaction between radiation and gas is accounted for, i.e. \( r_{\text{gas}} = 0 \). By using these simplifications together with the right-hand part of (3a), one can transform (1c) into

\[
\rho^{(\text{Ar})} z^{(\text{Ar})} \frac{R}{M^{(\text{Ar})}} \frac{\partial T}{\partial t} - \text{div} \left( \kappa^{(\text{Ar})} \text{grad} T \right) = 0. \tag{13}
\]

Temperature is assumed to be continuous in every part of the growth apparatus. Heat flux is presumed to be continuous at interfaces between different solid components except if one component is the SiC crystal, where a discontinuity occurs due to semi-transparency. Between solid materials and gas cavities, (4) and (13) are coupled by (11).

On outer boundaries we impose boundary conditions by

\[
q^{[\beta]} \cdot \mathbf{n}_\beta = \sigma \left( \varepsilon^{[\beta]}(T) T^4 - \varepsilon T_{\text{room}}^4 \right), \tag{14}
\]

\( T_{\text{room}} \) denoting the average temperature of the surroundings of the growth apparatus, emitting radiation with an average emissivity \( \varepsilon \). We also use artificial edges, such as \( \Gamma_1 \) and \( \Gamma_b \), depicted in Fig. 2, which radiate at room temperature, but do not absorb or reflect radiation.

We consider a geometric set-up where the seed and the source are inside two different chambers of the growth apparatus, separated by a porous graphite wall, as portrayed in Fig. 2. In this set-up the gas molecules diffuse through the pores of the graphite wall into the growth chamber, where the single crystal seed is typically placed at the bottom (cf. [Kon95]).

### 3.2 Discretization

Equations (4) and (13), as well as the corresponding interface and boundary conditions, are discretized in time using an Euler scheme, which is fully implicit, except for the approximation of emissivities, where the temperature of the previous time step is used. A finite volume scheme is then used for spatial discretization on an unstructured mesh, following the treatment in [Fuh97]. For a description of the methods employed for the computation of discretized radiation terms, visibility and view factors, we refer to [KPSW00, Sect. 3.2].

The result of the discretization procedure is a non-linear system of equations, involving a non-local coupling between the vertices on boundaries of radiation regions. This system is then solved by a Newton–Krylov method.

The numerical algorithms that have been used for the simulations of this article have been implemented in the framework of the PDELIB program package, being developed at the Weierstrass Institute of Applied Analysis and Stochastics, Berlin (cf. [FKL98]).
3.3 Results

Since the material data can vary considerably even between two different growth cycles occurring in the same growth apparatus, it is not feasible to exactly simulate one specific run of a growth experiment. We therefore aim at simulating an idealized growth apparatus, using typical material data where available. The material data that have been used during the following numerical experiments are listed in [KPSW00, App. A].

We discuss four numerical experiments demonstrating the respective influence of radiative and convective heat transfer inside a growth apparatus of height 0.25 m and radius 0.15 m. Each experiment starts at 293 K. The system is then heated inside the oblong area labeled with "induction heating" in Fig. 2 using a constant heating power of 10 kW. Figure 3 shows three temporal snapshots of the temperature distribution evolution of each experiment, the first at 300 s, the second at 3000 s and the third at 30000 s, when the simulations have reached a quasi-stationary state. Each column belongs to the same experiment.

The two simulations depicted in Fig. 3 differ in their treatment of radiative heat transport. Unlike the simulation belonging to the left-hand column, the
Fig. 3. Left-hand column: simulation with radiation. Right-hand column: simulation without radiation. Temperature difference between neighboring isolines is 20 K.
simulation corresponding to the right-hand column does not take into account radiative heat transfer between surfaces of gas cavities and likewise for the two blind holes. For the blind holes, the right-hand simulation employs additional emitting boundary conditions on the edges labeled $I_{\text{em}}$ in the upper picture.

The minimal temperature $T_{\text{min}}$ does always occur at the outside of the outer insulation layer of the apparatus, where the isolines become extremely dense at higher temperatures, appearing as continuously colored areas in the last two stages of Fig. 3. $T_{\text{max}}$ does always occur inside the rectangular heating region. Since the temperature difference between neighboring isolines is 20 K, the pictures allow to determine the temperature in most of the apparatus.

Comparing the two pictures in Fig. 3, one finds that the physical expectation of radiative heat transport dominating under growth conditions is confirmed drastically by the simulations. Without radiative heat transfer between solid surfaces of gas regions, one finds large temperature gradients inside the cavity where the temperature distribution is actually almost uniform. The error in the absolute temperatures is also considerable, especially close to the single crystal surface, where temperature differences of several 100 K occur between the two simulations.

The numerical results thus demonstrate that it is of paramount importance to simulate radiative heat transport in order to achieve a physically acceptable outcome.

4 Conclusions and Prospects

The numerical experiments of the preceding section agree with physical expectations, supporting the reliability of the presented numerical model and of the employed numerical algorithms, which at the current stage provide an effective tool to test different geometrical set-ups with respect to their influence on the resulting temperature evolution inside the growth apparatus. Including radiative heat transport through cavities was an essential step to achieve results of satisfactory accuracy.

So far, our simulations assumed that the heat source is distributed uniformly over some domain, where the induction coils couple to the growth apparatus (cf. Fig. 2). To improve the accuracy of the simulation of the induction heating, it is desirable to determine the distribution of the heat source by numerically solving Maxwell’s equations. The implementation of code using a cylindrically symmetric potential ansatz in two dimensions is currently under way.

In a possible extension of the project it is planned to use simple models for the temperature and concentration dependent growth of the crystal and the evaporation of the source powder to include these phenomena in the simulations. At the same time, it would be very interesting to calculate the stress distribution in the growing crystal, since the stress is supposedly a major driving force in the development of crystal defects.
The next steps would then include the solution of first the mass and momentum balance equations and later the reaction-diffusion equations. Before these tasks can be successfully approached, it is necessary to set up appropriate boundary conditions and to identify the relevant material functions and parameters in further close cooperation with experimentalists.

References


Mathematical Modelling and Numerical Simulation of Semiconductor Detectors

H. Gajewski¹, H.-Chr. Kaiser¹, H. Langmach¹, R. Nürnberg¹, and R. H. Richter²

¹ Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany
² MPI Halbleiterlabor, Otto Hahn Ring 6, 81739 München, Germany

Das diesem Artikel zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 03-GA7FV1-2 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Abstract. We report on a system of nonlinear partial differential equations describing signal conversion and amplification in semiconductor detectors. We explain the main ideas governing the numerical treatment of this system as they are implemented in our code WIAS-TeSCA. This software package has been used by the MPI Semiconductor Laboratory for numerical simulation of innovative radiation detectors. We present some simulation results focussing on three-dimensional effects in X-ray detectors for satellite missions.

1 Introduction

For observation of X-ray radiation in astrophysics and other applications semiconductor detectors play a rapidly growing role. Devices as the DEPFET (DEPleted Field Effect Transistor), in particular the MOS-type (DEPMOS) developed at the MPI Semiconductor Laboratory and the Max Planck Institutes for Physics and for Extraterrestrial Physics [9], offer very important advantages when compared to conventional detectors. These advantages are due to the intrinsic feature of the structure which at once has detector and amplifier properties. High amplification and low noise can be obtained and the charge generated by the signal can be read out at the place of origin, therefore avoiding problems connected with charge transport.

For development and optimization of such refined semiconductor detectors mathematical modelling and numerical simulation is indispensable. At the MPI Semiconductor Laboratory the simulation program WIAS-TeSCA is used to design detectors and to anticipate their qualitative and quantitative behavior. Device simulations on a prototype design of DEPFET confirm that the device is intrinsically fast and that it will function properly.

It is the aim of this article to describe the mathematical model for semiconductor detectors and the main ideas underlying our code for numerically
solving the model equations reliably and efficiently. Finally we present some simulation results concerning detector structures of current interest.

2 The Drift-Diffusion Model

Charge generation by radiation and charge transport in semiconductor detectors can be described adequately by the drift-diffusion model, consisting of a Poisson equation for the electrostatic potential $\varphi$:

$$-\nabla \cdot (\epsilon \nabla \varphi) = d + u_2 - u_1 \quad \text{in } \Omega ,$$

(1)

and continuity equations for the charge carriers, electrons $u_1$ and holes $u_2$:

$$\frac{\partial u_i}{\partial t} - \nabla \cdot \left( \mu_i \left( \nabla u_i + q_i u_i \nabla \varphi \right) \right) = g - r \quad \text{in } Q .$$

(2)

Here $\Omega \subset \mathbb{R}^n$, $n \leq 3$, is the Lipschitzian domain, occupied by the detector and $Q = (0, T) \times \Omega$ is a time cylinder. The source term $g = g(t, x)$, $(t, x) \in \Omega$, models charge generation by radiation and the remaining symbols have the following meaning:

- $\epsilon = \epsilon(x)$ dielectric permittivity,
- $d = d(x)$ density of impurities,
- $\mu_i = \mu_i(x, |\nabla \phi_i|)$ mobility,
- $\phi_i = \varphi + q_i \log u_i$ electro-chemical potential,
- $q_1 = -1$, $q_2 = +1$ sign of charges,
- $r = r_0(u)(u_1 u_2 - 1)$ recombination rate,

where the index $i = 1, 2$ refers to electrons and holes respectively, and $u = (u_1, u_2)$ is the vector of charge densities. This system has to be completed by initial conditions

$$u_i(0, x) = u_{i0}(x) \geq 0, \quad x \in \Omega, \quad i = 1, 2 ,$$

(3)

and boundary conditions

$$\varphi = \varphi_\Gamma, \quad u_i = u_{i\Gamma} \quad \text{on } (0, T) \times \Gamma_D, \quad i = 1, 2,$$

$$\nu \cdot \nabla \varphi = \nu \cdot \nabla \phi_i = 0 \quad \text{on } (0, T) \times \Gamma_N, \quad i = 1, 2 ,$$

(4)

where $\Gamma = \partial \Omega = \Gamma_D \cup \Gamma_N$ and $\nu$ is the outer unit normal on $\Gamma$.

The drift-diffusion model was derived by van Roosbroeck [13] in 1950 and is now generally accepted. The first significant report on using numerical techniques to solve these equations for carriers in an operating semiconductor device structure has been published by Gummel [7] in 1964. Since then the numerical modelling of semiconductor devices proved to be a powerful tool for device designers.
First mathematical papers devoted to the drift-diffusion equations of carrier transport in semiconductors appeared at the beginning of the seventies [10,11]. Global existence and uniqueness of solutions under realistic physical and geometrical conditions was proved firstly in [5].

The key for proving these results and also for establishing stable numerical solving procedures is the existence of a Lyapunov function for (1–2), provided

\[ u_i r = \exp(-q_i \varphi r), \quad g = 0. \]

Indeed, under these assumptions the free energy

\[ \mathcal{F}(\varphi, u) = \int_\Omega \left[ \frac{\varepsilon}{2} |\nabla(\varphi - \varphi^*)|^2 + \sum_{i=1}^2 u_i (\log \frac{u_i}{u_i^*} - 1) \right] \, dx \]

satisfies

\[ \frac{d\mathcal{F}}{dt} = -\int_\Omega \left[ \sum_{i=1}^2 \mu_i u_i |\nabla \phi_i|^2 + (\phi_2 - \phi_1) r \right] \, dx \leq 0, \tag{5} \]

where the thermal equilibrium potential \( \varphi^* \) is solution of the nonlinear Poisson equation

\[-\nabla \cdot (\varepsilon \nabla \varphi^*) = d + \exp(-\varphi^*) - \exp(\varphi^*) \quad \text{in} \; \Omega, \]

\[ \varphi^* = \varphi r \quad \text{on} \; \Gamma_D, \quad \nu \cdot \nabla \varphi^* = 0 \quad \text{on} \; \Gamma_N \]

and the equilibrium charge densities are given by

\[ u_i^* = \exp(-q_i \varphi^*). \]

In particular, (5) implies exponential decay of the solution \((\varphi, u)\) to (1–4) to the thermal equilibrium \((\varphi^*, u^*)\).

3 Numerics for the Drift-Diffusion Model

3.1 Calculation of Contact Currents

In view of real detector structures we suppose that

\[ \Gamma_D = \bigcup_{j=1}^{k_0} K_j. \]

An essential task of device simulation in general and of detector simulation in particular, is accurate calculation of the currents \( J_{K_j} \) through the device contacts \( K_j \). Since in general, by analytical reasons, only weakly regular solutions can be expected, we calculate the \( J_{K_j} \)'s in accordance with the
definition of weak solutions. To this end, we provide suitable test functions $h_j$ as solutions of the boundary value problems

$$\Delta h_j = 0 \quad \text{in } \Omega, \quad \nu \cdot \nabla h_j = 0 \quad \text{on } \Gamma_N, \quad \delta_{jk} \quad \text{on } K_k,$$

such that, evidently,

$$\sum_{j=1}^{k_0} h_j \equiv 1 \quad \text{in } \Omega \ . \quad (6)$$

Now, differentiating (1) with respect to time and adding to (2), we get the conservation of total current

$$\nabla \cdot J = 0, \quad J = e\nabla \frac{\partial \varphi}{\partial t} + \sum_{i=1}^{2} \mu_i \left( q_i \nabla u_i + u_i \nabla \varphi \right) \ .$$

($J$ sums up the dielectric displacement current, the electron and hole current.) Hence, the Gauss theorem implies

$$J_{K_j} = \int_{K_j} J \cdot \nu \, ds = \int_{\Omega} J \cdot \nabla h_j \, dx \ ,$$

and as a consequence of (6)

$$\sum_{j=1}^{k_0} J_{K_j} = 0 \ .$$

### 3.2 Time Discretization

For time discretization we apply Euler’s implicit method and we regard a partition of the time interval:

$$[0, T] = \cup_j S_j, \quad S_j = [t_{j-1}, t_j], \quad \tau_j := t_j - t_{j-1} > 0, \quad t_0 = 0 \ .$$

Then our time discrete version of the drift-diffusion system reads

$$-\nabla \cdot (\varepsilon \nabla \varphi^j) = d + \frac{u_i^j - u_i^{j-1}}{\tau_j} - \nabla \cdot \left( \mu_i^{j-1} \left( \nabla u_i^j + q_i u_i^j \nabla \varphi^j \right) \right) = g^j - r^j \ , \quad (7)$$

$$u_i^0 = u_{0i} \ ,$$

where $j = 1, 2, \ldots$ indexes the discrete times, and $i = 1, 2$ indicates the species.

An essential feature of this discretization is that, as in the continuous case, the free energy is Lyapunov function. More precisely, there is the following
estimate [4]:

\[ \mathcal{F}(\varphi^j, u^j) - \mathcal{F}(\varphi^{j-1}, u^{j-1}) \]

\[ \leq - \int_{\Omega} \left[ \sum_{i=1}^{2} u^{j-1}_i u^j_i |\nabla \varphi^j_i|^2 + (\varphi^j_i - \varphi^{j-1}_i) r^j \right] \, dx \leq 0. \]

### 3.3 Space Discretization

For space discretization we use \( n \)-dimensional simplex elements \( E_l \) such that

\[ \Omega = \bigcup_l E_l. \]

By \( E \) we denote the set of all edges \( e_{kl} = x_k - x_l \) connecting vertices \( x_k \) and \( x_l \) of our triangulation. Let

\[ V_k = \{ x \in \mathbb{R}^n : \| x - x_k \| \leq \| x - x_j \| \text{ for all vertices } x_j \in \Omega \} \]

be the Voronoï cell assigned to vertex \( x_k \) and \( \partial V_k \) its surface. Our main hypothesis with respect to space discretization is that the electron and hole current \( J_i = q_i \mu_i \left( \nabla u_i + q_i u_i \nabla \varphi \right), \ i = 1, 2 \), have to be constant along simplex edges \( e_{kl} \):

\[ J_{ikl}(s) = \mu_{ikl} q_i \left( \nabla u_i + q_i u_i \nabla \varphi \right)(s) = \text{ const. for all } s \in e_{kl}. \]

Replacing \( \nabla \) by \( d/ds \), we get ordinary differential equations with respect to the edge parameter \( s \). After integration we can express \( J_{ikl} \) in terms of the vertex values \( \varphi_k = \varphi(x_k) \) and \( u_{ik} = u_i(x_k) \):

\[ J_{ikl} = \frac{\mu_{ikl} q_i}{|e_{kl}|} \left( b(q_i(\varphi_k - \varphi_l)) u_{il} - b(-q_i(\varphi_k - \varphi_l)) u_{ik} \right), \]

where

\[ b(s) = \frac{s}{\exp(s) - 1} \]

is the Bernoulli function. In order to derive a space discrete version of the drift-diffusion system, we test (1) and (2) with the characteristic function \( \chi_{V_k} \) of the Voronoï cell \( V_k \). Thus, applying the Gauss theorem, we arrive at

\[ \epsilon_k \sum_{\{l: e_{kl} \in E\}} \frac{\varphi_k - \varphi_l}{|e_{kl}|} \left| \partial V_k \cap \partial V_l \right| = (d_k + u_{2k} - u_{1k}) |V_k|, \]

\[ \frac{\partial u_{ik}}{\partial t} |V_k| - q_i \sum_{\{l: e_{kl} \in E\}} J_{ikl} \left| \partial V_k \cap \partial V_l \right| = (g_k - r_k) |V_k|, \]

for all vertices \( x_k \) of our triangulation.
As the original drift-diffusion equations (1–2) and its time discrete variant (7), also the space discrete version (8) has the free energy as Lyapunov function. Indeed, there is the following estimate [4]:

\[
\frac{dF}{dt} \leq - \sum_{e_{kl} \in E} \left[ \sum_{i=1}^{2} u_{ik} \mu_{ikl} \frac{\partial V_k \cap \partial V_i}{|e_{kl}|} \frac{b(\varphi_i - \varphi_k)}{b(\phi_{il} - \phi_{ik})} |\phi_{ik} - \phi_{il}|^2 
+ \left| V_k (\phi_{2k} - \phi_{1k}) r_k \right| \right] \leq 0 .
\]

Now it is straightforward to combine time and space discretization in such a way that the free energy remains Lyapunov function of the fully discretized drift-diffusion system as it is implemented in our numerical code WIAS–TeSCA.

### 3.4 Decoupling, Linearization

A natural way of decoupling the drift-diffusion equations is due to Gummel [7]. It starts from the state equations

\[
u_i = \exp \left( q_i(\phi_i - \varphi) \right) .
\]  

(9)

Assuming the electro-chemical potentials \( \phi_i \) to be known from a preceding iteration, equation (1) can be seen as a nonlinear elliptic equation with monotone nonlinearity with respect to the electrostatic potential \( \varphi \) and can be solved via a globally convergent Newton procedure. In the second step we insert \( \varphi \) into the continuity equations and solve them for \( u_i \). Finally, from \( \varphi \) and \( u_i \) we update the electro-chemical potentials \( \phi_i \) via (9).

Roughly speaking, Gummel’s iteration method converges rapidly far away and slowly near the solution. Thus, we could combine it advantageously with Newton’s method having a complementary behavior.

### 3.5 Solution of Linear Algebraic Equations

After space and time discretization, decoupling and linearization we end up with sparse linear equation systems. The Poisson equation is symmetric and can be solved without any problems by conjugate gradient methods. The continuity equations, due to the (in general) strong drift term \( \nabla \varphi \), are only structurally symmetric and very stiff. Hence, we are often obliged to go back to direct solution procedures. Fortunately, we have available highly efficient factorization processes with complete super-node pivoting [12].

### 4 Software Design

Originally WIAS–TeSCA [3] is a two-dimensional semiconductor device simulation package written in FORTRAN. The present redesign of WIAS–TeSCA
Mathematical Modelling . . . of Semiconductor Detectors

aims at a tool which operates on two and three dimensional spatial domains. WIAS–TeSCA shall deal with a variable set of coupled model equations, which are defined and have to be solved on specific parts of the simulation domain, while referring to and communicating via the underlying discretization of the whole simulation domain. The management of material properties also works within this concept.

In the sense of structured programming we adopt an object oriented approach [6]. The implementation is made in C, while using numerical kernels in FORTRAN. All features which might impede portability, such as special C++ constructions are omitted. We reuse pdelib-components [1], in particular the interface for grid management and the interface for linear algebra. Moreover, we make use of the interactive graphics package gltools [2], and there is an interface to the extension language lua [8] for the description of the simulation problem.

In the following we describe the calculation of the electrostatic potential in a doped semiconductor hetero-structured detector. To that end a nonlinear Poisson equation with prescribed spatially varying quasi Fermi potentials of the carrier densities has to be solved on a fully three dimensional spatial domain.

5 Simulation Results

The structure and operation principle of the MOS-type DEPFET is shown in Fig. 1. It is based on the sideward depletion as used in the semiconductor drift chamber and the field effect transistor. The transistor is located atop a low doped n-type semiconductor substrate. It becomes fully depleted by applying a sufficiently high negative voltage to the backside p⁺ contact. By suitable doping, a potential minimum for electrons is formed below the transistor channel. The fully depleted bulk is the sensitive volume of the detector in which electron–hole pairs created by the incident radiation are separated by the electric field. While the holes are moved to the negatively biased back-contact, the electrons are collected in the local potential minimum below the channel of the transistor (the so called “internal gate”) and thus, increase its charge density by induction. As a consequence, the transistor current is increased as long as the signal charge is not removed from the internal gate. The removal of the signal charge (emptying of the internal gate) can be performed in a way, which will be described below. Arranging many DEPFETs over an extended area leads to a pixel detector array with each single DEPFET providing one pixel.

Our two- and three-dimensional device simulations of pixel cells with WIAS–TeSCA concern the response of the device to radiation (the collection mechanism), the functionality of the emptying of the internal gate (the clearing mechanism), and the transfer of charge between two internal gates. The latter mechanism plays a role in detector arrays, when the option of noise reduction by switching and repeated read out is used, see [9].
A perspective view on the electrostatic potential and the carrier densities in two operating states of a single pixel cell is shown in Fig. 2 and Fig. 3. The pixel cell measures 18\,\mu\text{m} \times 15\,\mu\text{m} \times 280\,\mu\text{m} and the figures show a section of the three dimensional simulation domain up to a depth of 12\,\mu\text{m}. For this type of topologies care has to be taken to avoid detrimental effects due to the sideward limits of the structure. In the three dimensional plot of the hole density, see Fig. 3, atop the structure the p\textsuperscript{+} contacts are visible red shaded. The drain (D) stretches over the full width of the device and is connected with the source (S) through the MOS enhancement channel below the gate (G), see Fig. 1. In the three dimensional plot of the electron density, see Fig. 2, atop the structure the n\textsuperscript{+} clear contact (Cl) is visible red shaded. It is separated from the source by the clear gate, see also Fig. 1. The signal charge is stored within the internal gates located below the transistor channel. As the DEPFET transistors are built on detector grade low doped silicon, additional buried n-type doping fairly close to the surface is necessary in order to move the position of the internal gate close to the surface (at a distance smaller than the gate length) and simultaneously prevent the flow of holes from the transistor through the bulk towards the strongly negatively biased backside diode.

Single photon detection of X-rays with a high energy resolution requires the collection of the whole generated signal charge without any losses. Since in the charge collection mode all generated electrons have to drift into the internal gates the still positively biased clear contacts have to be shielded from the sensor region. Fig. 2 shows the build up of a potential barrier beneath the n-doped clear contact by a buried p-doped layer. This layer becomes completely depleted from holes during the clearing phase, where a high positive voltage is applied to the clear contact, see Fig. 3. After turning off the clear voltage the p-doped region remains depleted forming the required potential barrier by the influence of the negative space charge of the acceptor ions. This non steady state can be simulated by applying a locally fixed quasi Fermi potential for holes in the buried p-doped layer.

After each readout the internal gate has to be emptied. Complete clearing of the internal gate avoids noise due to fluctuations in the left-over charge. Clearing of the internal gate is performed by the application of a positive
Fig. 2. DEPMOS pixel cell in collection mode: potential and electron density. Bottom: three dimensional plots. Top: level plots at a depth of 500 nm

Fig. 3. DEPMOS pixel cell in clearing mode: potential and hole density. Bottom: three dimensional plots. Top: level plots at a depth of 200 nm

voltage pulse of approximately 12V to the clear contact. Fig. 3 displays a potential barrier between the internal gate and the n$^+$ clear contact. Such a barrier hinders the clearing mechanism and its removal was a task for further
optimization of the device. Proper adjustment of device layout, doping profiles and applied voltages leads to detector structures which are intrinsically very fast such that the read out and clearing speed will rather be limited by effects in the signal routing.

Acknowledgments. We thank Gerd Reinhardt (WIAS Berlin) for his support with the graphics.

References


Optimal Design of High Power Electronic Devices by Topology Optimization

Ronald H.W. Hoppe¹, Peter Böhm², George Mazurkevitch¹, Svetozara Petrova¹,⁴, Gerhard Wachutka², and Eckhard Wolfgang³

¹ Institute of Mathematics, University of Augsburg, 86159 Augsburg, Germany
² Physics of Electrotechn., Munich University of Technology, 80290 Munich, Germany
³ Siemens AG, Dep. ZT MS 4, 81739 Munich, Germany
⁴ on leave from the Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. High power electronic devices such as converter modules are frequently used as electric drives for high power electromotors. The efficient and reliable operating behaviour of such devices requires an optimal design with regard to a minimization of power losses due to parasitic inductivities caused by eddy currents. The mathematical modelling gives rise to a topology optimization problem where the state variables are required to satisfy the quasistationary limit of Maxwell’s equations and the design variables are subject to both equality and inequality constraints. Based on appropriate finite element approximations involving domain decomposition techniques, the discretized optimization problem is solved by a primal-dual Newton interior-point method.

1 Introduction

The design of innovative high power electronic devices and systems based on the pulse width modulation technique has become an industrially relevant issue in recent years due to the wide range of applications. In particular, pulsed DC-AC converter modules are used both for energy generation and/or transmission and as electric drives for high power electromotors in public transportation systems such as trams and, as the most spectacular example, in high speed trains (cf. Fig. 1). Such converter modules consist of modern fast switching semiconductor devices, e.g. IGBTs (Insulated Gate Bipolar Transistors) or GTOs (Gate Turn-Off Thyristors) interconnected and linked with the power source and load by copper made bus bars (see Fig. 2). Due to the use of the IGBTs or GTOs, switching times of less than 100 ns and switched currents of one up to five kA can be realized. However, as a side-effect of the fast switching times and steep current ramps, eddy currents build up inside the bus bars that lead to parasitic inductivities causing possible overvoltages and significant power losses (cf. [9,11]). Therefore, one of the prime objectives of the electrical engineers is to design the bus bars in such a way that the total inductivity is minimized. It is known that the geometrical shape and topology of the bus bars play a prominent role in so far as they have a significant impact on the distribution and size of the generated eddy...
Fig. 1. Applications of high power converter modules: Energy generation (left), public transportation (top right), high speed trains (bottom right)

currents. Consequently, the task is to distribute the material in an optimal way. From a mathematical point of view, the problem can be stated as a topology optimization problem with constraints on the state and design variables. Here, the state variables are the generated electromagnetic fields, or associated potentials, and the design variable is chosen as the conductivity of the material. In this paper, we will present a primal-dual Newton-type interior-point method for the solution of the topology optimization problem (Sect. 4). Since this algorithm requires the frequent solution of the underlying field equations, we need advanced numerical solution methods for their efficient computation. In particular, we will discuss in some detail multilevel and domain decomposition techniques using adaptive curl-conforming edge
element discretizations (Sect. 3). These are applied to potential equations for the scalar electric potential and the magnetic vector potential which can be derived from the quasi-stationary limit of Maxwell’s equations as shown in Sect. 2. Finally, in Sect. 5 we will document the results of numerical computations that can lead to a considerable reduction of the parasitic inductivities.

2 Parasitic Inductivities in Converter Modules

We consider a converter module consisting of \( N \) bus bars \( \Omega_\nu, 1 \leq \nu \leq N \), and \( M \) IGBTs (Insulated Gate Bipolar Transistors) connecting a power source with the load (cf. Fig. 2 (left)). Each bus bar contains \( N_\nu \) ports \( T_{\nu \alpha}, 1 \leq \alpha \leq N_\nu \), where currents are either supplied or taken off the bar (cf. Fig. 2 (right); the ports are marked by different colours). During operation of the module, eddy currents are generated that can be described by the quasistationary limit of Maxwell’s equations

\[
\frac{\partial \mathbf{B}}{\partial t} + \text{curl} \mathbf{E} = 0, \quad \text{div} \mathbf{B} = 0, \quad \text{curl} \mathbf{H} = \mathbf{J},
\]

\[
\mathbf{B} = \mu \mathbf{H}, \quad \mathbf{J} = \sigma \mathbf{E}.
\]

Here, \( \mathbf{E} \) and \( \mathbf{H} \) denote the electric and the magnetic field, \( \mathbf{B} \) and \( \mathbf{J} \) stand for the magnetic induction and the current density, \( \mu \) is the magnetic permeability and \( \sigma \) refers to the electric conductivity (cf. [1] for a justification of the quasistationary limit in the computation of eddy currents).

We use a potential formulation by introducing a scalar electric potential \( \varphi \) and a magnetic vector potential \( \mathbf{A} \) according to

\[
\mathbf{E} = -\text{grad} \varphi - \frac{\partial \mathbf{A}}{\partial t}, \quad \mathbf{B} = \text{curl} \mathbf{A}
\]

Fig. 2. Converter Module (left); Geometry of a bus bar (right)
(cf., e.g., [6]). For the electromagnetic potentials \( \varphi \) and \( \mathbf{A} \), from (1),(2) we obtain a coupled system of PDEs consisting of an elliptic boundary value problem

\[
\text{div} (\sigma \, \text{grad} \, \varphi) = 0 \quad \text{in} \quad \Omega, \quad (3)
\]

\[
\sigma \, \mathbf{n} \cdot \text{grad} \, \varphi = \begin{cases} -I_{\nu \alpha}(t) & \text{on } T_{\nu \alpha} \\ 0 & \text{else} \end{cases} \quad (4)
\]

with \( \sum_{T_{\nu \alpha}} I_{\nu \alpha} = 0 \), and a parabolic PDE

\[
\sigma \frac{\partial \mathbf{A}}{\partial t} + \text{curl} \, \mu^{-1} \, \text{curl} \, \mathbf{A} = \begin{cases} -\sigma \text{grad} \varphi & \text{in } \Omega \\ 0 & \text{in } \mathbb{R}^3 \setminus \Omega \end{cases} \quad (5)
\]

with appropriate initial and boundary conditions. Note that (5) has to be considered in the interior and exterior domain.

The total inductivity is given by

\[
L_{\text{tot}} := \left( \sum_{\nu, \alpha} \sum_{\mu, \beta} \int_{0}^{T} |L_{\nu \alpha, \mu \beta}(t)|^2 \, dt \right)^{1/2} \quad (6)
\]

with the generalized transient inductivity coefficients

\[
L_{\nu \alpha, \mu \beta}(t) := \sigma^{-1} \int_{\Omega_{\nu}} \mathbf{J}_{\nu \alpha}(x) \cdot \mathbf{S}(t) \mathbf{J}_{\mu \beta}(x) \, dx
\]

where \( \mathbf{J}_{\nu \alpha} \) denotes the current associated with the bar \( \Omega_{\nu} \) at the port \( T_{\nu \alpha} \) and \( \mathbf{S}(\cdot) \) is the solution operator associated with (5).

### 3 Numerical Solution of the Field Equations

Structural optimization algorithms with PDE constraints like the primal-dual Newton interior-point method to be described in the subsequent section require the frequent solution of the PDEs, i.e., in this context the solution of the equations (3),(4) and (5) for the electric potential \( \varphi \) and the magnetic vector potential \( \mathbf{A} \). Therefore, we have to provide efficient numerical tools for their iterative solution. For that purpose we will use adaptive multilevel and domain decomposition methods based on curl-conforming edge element discretizations for the computation of the magnetic vector potential \( \mathbf{A} \) and on nonconforming P1 approximations for the electric potential \( \varphi \).

Discretizing the interior domain problem associated with (5) implicitly in time by the backward Euler scheme, at each time step \( t_m := t_{m-1} + \Delta t \) we are
faced with an elliptic boundary value problem for the double curl-operator whose variational formulation takes the form:

Find $j^m \in V \subset H(curl; \Omega) := \{ q \in L^2(\Omega)^3 \mid \text{curl} \ q \in L^2(\Omega)^3 \}$ such that

$$a(j^m, q) = \ell(q) \quad , \quad q \in V \ , \quad (7)$$

where

$$a(j, q) := \int_{\Omega} [\Delta t \mu^{-1} \text{curl} j \cdot \text{curl} q + \sigma j \cdot q] \ dx \quad , \quad j, q \in V \ ,$$

$$\ell(q) := \int_{\Omega} \sigma [j^{m-1} \cdot q - \Delta t \ \text{grad} \varphi_m \cdot q] \ dx \quad , \quad q \in V .$$

Edge elements, originally due to Whitney, have been systematically introduced into the finite element methodology by Nédélec (cf. [18,19]) and have become an appropriate tool in the computation of electromagnetic fields (see e.g. [7,8,10]).

We consider the lowest order edge elements with respect to a simplicial triangulation $T_h$ of the computational domain $\Omega$

$$Nd_1(K) := \{ q = a + b \wedge x \mid a, b \in \mathbb{R}^3 \} \quad , \quad K \in T_h$$

with the degrees of freedom given by the tangential components with respect to the edges $E_\nu$ of $K$

$$\ell_{E_\nu}(q) := \int_{E_\nu} t_{E_\nu} \cdot q \ d\sigma \quad , \quad 1 \leq \nu \leq 6 .$$

The global edge element space

$$Nd_1(\Omega; T_h) := \{ q : \Omega \to \mathbb{R}^3 \mid q |_K \in Nd_1(K), K \in T_h \}$$

is then a proper subspace of $H(curl; \Omega)$ and we choose $V_h := Nd_1(\Omega; T_h) \cap V$. The problem with the solution of the finite dimensional analogue of (7)

$$a(j_h^m, q_h) = \ell(q_h) \quad , \quad q_h \in V_h \ , \quad (8)$$

is that the curl-operator has a nontrivial kernel which is the subspace of irrotational vector fields. We take advantage of the fact that in the discrete regime this very subspace can be identified with $\text{grad} V_h$ for some $V_h \subset S_1(\Omega; T_h)$, where $S_1(\Omega; T_h)$ is the finite element space associated with the standard conforming P1 approximation. To be more precise, we can set up the following hybrid iterative scheme:

Given some iterate $j_{h,\nu}^m \in V_h , \nu \in N_0$, we first perform some SOR sweeps on (8) resulting in $\hat{j}_{h,\nu}^m$ and consider the defect correction problem on
the subspace of irrotational vector fields

\[ \int_{\Omega} \sigma \text{grad} u_h^m \cdot \text{grad} v_h \, dx \]

\[ = \ell(\text{grad} v_h) - a(\tilde{j}_h^{m,\nu}, \text{grad} v_h) , \quad v_h \in V_h . \] (9)

Performing some SOR sweeps on (9) gives \( \tilde{u}_h^m \), and we obtain \( \tilde{j}_h^{m,\nu+1} := \tilde{j}_h^{m,\nu} + \text{grad} \tilde{u}_h^m \).

In a multilevel framework, having a hierarchy \( (T_k)_{k=0}^{\ell} \) of simplicial triangulations at our disposal, the above hybrid iteration is used both as a smoother on all levels \( 1 \leq k \leq \ell \) and as an iterative solver on the coarsest grid \( k = 0 \) whereas the intergrid transfers are handled in a canonical way (cf. [12] for a detailed analysis and [2] for various applications).

Mortar edge element methods (cf. [4,13]) are appropriate for an FEM-FEM coupling of the interior and exterior domain problems as well as for a domain decomposition approach to the interior domain problem with respect to a nonoverlapping geometrically conforming decomposition

\[ \tilde{\Omega} = \bigcup_{i=1}^{n} \tilde{\Omega}_i , \quad \Omega_i \cap \Omega_j = \emptyset , \quad i \neq j \]

according to the geometrical structure of the bus bars. We refer to the union of the interfaces between adjacent subdomains

\[ S := \bigcup_{i \neq j} \Gamma_{ij} , \quad \Gamma_{ij} = \tilde{\Omega}_i \cap \tilde{\Omega}_j \]

as the skeleton of the decomposition.

We further consider individual simplicial triangulations of the subdomains \( \Omega_i , \ 1 \leq i \leq n \), allowing nonconforming nodal points on the interfaces \( \Gamma_{ij} \subset S \) and discretize the subdomain problems by the lowest order curl-conforming edge elements. In order to guarantee consistency of the overall approximation, we have to impose weak continuity constraints on the skeleton \( S \) which can be realized by appropriately chosen Lagrangian multipliers. We denote by \( V_h(\Omega_i) , \ 1 \leq i \leq n \), the subdomain based edge element spaces and by \( M_h(\Gamma_{ij}) , \ \Gamma_{ij} \subset S \), the local multiplier spaces. We define the product spaces

\[ V_h(\Omega) := \prod_{i=1}^{n} V_h(\Omega_i) , \quad M_h(S) := \prod_{\Gamma_{ij} \subset S} M_h(\Gamma_{ij}) \] (10)

as well as bilinear forms \( a_h : V_h(\Omega) \times V_h(\Omega) \rightarrow \mathbb{R} \)

\[ a_h(u_h, v_h) := \sum_{i=1}^{n} a_{\Omega_i}(u_h, v_h) , \quad u_h, v_h \in V_h(\Omega) , \] (11)
where \(a_{\Omega_i} := a|_{\Omega_i}\), and \(b_h : V_h(\Omega) \times M_h(S) \rightarrow \mathbb{R}\)

\[
b_h(q_h, \mu_h) := \sum_{\Gamma_{ij} \subset S} \int_{\Gamma_{ij}} \mu_h \cdot [n \wedge q_h] |_{\Gamma_{ij}} \, ds \tag{12}\]

for edge element discretizations where \([n \wedge q_h] |_{\Gamma_{ij}}\) denote the jumps of the tangential components of \(q_h\) across the interfaces \(\Gamma_{ij} \subset S\).

The domain decomposition approaches give rise to the discrete saddle point problems:

Find \((u_h, \lambda_h) \in V_h(\Omega) \times M_h(S)\) such that

\[
ah(u_h, v_h) + b_h(v_h, \lambda_h) = \ell(v_h) \quad , \quad v_h \in V_h(\Omega) \quad , \quad \lambda_h \in M_h(S) \tag{13}
\]

\[
b_h(u_h, \mu_h) = 0 \quad , \quad \mu_h \in M_h(S) . \tag{14}\]

In order to guarantee the ellipticity of the bilinear form \(a_h(\cdot, \cdot)\) on \(\text{Ker} \ B_h\)

\[
a_h(v_h, v_h) \geq \alpha \|v_h\|_{V_h(\Omega)}^2 \quad , \quad v_h \in \text{Ker} \ B_h \quad , \quad \alpha > 0 , \tag{15}\]

where \(B_h : V_h(\Omega) \rightarrow (M_h(S))^*\) is the operator associated with \(b_h(\cdot, \cdot)\) as well as the discrete inf-sup condition (LBB-condition)

\[
\inf_{\mu_h \in M_h(S)} \sup_{v_h \in V_h(\Omega)} \frac{b_h(v_h, \mu_h)}{\|v_h\|_{V_h(\Omega)} \|\mu_h\|_{M_h(S)}} \geq \beta > 0 , \tag{16}\]

the multiplier spaces \(M_h(\Gamma_{ij})\), \(\Gamma_{ij} \subset S\), have to be chosen in an appropriate way. This can be done by a suitable modification of the basis functions associated with edges in the interior of \(\Gamma_{ij}\) that have neighboring edges on \(\partial \Gamma_{ij}\). The modification consists in adding a convex combination of the basis functions corresponding to the neighbors on \(\partial \Gamma_{ij}\). For details of the construction as well as a verification of (15) and (16) we refer to [4,13].

The numerical solution of (13), (14) has again to take into account a defect correction in subspaces of irrotational vector fields which now has to be adapted to the domain decomposition setting. We refer to [14] for details as well as to [3,14] for the realization of grid adaptation strategies based on efficient and reliable residual type a posteriori error estimators.

### 4 Minimization of the Total Inductivity by Topology Optimization

Significant power losses during operation of the converter module can be avoided by a reduction of the parasitic inductivities that can be achieved by an optimal distribution of the material in the bus bars. Such topology optimization problems are well-known in structural mechanics (cf., e.g., [5] and the references therein), but have not yet been investigated in the framework of an optimal design of high power electronic devices.
We use the conductivity $\sigma$ as a material parameter and choose an SIMP-methodology (Simple Isotropic Material with Penalization) where we consider a scaled version of the potential equations and their discrete analogs, respectively, with $\eta(\sigma) := ((\sigma - \sigma_{\text{min}})/(\sigma_{\text{max}} - \sigma_{\text{min}}))^q$ and an appropriately chosen $q \in \mathbb{N}$ (e.g., $q = 2$) for a penalization of intermediate conductivities. Here, $\sigma_{\text{max}}$ stands for the conductivity of the basic material (copper) and $0 < \varepsilon := q_{\text{min}} \ll 1$ to avoid difficulties due to a loss in ellipticity.

The potential equations are discretized implicitly in time by the backward Euler scheme and by Nédélec’s lowest order curl-conforming edge elements resp. the nonconforming P1 approximation in space with respect to a simplicial triangulation $\mathcal{T}_h$ as described in the previous section. Denoting the discretized total inductivity by $L_h$ and comprising the discrete state variables $\varphi_h$ and $\mathbf{A}_h$ in a single vector $\mathbf{u}_h := (\varphi_h, \mathbf{A}_h)^T$ for the sake of notational convenience, we are thus led to the following nonlinear minimization problem:

Find $\sigma_h^*, \mathbf{u}_h^*$ such that

$$ L_h(\sigma_h^*, \mathbf{u}_h^*) = \min_{\sigma_h, \mathbf{u}_h} L_h(\sigma_h, \mathbf{u}_h) \tag{17} $$

subject to the equality constraints

$$ C_h(\sigma_h) \mathbf{u}_h = \mathbf{b}_h, \tag{18} $$

$$ g(\sigma_h) := \sum_{K \in \mathcal{T}_h} \text{meas}(K) \sigma_h |_K = C \tag{19} $$

and the inequality constraints

$$ \sigma_{\text{min}} \mathbf{e}_h \leq \sigma_h \leq \sigma_{\text{max}} \mathbf{e}_h, \tag{20} $$

where $\mathbf{e}_h := (e_{h,i})_{i=1}^{d_h}, e_{h,i} = 1, 1 \leq i \leq d_h := \text{card} \mathcal{T}_h$. Note that the system (18) represents the discretized potential equations in compact form.

For the numerical solution of (17)–(20) we use a primal-dual Newton interior-point method where the inequality constraints (20) are taken care of by logarithmic barrier functions. To be more precise, we consider a sequence of minimization subproblems of the form

$$ \min_{\sigma_h, \mathbf{u}_h} B_h(\sigma_h, \mathbf{u}_h, \rho_h), \tag{21} $$

$$ B_h(\sigma_h, \mathbf{u}_h, \rho_h) := L_h(\sigma_h, \mathbf{u}_h) - \rho_h [\log (\sigma_h - \sigma_{\text{min}} \mathbf{e}_h) + \log (\sigma_{\text{max}} \mathbf{e}_h - \sigma_h)] $$

subject to the equality constraints (18), (19) where $B_h(\sigma_h, \mathbf{u}_h, \rho_h)$ is the barrier function and $\rho_h > 0$ the barrier parameter.
The equality constraints are coupled by Lagrangian multipliers giving rise to the Lagrangian

\[ \mathcal{L}_h(\sigma_h, u_h, \lambda_h, \mu_h) := B_h(\sigma_h, u_h, \rho_h) + \lambda_h \cdot (C_h(\sigma_h) u_h - b_h) + \]
\[ + \mu_h (g_h(\sigma_h) - C). \]

The first-order Karush–Kuhn–Tucker (KKT) conditions for the associated saddle point problem are solved for decreasing values of the barrier parameter \( \rho_h \) by a Newton-type method with line-search. We choose a primary merit function based on the logarithmic barrier function and an augmented Lagrangian as well as a secondary merit function in terms of the \( \ell_2 \)-norm of the residual with respect to the KKT conditions combined with a watchdog strategy for convergence monitoring (for details we refer to [16,17]).

5 Numerical Results

The application of the design algorithm typically results in a material distribution which can be displayed by a grey-scale ranging from black (\( \sigma = \sigma_{\text{max}} \)) to white (\( \sigma = \sigma_{\text{min}} \)). Fig. 3 shows the results in a 2D situation where the design objective is to minimize the total amount of dissipated electric energy. The initial design was chosen as a uniform distribution. The performance of the primal-dual Newton interior-point method as described in Sect. 4 depends on the number of ports and individual contact currents (for details we refer to [16,17]). For an individual bus bar with prescribed currents at the ports, Fig. 4 displays the distribution of both the computed electric potential (left) and the computed electric currents (right). We refer to [9,11] for a detailed documentation.

Finally, Fig. 5 contains the visualization of the computed magnetic induction between two ports of the bus bar where the computation has been done by the adaptive multilevel method described in Sect. 3. Again, more details can be found, e.g., in [15].

![Fig. 3. Material distribution: 3 contacts (left); 5 contacts (right)](image-url)
6 Conclusions

In a combined way, we have used modern mathematical methods from structural optimization and the numerical solution of PDEs in general and the eddy current equations in particular to develop an efficient algorithmic tool for the optimal design of high power electronic devices. The numerical computations reveal that the design can be improved by a margin ranging between 10% and 20%.

Acknowledgments. The work has been supported by a grant from the Federal Ministry for Education and Research (BMBF) under Grant No. 03H07AU1-8. The fourth author is indebted to the Alexander-von-Humboldt Foundation for an AvH-Fellowship.

References

Modelling and Simulation of Strained Quantum Wells in Semiconductor Lasers

Hans-Christoph Kaiser, Uwe Bandelow, Thomas Koprucki, and Joachim Rehberg

Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS), Mohrenstrasse 39, 10117 Berlin, Germany

Abstract. A model allowing for efficiently obtaining band structure information on semiconductor Quantum Well structures will be demonstrated which is based on matrix-valued kp-Schrödinger operators. Effects such as confinement, band mixing, spin-orbit interaction and strain can be treated consistently. The impact of prominent Coulomb effects can be calculated by including the Hartree interaction via the Poisson equation and the bandgap renormalization via exchange-correlation potentials, resulting in generalized (matrix-valued) Schrödinger–Poisson systems. Band structure information enters via densities and the optical response function into comprehensive simulations of Multi Quantum Well lasers. These device simulations yield valuable information on device characteristics, including effects of carrier transport, waveguiding and heating and can be used for optimization.

1 Introduction

Semiconductor lasers are the most promising optoelectronic devices for generation of intense light in very small spectral domains. With emission wavelengths ranging now from near UV to far IR, these laser diodes are used in a wide range of applications. Their operating state is usually characterized by conduction bands occupied with electrons and valence bands occupied with holes which are separated by an energy gap. Emission of light, the wavelength of which is roughly given by this bandgap, results from recombination of electron-hole pairs. Applications depending on proper adjustment of the emission wavelength require bandgap engineering.

Laser action requires optical gain, the description of which is the heart of semiconductor laser modelling. The gain corresponds to the imaginary part of the optical response function, which itself depends on almost all properties of the semiconductor material and the operating state of the device as well as on properties of the optical field, e.g. its wavelength and polarization. To increase the gain in a laser, carriers and photons should be confined together in a small (optical active) region (see Fig. 1), requiring semiconductor heterostructures, in general nanostructures. During the last decade sophisticated Quantum Well structures have been developed to achieve good confinement and high optical gain. This has been done by band structure engineering, using in particular geometry, band offset and strain. Widely applied, Strained Multi Quantum Wells (SMQW) are subject to intense research – as in this project.
Fig. 1. Scheme of a Ridge-Waveguide (RW) Strained Multi Quantum Well (SMQW) laser diode, by Heinrich Hertz Institut für Nachrichtentechnik, Berlin. By forward biasing the diode, holes are injected from the p-contact, electrons from the n-contact, and get confined in the optical active SMQW region, which is enlarged on the right. Supported by the waveguide layers and the ridge optical modes are guided along the active region to achieve a significant optical gain.

For state of the art optoelectronic devices mainly two material classes are important which differ in their crystal symmetry: cubic and wurtzite. Cubic materials can be found in long wavelength lasers based on e.g. Indium Phosphide (InP) and/or Gallium Arsenide (GaAs). The more recent blue/UV diode lasers are based on Gallium Nitride (GaN), usually grown in wurtzite configuration. Gain spectra for a cubic (left) and a wurtzite material system (right) are drawn in Fig. 2, calculated with KPLIB[1]. The optical response is extremely sensitive to the electronic band structure and to the transition matrix elements. The calculation of these quantities for the optical active SMQW region (see Fig. 1) will be subject of Sect. 2. We have used these data for predicting the characteristics of devices as drawn in Fig. 1. Corresponding models and simulation results will be presented in Sect. 3.

Fig. 2. Examples of material gain dispersion in Strained Quantum Wells at room temperature for different sheet concentrations. Left: InGaAsP, Right: AlGaN/GaN
Throughout this paper our example device is the SMQW-laser depicted in Fig. 1. It is a 470 \( \mu \text{m} \) long single section Fabry–Perot laser with cleaved facets, designed for emitting at 1.55 \( \mu \text{m} \). The optical active region (enlarged on the right of Fig. 1) consists of six 7 nm thick compressively strained Quantum Wells, sandwiched by 10 nm thick tensile strained barriers.

2 Quantum States

Strained Multi Quantum Wells give rise to strong band mixing effects. These affect not only the Density of States, but also the transition matrix elements between these states and hence the optical response function.

We treat band mixing effects based on \( \text{k}p \)-calculations with the quantum states \( \Psi_{l,k_{||}}(r) \) described within the Envelope Function Approach:

\[
\Psi_{l,k_{||}}(r) = \exp(ik_{||}r_{||}) \sum_{\nu} F_{\nu l}(z; k_{||}) u_{\nu,k=0}(r). \tag{1}
\]

The index \( || \) indicates in-plane vectors and \( z \) denotes the quantization direction. The \( u_{\nu,k=0}(r) \) are zone-center Bloch functions, varying on the atomic scale, \( F_{\nu l}(z; k_{||}) \) are envelope functions, varying on the nanoscale. Based on the conventional \( \text{k}p \) method [2, ch.4] applied here this approach yields a matrix Hamiltonian \( H_{\mu \nu}(k_{||}, k_z = -i \frac{\partial}{\partial z}, V(z), \epsilon) \). In this simple form this \( \text{k}p \)-Schrödinger operator covers the crystal symmetry and the spin–orbit interaction, expressed by measurable Luttinger parameters, the confinement \( V(z) \) imposed by the band offsets of the adjacent materials, and the impact of strain \( \epsilon \) induced by their different solitary lattice constants.

The envelope functions \( F_{\nu l} \) of the quantum states and the corresponding subband dispersions \( E_{l}(k_{||}) \) are solutions of the respective eigenvalue problem

\[
\sum_{\nu} H_{\mu \nu}(k_{||}, k_z = -i \frac{\partial}{\partial z}; \ldots) F_{\nu l}(z; k_{||}) = E_{l}(k_{||}) F_{\mu l}(z; k_{||}). \tag{2}
\]

In KPLIB [1], we have implemented several stages in the hierarchy of \( \text{k}p \)-models. The simplest stage applies to Light Hole (LH) – Heavy Hole (HH) band mixing based on a 4x4 matrix Hamiltonian. Additional coupling to Spin-Split Off (SO) bands enlarges the problem to a 6x6 matrix Hamiltonian. Both are purely valence-band Hamiltonians valid for describing the hole dispersion to some extent. Valence-band (VB) Hamiltonians are mathematically advantageous, because they are always semibounded [3].

For consistently modelling mixing of conduction and valence bands we use 8x8 Hamiltonians. This causes nonparabolic dispersion for the conduction subbands as well and allows for consistently predicting optoelectronic properties. 8x8 Hamiltonians are not semibounded and their mathematical and numerical treatment is much more difficult, because standard comparison arguments for eigenvalues do not apply. Indeed, spurious solutions and their
elimination are discussed in the literature [4]. We have performed a spectral analysis [3], showing, that spurious solutions don't appear if the interband coupling remains weak enough. This is the case in wide gap materials as GaN, whereas GaAs and InP based materials are situated close above this limit.

2.1 Multi Quantum Wells

We have investigated the Multi Quantum Wells based on (8x8) kp-calculations. With decreasing barrier width the states localized in the individual Quantum Wells start to couple with each other which gives rise to miniband formation. We have studied this behaviour, which is illustrated in Fig. 3. In our example, no dispersion is observed for the lowest conduction and highest valence minibands, which contribute most to the optical response. This is a typical feature, indicating a good laser design for achieving small linewidths in the laser emission. In such cases we may confine to the single quantum well

Fig. 3. Miniband formation in a SMQW structure consisting of 6 Quantum Wells. Due to symmetry all eigenvalues are twice degenerate, yielding groups of 12 eigenvalues

Fig. 4. Subband dispersion and warping in a 7nm InGaAsP compressively strained Quantum Well, calculated with KPLIB for different crystallographic directions in the k_\parallel plane
case, which considerably reduces the numerical effort. As is visible in Fig. 3, the other minibands exhibit dispersion, corresponding to less localized states which interfere in Multi Quantum Wells.

2.2 Single Quantum Wells

The conduction and valence subband dispersion $E_l(k_{||})$ for our example Quantum Well is drawn in Fig. 4. Due to the compressive strain in the Quantum Well the LH-HH degeneracy at $k = 0$ is lifted, and the upper two valence subbands correspond to HH-like states. The dispersion is shown for two different crystallographic directions ([100] and [110]) in the $k_{||}$-plane, which exhibits the warping (angular dependence) effect. Fig. 4 displays weak warping for conduction subbands, but strong warping for the valence subbands.

2.3 Transition Matrix Elements

Taking the $k$-gradient of $H_{\mu\nu}(k;\ldots)$ and the solutions $F_{\nu i} = F_{\nu i}(z; k_{||})$ of (2) we get the momentum matrix elements $p_{ij} = p_{ij}(k_{||})$:

$$p_{ij} = \frac{m_0}{\hbar} \sum_{\mu,\nu} \left( F_{\mu i} \nabla_k H_{\mu\nu} (k_{||}, k_z; \ldots) \right)_{ik_z = \frac{\partial}{\partial z}} F_{\nu j},$$  \hspace{1cm} (3)

($m_0 = \text{free electron mass}$) which quantify the transition rate between the states involved under presence of optical excitation. This transition rate enters, e.g., the optical gain, expressed by (7). The corresponding oscillator strengths for our quantum well are depicted in Fig. 5, both for the TE-polarized ($p_{ij} \perp e_z$) and the TM-polarized case ($p_{ij} \parallel e_z$). Obviously, the TE-polarization is favoured by our particular structure, which is due to the

![Fig. 5. Interband oscillator strength dispersion for transitions between the lowest conduction subband and the upper valence subbands for different crystallographic directions, corresponding to Fig. 4, for TE- and TM-polarization](image-url)
support for the CB-HH like transitions by the compressive strain in the wells (see Fig. 4). Especially for the main transitions we have to state a strong influence of warping and it should be noted that calculations restricted to only [100] or [110] directions are expected to give wrong results for the response functions.

2.4 Carrier Densities and Optical Response Function

If the Fermi levels $\phi_n$ and $\phi_p$ of the quantum confined electrons $n$ and holes $p$ are given, or, alternatively, their sheet concentrations, then the quantum confined carrier densities are obtained by

$$n(z) = \sum_{i \in e} \int \frac{d^2k_\parallel}{(2\pi)^2} \frac{1}{f(E_i(k_\parallel) - \phi_n)} \sum_{\nu} |F_{\nu l}(z; k_\parallel)|^2,$$

$$p(z) = \sum_{i \in v} \int \frac{d^2k_\parallel}{(2\pi)^2} \left( 1 - f\left(E_i(k_\parallel) - \phi_p\right) \right) \sum_{\nu} |F_{\nu l}(z; k_\parallel)|^2$$

with the Fermi distribution

$$f(E) = \left( 1 + \exp\left(\frac{E}{k_B T}\right) \right)^{-1},$$

where $k_B$ denotes the Boltzmann constant and $T$ the temperature. An example for the corresponding density distributions in our quantum well structure is shown in Fig. 7.

In addition, we can calculate the optical response function $\varepsilon_{opt}$. In our model the real part of $\varepsilon_{opt}$, which corresponds to the (squared) refractive index, consists of different contributions, stemming from a background as well as from interband and intraband transitions which are calculated according to Wenzel et al. [5]. The resulting dispersion of the real part of $\varepsilon_{opt}$ is shown for different sheet concentrations in the left part of Fig. 6. Besides a (frequency dependent) prefactor the interband contribution to the imaginary part of $\varepsilon_{opt}$ corresponds to the material gain, which we have calculated according to [6]:

$$g(\omega) = \frac{\pi \hbar q^2}{\varepsilon_0 n_r^2 n c L_z} \sum_{i \in e} \sum_{j \in v} \frac{d^2k_\parallel}{(2\pi)^2} \frac{|P_{ij} e|^2}{E_i - E_j} f(E_i - \phi_n) \left( 1 - f(E_j - \phi_p) \right) \times$$

$$\times \left[ 1 - \exp\left(\frac{\hbar \omega - (\phi_n - \phi_p)}{k_B T}\right) \right] \frac{1}{\pi \left[ (E_i - E_j) - \hbar \omega \right]^2 + \Gamma^2}$$

where the last factor includes broadening processes, modeled with a characteristic intraband relaxation time $\tau$ of 60fs ($\Gamma = \hbar / \tau$). $e$ is the polarization direction of the optical field, $q$ denotes the elementary charge, $c$ the speed of light, $n_r$ the refractive index and $L_z$ is a normalization length. Examples for the dispersion of the imaginary part of $\varepsilon_{opt}$ are drawn in Fig. 6, which exhibit significant maxima, a prerequisite for laser action (see also the gain maxima in Fig. 2).
2.5 Selfconsistent Quantum States

The carriers are charged particles and therefore exert Coulomb interaction on each other which is additionally modified by their fermionic character. The Coulomb interaction would guide us to much more complicated and nonlinear problems and results in important physical effects [7, ch.4] – the most prominent of which being the reduction of the bandgap to which we confine our analysis here. We include into our kp-model a Hartree potential and, motivated by the Density Functional Formalism [8], an exchange-correlation potential.

As can be seen from the thin lines in Fig. 7 the quantum confined carriers exhibit different localization behaviour in the Quantum Well, which is due to the different band offsets, Luttinger parameters and the presence of strain. In effect, an electrostatic potential builds up (see the full line in Fig. 8), which attracts the holes and repulses the electrons from the well in our particular case. Therefore we expect more equidistributed carriers than suggested by pure kp-calculations. This behaviour is covered by including the Hartree interaction into the kp-Schrödinger operator. The Hartree potential $\varphi_H$ is a solution of the Poisson equation:

$$\frac{d}{dz} \left( \varepsilon \frac{d}{dz} \varphi_H \right) = q(n - p)$$

where the carrier densities according to (4),(5) enter the r.h.s., $\varepsilon$ is the static dielectric permittivity. The arising system (2),(8) accomplished with (4),(5) can be viewed as a generalized (matrix-valued) Schrödinger–Poisson system. The bandgap shift now enters our model via density dependent exchange-correlation potentials $V_{xc}(n, p)$. Typical potentials are plotted in Fig. 8.

The mathematical analysis [9], [10] of the case without band mixing displays an unique solution for certain $V_{xc}$-potentials. Fortunately, in our example we also observed rapid convergence. In our particular case the wells
**Fig. 7.** Density distribution in a 7nm Quantum Well, sandwiched between 10nm barriers for a sheet concentration of $4 \cdot 10^{12}/\text{cm}^2$. Thin: pure 8x8 kp. Thick: selfconsistent 8x8 kp with Hartree and exchange-correlation potential.

**Fig. 8.** Spatial distribution of Hartree- and exchange-correlation potentials in a 7 nm InGaAsP strained Quantum Well, selfconsistently calculated for a sheet concentration of $4 \cdot 10^{12}/\text{cm}^2$.

**Fig. 9.** Spectrum of the material gain for TE- and TM polarization, corresponding to the case in Fig. 8. Thick: selfconsistent 8x8 kp, thin: pure 8x8 kp-calculations.
and barriers are undoped. Therefore, we are close to the local charge neutrality \( n = p \). For this case we have adopted a formula which is based on a result by Zimmermann [11] for parabolic bands:

\[
V_{xc}^e = V_{xc}^h \propto \sqrt{\frac{n_s}{n_s + T/T_0}}, \quad n_s \propto \frac{n + p}{2},
\]

which is also valid for finite temperatures. Fig. 7 shows the impact of the Hartree-contribution together with the exchange-correlation effects on the density distributions.

Exchange-correlation effects enhance the tendency for localization of the carriers in the Quantum Well, uniformly for electrons and holes. At the chosen sheet concentration the exchange-correlation potential \( V_{xc} \) is approximately of the same amount as the Hartree potential (see Fig. 8). We have no net effect on the electrons because \( \varphi_H \) and \( V_{xc} \) nearly cancel each other, which is reflected in the similarity of the pure-\( kp \) (thin) and the selfconsistent (thick) electron density distribution curve in Fig. 7. For the holes the two potentials sum up to approximately twice the Hartree potential effect. This yields a strong enhanced localization for the holes, as reflected by the thick dashed curve in Fig. 7. The resulting net gap shrinkage of about 30meV is finally reflected in the shift of the gain spectrum in Fig. 9. The redshift, approximately 50nm, of the TE- as well as the TM-gain spectra corresponds very good to that value. The relative enhancement of the gain maxima visible in Fig. 9 reflects the more equidistributed electron- and hole wave functions, which results in a better overlap between them and hence increases the oscillator strength.

3 Device Simulation

A schematic overview on our simulations based on the packages KPLIB [1] and WIAS-TeSCA [12] is depicted in Fig. 10. The carrier transport is governed by Drift-Diffusion equations for electrons and holes:

\[
q \frac{\partial}{\partial t} n - \nabla J_n(n, \varphi) = qR,
\]

\[
q \frac{\partial}{\partial t} p + \nabla J_p(p, \varphi) = qR,
\]

which are coupled to the Poisson equation for the electrostatic potential \( \varphi \)

\[
-\nabla(\varepsilon \nabla \varphi) = q(N_D - N_A + p - n).
\]

In the above equations \( N_D - N_A \) is the net doping profile and \( R \) in (10),(11) is short for all recombination terms, which depend on the carrier densities, the electrostatic potential, the optical field and much other quantities. The
relation between the carrier densities and the potentials is described by state
equations in terms of Fermi-Dirac statistics

\[ n = N_c \mathcal{F}_{1/2} \left( \frac{E_c + qF_n - g \varphi}{k_B T} \right), \quad p = N_v \mathcal{F}_{1/2} \left( \frac{qF_p - E_v - q \varphi}{k_B T} \right), \quad (13) \]

where \( E_c \) denotes the conduction and \( E_v \) the valence band edge. \( F_n \) and \( F_p \) are the quasi Fermi potentials of electrons and holes, the gradients of which

drive the current densities \( J_n, J_p \), respectively. \( \mathcal{F}_{1/2} \) is the Fermi-integral of order 1/2.

According to (13) the carriers are treated bulk-like with band-edge Density of States \( N_c \) and \( N_v \). For the quantum confined carriers the \( kp \)-calculations yield at least a modified Density of States \( N_c \) and \( N_v \). In an advanced stage of the model the quantum confined carriers will be established as a separate species coupled to the free carriers by capture-escape mechanisms.

3.1 Optics

The above carrier equations are coupled to equations for the optical field. Due
to properties of the laser resonator the latter are scalar Helmholtz equations

\[ \left[ \Delta + \frac{\omega^2}{c^2} \varepsilon_{\text{opt}}(\omega, n, p) \right] \Phi_j = \beta_j^2 \Phi_j, \quad (14) \]

which describe the spatial distribution of the optical modes \( \Phi_j \) within the sim-
ulation plane (see Fig. 11). These modes are characterized by their respective
(complex) eigenvalues \( \beta_j \). The imaginary part of \( \varepsilon_{\text{opt}}(\omega, n, p) \) in (14) contains
the material gain contribution, which is fitted to our \((8x8)\) \( kp \)-calculations
(see (7) and Fig. 6), as well as contributions from the Inter Valence Band Abs-
sorption (IVBA), described by phenomenological models in WIAS-TeSCA.
As argued earlier and indicated by Fig. 9, we have strong support for the TE-polarization, but nearly no support for the TM-polarization in our structure. For this reason, calculations here are confined to TE modes, drawn in Fig. 11.

In 2D-simulations with WIAS-TeSCA, longitudinal properties are considered by assuming a longitudinally homogeneous power distribution, which is approximately met in Fabry-Perot lasers or in edge-emitting lasers with properly designed Bragg gratings [13]. In our calculations the two modes shown in Fig. 11 have been involved, the number of photons $S_j$ of which are balanced by corresponding photon rate equations

$$\dot{S}_j = v_{gj} (2 \Im m \beta_j - \alpha_j) S_j + \dot{S}^{\text{spont}}_j,$$

where $\dot{S}^{\text{spont}}_j$ is short for the spontaneous emission into the mode, $v_{gj}$ denotes the modal group velocity, $\alpha_j$ is the sum of longitudinal scattering- and output losses at the facets and $\Im m \beta_j$ is the imaginary part of the corresponding eigenvalue subject to (14).

### 3.2 Heating

For a realistic estimation of the device performance heating effects are considered within the model. To that end WIAS-TeSCA comprises an energy transport equation [14], [15].

In Fig. 12 a stationary spatial temperature distribution for our example device is shown. The temperature of the heat sink was adjusted to room temperature (300K). The main heat source is located within the active region, corresponding to the profile of the main mode. Accordingly, we address the main heating processes to the IVBA.

For estimating the thermal stability of the device operation we have studied the influence of different temperatures of the heat sink on the power-current (PI) characteristics. Results of our simulations are shown in Fig. 13.
Two important effects are observed, which are in agreement with the measurements. First, we can reproduce the significant shift of the laser threshold with temperature caused by decreasing material gain accompanied by an increase of IVBA. Due to these processes higher carrier densities are required for lasing, which enhances the recombination and therefore requires a larger threshold current (Fig. 13). Second, we can computationally reproduce a thermal roll-over at higher temperatures, which can be explained mainly by the significant increase of the Auger recombination which strongly decreases the quantum efficiency.

4 Summary

Many modern optoelectronic applications essentially rely on Quantum Well structures which need to be properly designed. Based on the Envelope Func-
tion Approach and kp-theory, a model has been developed which allows to simulate such structures consistently. Important information provided by the model are the nonparabolic band structure, the quantum confined states, their respective transition matrix elements, carrier densities and the optical response function, which is crucial for semiconductor laser modelling. Furthermore, this model has been extended for the Hartree interaction as well as for the density dependent bandgap shift to become as realistic as possible. Based on such results, more comprehensive device simulations have been performed. These simulations additionally included effects of carrier transport, doping, optical waveguiding and heating. The simulation results are in good agreement with the experiments, indicating their applicability for designing modern optoelectronic devices.

Acknowledgments. We acknowledge the cooperation with the Heinrich-Hertz Institut für Nachrichtentechnik Berlin GmbH, in particular Martin Möhrle. We thank H. Wenzel (FBH Berlin) and H.-J. Wünsche (HU Berlin) for many fruitful discussions on laser modelling. Uwe Bandelow gratefully acknowledges the German Federal Ministry for Education, Sciences, Research and Technology for support of this work under grant no. 03-KA7FV1-4. The work of Thomas Koprucki has been supported by the Deutsche Forschungsgemeinschaft under grant no. HA 1807/5-1.

References


VI. Electronic Circuits

Efficient Analysis of Oscillatory Circuits
   *R. Bulirsch, R. Neubert, A. Schwarz*

Modelling and Simulation of Power Devices for High-Voltage Integrated Circuits
   *R. Hünlich, G. Albinus, H. Gajewski, A. Glitzky, W. Röpke, J. Knopke*

Finding Beneficial DAE Structures in Circuit Simulation

CHORAL – a Charge-Oriented Algorithm for the Numerical Integration of Electrical Circuits
   *P. Rentrop, M. Günther, M. Hoschek, U. Feldmann*
Efficient Analysis of Oscillatory Circuits

Roland Bulirsch\textsuperscript{1}, Rolf Neubert\textsuperscript{2}, and Angela Schwarz\textsuperscript{1}

\textsuperscript{1} Technische Universität München, Zentrum Mathematik, 80290 München, Germany
\textsuperscript{2} Infineon Technologie AG, MP PTS, 81730 München, Germany

Abstract. Circuit simulation is a standard task in the computer-aided design of electronic circuits. The generation of the circuit equations leads to differential-algebraic equations (DAE). With respect to the analysis of oscillatory circuits, periodic solutions are of major interest. For highly oscillatory circuits, direct methods must be used to obtain the limit cycle efficiently. A specialized shooting method for solving the associated boundary-value problem of the differential-algebraic network equations is presented.

1 Introduction

Due their rapid development, the design of integrated circuits can be realized only with the aid of powerful simulation packages. The modelling of the electrical circuits are treated as networks consisting of nodes and elements and their interconnections. The topology of the network determines how the elements are connected. The model equations are derived from the element equations and Kirchhoff’s law. The charge/flux-oriented modified nodal analysis leads to a system of quasi-linear differential-algebraic equations of high dimension with the form

$$A \cdot q(x(t)) + f(x(t), t) = 0.$$  \hspace{1cm} (1)

$x(t)$ is the vector of unknowns consisting of the node potentials and the branch currents of voltage controlling elements, $q(x)$ is the vector of the electrical charges and the magnetic fluxes, $A$ is the incidence matrix that describes the network topology and is normally singular, and $f(x,t)$ includes the static linear and nonlinear parts of the circuit and the time dependent input signals. In general the equations are of index 1 or index 2. The special structure is analyzed in detail in [6], [7], [3], [12].

Oscillatory circuits present a special task in the field of circuit simulation. The periodic oscillation is important e.g. for radio technology, analog-to-digital conversion and as clock signals in microprocessors. The analysis of the oscillatory circuits includes the computation of periodic steady state solutions, their stability, periodic noise and distortion analysis. This article deals exclusively with the determination of the periodic solution. To avoid the expansive transient integration of numerous periods, direct methods are utilized to approximate the limit cycles. The most common methods are harmonic balance and shooting.
The method of harmonic balance approximates the periodic solution using a Fourier-polynomial and solves for the coefficients by means of a Galerkin approach [10]. Harmonic balance has been shown to be efficient for a wide range of circuits but for very large circuits or for circuits having pulse like excitations (which force the usage of many Fourier coefficients within the approximation), the method may run into efficiency or accuracy problems. An alternative approach is to formulate and solve a periodic boundary-value problem using a shooting method. The periodic boundary condition reads

$$x(0) - x(\tau) = 0,$$

where $\tau$ is the period of the input and output signal. If the system (1) is autonomous, the period is unknown. For autonomous systems, the boundary-value problem can be transformed to the interval $[0, 1]$ and the period is then treated as an additional unknown variable. Besides the periodicity condition, a phase condition must be specified. Here the non-autonomous case is treated exclusively.

A shooting method for periodic boundary-value problems (1), (2) of index 1 is presented. The concept can be transferred to the index 2 case, but this will not be illustrated here.

2 A Shooting Method for the DAE System of the Network Equation

For a solution $\varphi(t; x_0)$ with $\varphi(0; x_0) = x_0$ of the given implicit DAE-system

$$F(\dot{x}(t), x(t), t) = 0,$$

a consistent initial value $x_0$ must be found such that the periodicity condition (2) is fulfilled.

Similar to multiple shooting methods for high dimensional ODE systems, the nonlinear equation for the unknown initial state $x_0$ is formulated as the minimization problem

$$\min_{x_0 \in \mathbb{R}^n} \| \varphi(\tau; x_0) - x_0 \|_2^2.$$ 

For the initial state of a periodical solution the residual vanishes.

A Gauss–Newton method is used to solve the minimization problem. It consists of a sequence of linear approximations of the residual function to be minimized. For the actual approximation of the initial state $x_0^i$, a correction $\Delta x^i$ is determined as the solution of the linearization of the argument of the minimization problem:

$$x_0^{i+1} = x_0^i + \Delta x^i, \quad i = 0, 1, ...$$
Efficient Analysis of Oscillatory Circuits

\[
\min_{\Delta x^i \in \mathbb{R}^n} \| (X(\tau) - X(0)) \Delta x^i + \varphi(\tau; x_0^i) - \varphi(0; x_0^i) \|_2^2.
\]  

(3)

\[ X(\tau) = \frac{\partial \varphi}{\partial x_0}(\tau; x_0^i) \] is the sensitivity of the solution \( \varphi \) with respect to the initial value \( x_0^i \) at \( t = \tau \) and \( X(0) = \frac{\partial \varphi}{\partial x_0}(0; x_0^i) \) is the sensitivity at the initial time \( t = 0 \). For DAEs the iteration matrix \( S_X \) is singular in which case (3) has no unique solution \( \Delta x^i \). This is exploited to improve the efficiency of the method. Instead of the square singular sensitivity matrix \( X \) with rank \( X = k \), a matrix \( X_\alpha \in \mathbb{R}^{n \times k} \) with the same image is used. This results in reduced costs for the method. To enlarge the convergence region of the method a damped Gauss–Newton method is used in practice (cf. [11]).

The fundamental matrix \( X(\tau) \) is the solution of the variational problem

\[
F_x(\varphi, \varphi, t) \dot{X}(t) + F_x(\varphi, \varphi, t) X(t) = 0
\]  

(4)

(cf. (5) for the network equations), where \((F_x, F_x)\) is a regular matrix pencil. The consistent initial conditions are given by \( P(X(0) - I) = 0 \), where \( P \) is a projector on the subspace

\[
S(t) := \{ z \in \mathbb{R}^n : F_x z \in \text{im } F_x \}.
\]

All solutions of (4) belong to \( S(t) \). \( P \) is constant for the network equations, because the leading nullspace of these equations is constant.

Due to [5], \( X(t) \) can be expressed as

\[
X(t) = K(t)P
\]


\[
S_X = (K(\tau) - K(0))P = S_X P, \quad \text{as } P^2 = P,
\]

the minimization problem can be written as (omitting the iteration index \( i \))

\[
\min_{\Delta x \in \mathbb{R}^n} \| S_X \Delta x - r \|_2^2 = \min_{\Delta x \in \mathbb{R}^n} \| S_X P \Delta x - r \|_2^2 = \min_{\Delta x \in S(t)} \| S_X \Delta x - r \|_2^2.
\]

Without increasing the minimum, the solution can be selected out of the image space of \( P \), \( \text{im } P = S(t) \).

Let \( V \) be an orthonormal basis of \( S(t) \). Then \( V V^T \) is a projector onto \( S(t) \), \( P := V V^T \). The definition \( \Delta \alpha := V^T \Delta x \) leads to:

\[
\min_{\Delta x \in \mathbb{R}^n} \| S_X \Delta x - r \|_2^2 = \min_{\Delta x \in S(t)} \| S_X V V^T \Delta x - r \|_2^2
\]

\[
= \min_{\Delta \alpha \in \mathbb{R}^k} \| S_X V \Delta \alpha - r \|_2^2.
\]
A minimal solution $\Delta x$ in $S(t)$ results from $\Delta x = V \Delta \alpha$. Using the abbreviation $X_\alpha(t) := X(t) \cdot V$ the minimization problem reads

$$\min_{\Delta \alpha \in \mathbb{R}^k} \| (X_\alpha(\tau) - V) \Delta \alpha - r \|^2, \quad \Delta x = V \Delta \alpha.$$ 

The minimization problem in the space $S(t)$ does not require the full square sensitivity matrix. The correction $\Delta x$ lies in the tangent space of the constrained manifold in the actual $x_0^i$, that is $S(t)$. With it the new iterative $x^{i+1} = x^i + \Delta x^i$ fulfills the system equations to first order. For the index 1 case, the system equations are the consistency condition. In contrast to this, a general statement about the consistency of the correction obtained by the direct solution of (3) is difficult. In general, without the restriction of $\Delta x$ onto $S(t)$ the $\Delta x$ with the smallest norm is chosen to get a unique solution.

To determine a full consistent initial value $x_0$, the new iterate $x^{i+1}$ is projected onto the constrained manifold in a second step. Therefore the new approximation of the initial value is defined as the $y$ which is closest to $x^{i+1}$ and also lies on the constrained manifold. $y$ is the solution of the constrained minimization problem

$$\min_{y \in \mathbb{R}^n} \| x^{i+1} - y \|^2_2$$

$$\phi(y) = 0$$

where $\phi(y) = 0$ describes the constrained manifold. This description can be reached using the projectors onto the corresponding subspaces (cf. [8]). In the special case of the network equations resulting from the charge/flux-oriented MNA, these projectors result from the network topology (cf. [3], [2]).

**The convergence of the method.** The sensitivity matrix $X(t)$ is supposed to be Lipschitz continuous. For the convergence of the Gauss–Newton method, the matrix $X_\alpha(\tau) - V$ must have full column rank $k$ (cf. [1]). This is true if the monodromy matrix $X(\tau)$ of the non-autonomous system has no eigenvalue equal to 1. This can be seen as follows:

$$X_\alpha(\tau) - V = (X(\tau) - I) V$$

where $I$ is the identity matrix in $\mathbb{R}^n$. $V \in \mathbb{R}^{n \times k}$ is a basis of $S(t)$, i.e. it has full column rank $k$. $X(\tau) - I$ is regular, that is it has no eigenvalue 0, if 1 is no eigenvalue of $X(\tau)$.

**Approximation of the sensitivity.** The approximation of the sensitivity $X$ is done by the integration of the linear variational problem. This is the derivation of system equations with respect to the initial values. For the network equations this reads

$$A \frac{d}{dt} (q_x(\varphi(t)) X(t)) + f_x(\varphi(t), t) X(t) =$$

$$= A q_x(\varphi(t)) \dot{X}(t) + (f_x(\varphi(t), t) + A \dot{q}_x(\varphi(t))) X(t) = 0$$

(5)
Efficient Analysis of Oscillatory Circuits

with $q_x = \frac{d}{dx} q$. $X_\alpha$ is the solution of the variational problem multiplied by $V$ from the right, that is the solution of (5) with the initial values

$$X_\alpha(0) = X(0)V = V.$$ 

In general low order BDF-methods, the trapezoidal rule or some versions are used to integrate the network equations. This is also true for the integration of the variational problem. The step size and the order of the integration method for the sensitivity is regulated by the integration of the trajectory. Using the discretization of a BDF-method

$$\frac{d}{dt}(q_x(t_m)X(t_m)) = \sum_{i=0}^{k} \alpha_i q_x(t_{m-i})X(t_{m-i}),$$

where $\alpha_i$ are the coefficients and $k$ is the order, leads to a linear equation for the unknown $X(t_m)$

$$(\alpha_0 A q_x(t_m) + f_x(t_m)) X(t_m) = -A \sum_{i=1}^{k} \alpha_i q_x(t_{m-i}) X(t_{m-i}).$$

Because the same step size and order are used for the integration of the sensitivity as for the integration of the trajectory, the matrix of the linear system is known already. It results from solving the nonlinear equation of the implicit discretization of the network equations. During the calculation of the trajectory the sparse matrices are saved. In this way the matrix $X_\alpha$ can be computed efficiently.

**Calculation of the basis $V$.** Determine the basis $V$ means calculating $k$ linear independent consistent initial values of (4). A general way to get the basis is solving a generalized eigenvalue problem. Inserting the particular solution $X(t) = ve^{\lambda t}$ and $\dot{X}(t) = \lambda ve^{\lambda t}$ in (4) leads to the equation

$$\lambda F_x v + F_x v = 0$$

at the initial times $t = 0$. $V = (v_1, ..., v_m)$ with $v_i, i = 1, ..., m$ are the linear independent eigenvectors for the eigenvalues $\lambda_i$.

In the case of the network equations the basis is also available from the topology of the network. Especially for high dimensional problems this is much cheaper than solving the generalized eigenvalue problem. From the topological investigations in [3], [2], this results in a projector $P$. $k$ linear independent columns of $P$ form the basis. In both cases $V$ must be orthonormalized.

**3 Numerical Examples**

The method has been implemented in the industrial circuit simulator TITAN of Infineon Technologies AG, [4]. Is has been tested on various examples. Here
the results for two circuits are presented. The first considered circuit is a mixer, as shown schematically in Fig. 1. The multi-tone system is clocked by a local oscillator of frequency 900 MHz. The output signal is the superposition of a radio frequent input signal, e.g. of frequency 950 MHz and the local oscillator. In this example the mixer is down converting, i.e. the output signal is of frequency 50 MHz. The dimension of the problem is 29 and the basis \( V \) has a dimension of 13. Figure 2 shows the time and phase diagram of the output signal.

![Fig. 1. Scheme of a mixer circuit](image)

Fig. 1. Scheme of a mixer circuit

![Fig. 2. Time and phase diagram of a low-noise, highly linear 0.25μm CMOS-mixer](image)

Fig. 2. Time and phase diagram of a low-noise, highly linear 0.25μm CMOS-mixer

The second example illustrates the differences between the harmonic balance and the shooting method. For these purposes a commercially available simple operational amplifier μA709 circuit is considered (cf. Fig. 3). The dimension of the problem is 43, and the dimension of \( V \) is 34. First, sinusoidal input is applied. The output signal obtained by the method of harmonic balance with nine calculated frequencies nearly coincides with the solution obtained by the shooting method. With 17 frequencies the solutions are identically (cf. Fig. 4, left diagram). The calculation time for both methods is the same. Second, the circuit is excited by a pulse like input signal. In this case the method of harmonic balance needs about 25 frequencies for a solution of satisfying accuracy (cf. Fig. 4, right diagram). For the shooting method the
pulse like input does not cause any problems. As expected in this case, it is superior to harmonic balance with respect to run time and accuracy.

**Fig. 3.** Schematic of an operational amplifier μA709

**Fig. 4.** Output signal of the operational amplifier μA709, sinusoidal (left) and pulse like (right) input signal

**References**


Modelling and Simulation of Power Devices for High-Voltage Integrated Circuits

Rolf Hünlich¹, Günter Albinus¹, Herbert Gajewski¹, Annegret Glitzky¹,*; Wilfried Röpke¹,**, and Jürgen Knopke²

¹ Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany
² alpha microelectronics gmbh, 15236 Frankfurt (Oder), Germany

Abstract. Process and device simulators turned out to be important tools in the design of high-voltage integrated circuits and in the development of their technology. The main goal of this project was the improvement of the device simulator WIAS-TeSCA in order to simulate different power devices in high-voltage integrated circuits developed by the industrial partner. Some simulation results are presented. Furthermore, we discuss some aspects of the mathematics of relevant model equations which device and process simulations are based on.

1 Introduction

In the computer aided design of high-voltage Integrated Circuits (ICs) and in the development of their technology process and device simulation programmes turned out to be important tools. Challenges of new technologies and devices require a permanent discussion of underlying physical models, the mathematical analysis of related model equations, as well as the improvement of simulation codes. The main goal of this project was the extension of the device simulator WIAS-TeSCA [Gaj] by a self-consistent coupling of the van Roosbroeck system with a heat flow equation since thermal and thermoelectric effects play some rôle in power devices.

The industrial partner of the project is the company alpha microelectronics gmbh Frankfurt (Oder). The company develops and produces application-specific ICs for high-voltage applications (e.g. driver ICs for coils, motors and relays). The basis for all high-voltage ICs is a proprietary 500 V dielectric isolated MOS technology [Kno] which uses thick-film bonded silicon-on-insulator wafers with dielectric trench isolation.

Different power devices in the high-voltage ICs of the company were used as test structures for the development and application of WIAS-TeSCA. Here we present simulation results for a dielectric isolated, double Diffused Metal Oxide Semiconductor Transistor (DMOST). Furthermore, we discuss some aspects of the mathematics of relevant model equations which device and process simulations are based on.

* Supported by the German Research Foundation (DFG) grant HU 868/1-1
** Supported by the Ministry of Education and Research (BMBF) grant HU7FV1
2 Process Simulation

The simulation of the technological process of manufacturing semiconductor devices is used to develop such processes and to optimize them with respect to the device geometry and doping, for instance. We applied the process simulator DIOS-ISE [Str]. In Fig. 1 the simulated cross-section of a typical axisymmetric DMOST test structure with two gate contacts is shown. For the grid generation a compromise between accuracy and effort was reached.

Fig. 1. DIOS-ISE simulation result for the DMOST: cross-section (top, the unit of length is micron, the axis of rotation lies on the left hand side, Si – silicon) and detail of the source and gate contact regions (bottom)
Modelling and Simulation of Power Devices

The defaults of the diffusion parameters could mostly be used; only the redistribution of dopants during oxidation needed some changes. For this structure the thickness of the silicon layer and the doping concentrations (in the silicon layer, buried layer, sinker region) were varied to get different devices. Process simulation results are taken as input for subsequent device simulations.

3 Device Simulation

As mentioned in the introduction the device simulator WIAS-TeSCA was improved in order to be able to simulate power devices successfully. Now we shortly discuss some simulation results.

The aim of device simulation is to evaluate the electrical behaviour of the device. Often stationary current–voltage characteristics are required showing e.g. the drain current versus the drain(-source) voltage with the gate(-source) voltage as a parameter (see Figs. 2, 3, 4). Important device properties are derived from these characteristics which have to be optimized. The on-resistance is given by the reciprocal slope in the origin of the characteristics. The breakdown voltage characterizes the electrical breakdown of the device caused by avalanche generation of carriers due to impact ionization in high field regions and resulting in a drastic increase of the current. The knowledge of regions of high electric field strength, and thus also of high avalanche generation rate, is very useful for the designer to optimize the breakdown voltage. Finally, inhomogeneities of the temperature caused by large power densities at higher drain voltages should also influence the device behaviour since decisive physical parameters depend strongly on the temperature.

One important task for the simulation was the investigation of the influence of different dielectric isolated substrates on the device behaviour. In

![Fig. 2. WIAS-TeSCA simulation results for the DMOST: current–voltage characteristics at low drain voltages (left, bl – buried layer, he – heat flow equation) and avalanche generation current at high drain voltages (right)](image_url)
Electrostatic potential
Isolines for a drain voltage of 600 V

Temperature
Isolines for a drain voltage of 600 V
(the maximal device temperature lies on the axis of rotation)

Current–voltage characteristics
Solid line: simulation with heat flow equation (in some points the maximal device temperature $T_{\text{max}}$ [K] is indicated)
Dashed line: simulation for fixed temperature $T = 300$ K

Fig. 3. Simulation results for the DMOST with a 10μm thick buried layer
order to demonstrate the effect of the electric field we chose two D莫斯T structures as in Fig. 1, one has a buried layer, the other has not. For estimating the effect of inhomogeneities of the temperature all simulations were done with and without using the heat flow equation. On all electrical contacts and on the bottom of the slice (more precisely, in a depth of 100 μm) the temperature was fixed at 300 K while on the other parts of the boundary homogeneous Neumann conditions were posed.

Figure 2 shows current–voltage characteristics which are used for the determination of the on-resistance and breakdown voltage. Isolines of the electrostatic potential as well as the current–voltage characteristics at 10 V gate voltage are given in Fig. 3 for the structure with a buried layer and in Fig. 4 for the structure without a buried layer. Figure 3 shows also isolines of the temperature. For the structure without a buried layer isolines of the

---

**Electrostatic potential**

Isolines for a drain voltage of 600 V

---

**Current–voltage characteristics**

Solid line: simulation with heat flow equation (in some points the maximal device temperature $T_{\text{max}}$ [K] is indicated)

Dashed line: simulation for fixed temperature $T = 300$ K

---

**Fig. 4.** Simulation results for the D莫斯T without buried layer
temperature look similarly, but the values of the temperature are much lower. The electrostatic potential differs clearly in both devices such that the regions of high avalanche generation rate are also quite different, see Fig. 5.

4 Device Modelling

Simulations with WIAS-TeSCA are based on some energy model that extents the basic drift-diffusion model [Gaj] by a heat flow equation. There is a large variety of energy models (see e.g. [Kel] and references therein). Our approach can be found in [AGH]. Starting with a generally accepted expression for the density of the free energy and applying only first principles like the entropy maximum principle and the principle of local equilibrium a system of evolution equations is derived for different variants of energy models (including the Boltzmann or Fermi–Dirac statistics, non-parabolic band structures, electron–hole scattering models, the kinetics of deep traps, and hot carrier effects). Particular attention is paid to include the electrostatic potential self-consistently.

Here we describe a simple version of such a model. We start with state equations of the form

\[ n_i = \bar{n}_i(T) e^{(\zeta_i - q_i \psi)/T}, \quad \bar{n}_i(T) = [m_i(T) T^{3/2} e^{q_i E_i(T)/T}, \quad i = 1, 2, \]

\[ u = u(n_1, n_2, T) = c_L T + \sum_{i=1}^{2} n_i T^2 \frac{\partial \log \bar{n}_i}{\partial T}(T), \]

\[ s = s(n_1, n_2, T) = c_L \log T + \sum_{i=1}^{2} n_i P_i(n_i, T), \quad P_i = 1 + T \frac{\partial \log \bar{n}_i}{\partial T} - \log \frac{n_i}{\bar{n}_i}. \]
where \( n_1 \) is the electron density, \( q_1 = -1 \), \( n_2 \) is the hole density, \( q_2 = 1 \), \( \zeta_i \) are the electrochemical potentials, \( \psi \) is the electrostatic potential, \( T \) is the lattice temperature, \( m_i \) and \( E_i \) are the effective masses (up to a constant factor) and energy band edges, \( u \) denotes the density of the internal energy, \( s \) the density of entropy, \( c_L \) is the heat capacity of the lattice, and \( P_i \) are the thermoelectric powers.

The potential \( \psi \) fulfills the Poisson equation

\[
-\nabla \cdot \left( \varepsilon \nabla \psi \right) = f - n_1 + n_2 \tag{1}
\]

where \( \varepsilon \) is the dielectric permittivity, \( f \) is a given doping profile and does not depend on time. A mixed boundary condition is used,

\[
\psi = \psi^D \text{ on } \Gamma_D, \quad \varepsilon \frac{\partial \psi}{\partial \nu} + \tau(\psi - \psi^N) = 0 \text{ on } \Gamma_N \tag{2}
\]

where \( \Gamma \) denotes the boundary of the domain \( \Omega \) occupied by the device, \( \Gamma_D \) and \( \Gamma_N \) are disjoint, relatively open parts of \( \Gamma \) with \( \text{mes} (\Gamma \setminus (\Gamma_D \cup \Gamma_N)) = 0 \).

The particle fluxes \( j_i \) and the reduced heat flux \( j_q \) are assumed to be given by

\[
j_i = -\sigma_i(n_1, n_2, T) \left( \nabla \zeta_i + P_i \nabla T \right), \quad j_q = -\lambda(n_1, n_2, T) \nabla T
\]

with conductivities \( \sigma_i, \lambda > 0 \). A first form of the system of evolution equations consists of two continuity equations and a heat flow equation,

\[
\frac{\partial n_i}{\partial t} + \nabla \cdot j_i = -R, \quad c(n_1, n_2, T) \frac{\partial T}{\partial t} + \nabla \cdot j_q = H \tag{3}
\]

with the net recombination rate

\[
R = \rho(n_1, n_2, T) \left( e^{(\zeta_1 + \zeta_2)/T} - 1 \right), \quad \rho > 0,
\]

the heat capacity \( c = \partial u / \partial T > 0 \), and the heat generation rate

\[
H = \sum_{i=1}^{2} \left[ \frac{|j_i|^2}{\sigma_i} - T \nabla P_i \cdot j_i + [\zeta_i + (P_i - 1)T] R - T \nabla \cdot j_i \right]
\]

representing especially the Joule, Thomson and Peltier heating rates as well as the recombination heat (cf. [Alb,Kel,Wac]). This form of the evolution equations was the starting point for the implementation of the energy model in WIAS-TeSCA. The heat flow equation can be replaced by other balance equations, too.

Firstly, defining the energy flux \( j_u = j_q + \sum_{i=1}^{2} (\zeta_i + P_i T) j_i \), the energy balance equation

\[
\frac{\partial u}{\partial t} + \nabla \cdot j_u = -\psi \nabla \cdot (j_1 - j_2)
\]
is obtained. If we introduce the total energy density $\tilde{u} = u + \frac{\varepsilon}{2} |\nabla \psi|^2$ and the total energy flux $j_{\tilde{u}} = j_u + \psi \partial D/\partial t$, $D = -\varepsilon \nabla \psi$, then the last equation becomes a conservation law,

$$\frac{\partial \tilde{u}}{\partial t} + \nabla \cdot j_{\tilde{u}} = 0 .$$

(4)

Concerning this result our approach differs from those given in [BS,Kel,Wac].

Secondly, defining the entropy flux $j_s = j_q/T + \sum_{i=1}^{2} P_i j_i$, the entropy balance equation

$$\frac{\partial s}{\partial t} + \nabla \cdot j_s = \sigma$$

(5)

is derived where $\sigma$ denotes the entropy production rate,

$$T \sigma = \sum_{i=1}^{2} \frac{1}{\sigma_i} |j_i|^2 + \frac{1}{\lambda T} |j_q|^2 + \sigma_R = \sum_{i=1}^{2} \sigma_i |\nabla \zeta_i + P_i \nabla T|^2 + \frac{\lambda}{T} |\nabla T|^2 + \sigma_R$$

$$= -T \sum_{i=1}^{2} j_i \cdot \nabla \frac{\zeta_i}{T} + T j_u \cdot \nabla \frac{1}{T} + \sigma_R$$

$$= -\sum_{i=1}^{2} j_i \cdot \nabla \zeta_i - j_s \cdot \nabla T + \sigma_R, \quad \sigma_R = \rho \left( e^{(\zeta_1 + \zeta_2)/T} - 1 \right) (\zeta_1 + \zeta_2) .$$

Obviously $\sigma \geq 0$ holds, and $\sigma = 0$ if and only if $\nabla \zeta_i = 0$, $\nabla T = 0$, $\zeta_1 + \zeta_2 = 0$ (thermodynamic equilibrium). The Onsager relations are valid if either the fluxes $(j_1, j_2, j_u)$ and the generalized forces $(\nabla [\zeta_1/T], \nabla [\zeta_2/T], -\nabla [1/T])$, or the fluxes $(j_1, j_2, j_s)$ and the generalized forces $(\nabla \zeta_1, \nabla \zeta_2, \nabla T)$ are used.

The equations (4), (5) reflect the First Law and Second Law of Thermodynamics in differential form. In order to get their integral form we introduce the functionals of total energy and total entropy,

$$U(n_1, n_2, T) = \int_{\Omega} \left[ \frac{\varepsilon}{2} |\nabla \psi|^2 + u(n_1, n_2, T) \right] \, dx + \int_{\Gamma_N} \frac{\tau}{2} \psi^2 \, d\Gamma ,$$

$$S(n_1, n_2, T) = \int_{\Omega} s(n_1, n_2, T) \, dx .$$

Assuming that the system is thermodynamically closed ($j_i \cdot \nu = 0$, $j_q \cdot \nu = 0$ on $\Gamma$) we obtain that

$$\frac{dU}{dt} = \int_{\Gamma_D} \varepsilon \nabla \frac{\partial \psi}{\partial t} \cdot \nu \, \psi^D \, d\Gamma + \int_{\Gamma_N} \tau \psi \frac{\partial \psi^N}{\partial t} \, d\Gamma , \quad \frac{dS}{dt} = \int_{\Omega} \sigma \, dx$$

along any solution of (3), (1), (2). The second equation shows that the negative total entropy $-S$ is a Lyapunov function of the evolution system. The
first equation implies that the total energy is preserved if $\psi^D = 0$ and $\psi^N$ does not depend on time.

These important properties of the evolution system remain valid for the corresponding discrete system what has been achieved by using sophisticated discretization schemes with respect to time and space coordinates. Thus the energy model implemented in WIAS-TeSCA turns out to be well posed from the thermodynamic point of view. From the mathematical point of view the evolution system is very complicated, and there are known only few results, at least under realistic assumptions on the state equations and kinetic coefficients. For the stationary case in [Gri] new existence and uniqueness results are obtained using the Implicit Function Theorem and properties of linear elliptic operators in a suitable scale of Sobolev–Campanato spaces.

5 Process Modelling

One of the main process steps in manufacturing semiconductor devices is the redistribution of dopants connected with or followed after the doping. In order to simulate this process, different models have been developed. Nowadays so-called pair diffusion models are preferred [Dun, GH, Höf, Str]. They consist in a set of reaction-diffusion equations for a lot of electrically charged species $X_1, \ldots, X_m$ (dopants, point defects, dopant–defect pairs, see

![Diagram](image)

**Fig. 6.** Species and reactions in a variant of pair diffusion models
Fig. 6). The kinetics of electrons and holes is assumed to be very fast. Then the electrochemical potential of the electrons does not depend on the spatial coordinates, and for determining their chemical potential, denoted by $\psi$ again, a boundary value problem for a nonlinear Poisson equation is obtained,

$$-\nabla \cdot (\varepsilon \nabla \psi) = f + e(\psi) + \sum_{i=1}^{m} n_i q_i(\psi) , \nabla \psi \cdot \nu \mid \Gamma = 0 . \quad (6)$$

Here $f$, $e(\psi)$, $n_i$ and $q_i(\psi)$ denote a fixed background doping, the charge density of electrons and holes, the density of the species $X_i$ and its charge number depending on $\psi$, respectively.

The initial boundary value problem for the reaction-diffusion system is

$$\frac{\partial n_i}{\partial t} + \nabla \cdot j_i = - \sum_{(\alpha, \beta) \in \mathcal{R}} (\alpha_i - \beta_i) R_{\alpha\beta} , \quad j_i \cdot \nu \mid \Gamma = 0 , \quad n_i(0) = N_i \quad (7)$$

where the particle fluxes $j_i$ and reaction rates $R_{\alpha\beta}$ are given by

$$j_i = -D_i(\psi) \left[ \nabla n_i + n_i q_i(\psi) \nabla \psi \right] ,$$

$$R_{\alpha\beta} = k_{\alpha\beta}(\psi) \left[ \prod_{i=1}^{m} \alpha_i^{a_i} - \prod_{i=1}^{m} \beta_i^{a_i} \right] , \quad a_i = \frac{n_i}{p_i(\psi)} , \quad p_i(\psi) = \overline{n}_i e^{-\int_{0}^{\psi} q_i(s) \, ds} .$$

Here the vector $(\alpha, \beta) = (\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m)$ represents the stoichiometric numbers of mass action type reactions of the form

$$\alpha_1 X_1 + \cdots + \alpha_m X_m \rightleftharpoons \beta_1 X_1 + \cdots + \beta_m X_m , \quad (\alpha, \beta) \in \mathcal{R} .$$

The kinetic coefficients $D_i(\psi)$ and $k_{\alpha\beta}(\psi)$ are generally positive (only for the dopants often $D_i = 0$ is proposed). Finally, $\overline{n}_i$ is a positive reference density, $N_i$ is the non-negative initial density of the $i$-th species.

If the charge numbers do not depend on $\psi$ we arrive at a model that we have studied in a previous project [HGGR]. In the more general situation considered now the mathematical investigation starts from thermodynamic principles, again. Since we deal only with isothermal processes, the energy functional that has to be chosen is the total free energy [HG],

$$F(n) = \int_{\Omega} \left\{ \frac{\varepsilon}{2} |\nabla \psi|^2 dx + \int_{0}^{\psi} [e(s) - e(\psi)] \, ds + \sum_{i=1}^{m} n_i \int_{0}^{\psi} [q_i(s) - q_i(\psi)] \, ds \right. \right.$$

$$+ \left. \sum_{i=1}^{m} \left[ n_i \left( \ln \frac{n_i}{\overline{n}_i} - 1 \right) + \overline{n}_i \right] \right\} \, dx$$

where $\psi$ is the solution of (6) for prescribed densities $n = (n_1, \ldots, n_m)$. Under some assumptions concerning the initial values and the structure of
the underlying reaction system we have proved in [HG] the following results. The time derivative of the free energy fulfills the inequality

$$\frac{dF}{dt} \leq - \int_\Omega d \, dx$$

along any solution of (7), (6), $d$ is a lower estimate of the dissipation rate,

$$d = \sum_{i=1}^{m} 4 D_i(\psi) \rho_i(\psi) |\nabla \sqrt{\alpha_i}|^2 + \sum_{(\alpha, \beta) \in \mathcal{R}} 2 k_{\alpha\beta}(\psi) \left| \prod_{i=1}^{m} \sqrt{\alpha_i^{-\alpha_i}} - \prod_{i=1}^{m} \sqrt{\alpha_i^{-\beta_i}} \right|^2.$$

Since $d \geq 0$ the free energy is a Lyapunov function of the evolution system. Moreover, there exists a steady state $n^*$ (with corresponding potential $\psi^*$) which is uniquely determined in the class

$$\left\{ n: \int_\Omega (n - N) \, dx \in \text{span} \left\{ \alpha - \beta: (\alpha, \beta) \in \mathcal{R} \right\} \right\}.$$

The difference $F(n) - F(n^*)$ decays exponentially to zero if the time tends to infinity. Based on these properties of the free energy further a priori estimates and existence results can be derived. First results are obtained in [MGHP].

References


Finding Beneficial DAE Structures in Circuit Simulation

Roswitha März\textsuperscript{1}, Diana Estévez Schwarz\textsuperscript{1}, Uwe Feldmann\textsuperscript{2}, Sandra Sturtzel\textsuperscript{1}, and Caren Tischendorf\textsuperscript{1}

\textsuperscript{1} Institute of Applied Mathematics, Humboldt-University Berlin, Germany
\textsuperscript{2} Infineon Technologies AG, Germany

Abstract. Circuit simulation is a standard task for the computer-aided design of electronic circuits. The transient analysis is well understood and realized in powerful simulation packages for conventional circuits. But further developments in the production engineering lead to new classes of circuits which may cause difficulties for the numerical integration. The dimension of circuit models can be quite large ($10^5$ equations). The complexity of the models demands a higher abstraction level. In this paper, we analyze electric circuits with respect to their structural properties. We discuss the relevant subspaces of the resulting differential algebraic equations (DAEs) and present algorithms for calculating the index as well as consistent initial values.

1 Introduction

The modern simulation of electric networks is based on modelling techniques that allow an automatic generation of the model equations. One of the mostly used techniques is the Modified Nodal Analysis (MNA). It leads to Differential-Algebraic Equations (DAEs)

$$f(x'(t), x(t), t) = 0.$$ (1)

where $f'_x$ is in general singular. For the numerical solution of these special systems arising from circuit simulation, Gear [8] proposed the BDF (Backward Difference Formulae). Their robustness and reliability have made the BDF methods to become a standard in simulation packages up to now. However, they failed in certain situations. The study of the theory of DAEs (cf. e.g. [9], [1], [12]) has shown that DAE systems represent not only integration problems but may also involve differentiation problems. DAEs of an index $\geq 2$ provide solutions including inherent differentiations of the input signals (cf. [10], [15], [16], [6]). These inherent differentiations are not always properly reflected by the BDF methods. Consequently, the variable stepsize BDF usually do not work well for index-3 problems (see e.g. [13]). In the index-2 case, convergence and stability results were obtained for a large class of quasilinear DAEs, e.g. [14], [18], [17].

Naturally, in the analysis of DAEs, various subspaces (and projectors) as e.g. the leading nullspace $f'_x(x', x, t)$ and the tangent spaces to the obvious

W. Jäger et al. (eds.), Mathematics - Key Technology for the Future
© Springer-Verlag Berlin Heidelberg 2003
and hidden manifolds play an important role. Relevant results – from index characterization up to error propagation – are usually given in terms of those subspaces and projector matrices.

The study of various circuit examples led to the hypothesis that the space of the algebraic components as well as the space of the index-2 components are constant for standard circuit simulation problems. Under these assumptions, convergence and weak instability of the BDF methods is satisfied for index-2 problems [18]. The detailed analysis of circuit systems that we present here enables us to prove the hypothesis to be true for a wide class of circuits (restricted by some conditions for controlled sources, see [4]). Additionally we see that the circuit systems have a special structure that provides fast algorithms for an index check basing on network graphs. This result is really surprising since all other attempts for index tests (using numerical linear algebra) were not always reliable and prohibitively expensive with respect to the circuit simulation itself.

A further nice result of the structural investigations is that in index-2 circuit systems the index-2 components appear only linearly. This can effectively be used for another practical simulation problem, the calculation of consistent initial values. In Sect. 5 we present a fast and reliable algorithm for computing consistent initial values for circuit systems of index 2.

2 Modelling

The index and the structure of the equations we obtain in electric circuit simulation depend, among other things, on the scheme for setting up the equations (cf. [10], [16]). We will restrict ourselves to two different formulations of the most frequently used modelling technique, the Modified Nodal Analysis.

2.1 Conventional Modified Nodal Analysis

Circuits are usually described in terms of all nodal potentials and currents of current controlled elements. The equation system

$$C(x, t)x' + f(x, t) = 0$$

contains

- the KCL (Kirchhoff’s Current Law) equations and
- the constitutive relations of inductors and voltage sources.

Since the matrix $C(x, t)$ is usually singular, the system (2) represents a quasilinear DAE.

\footnote{For the definition of consistent initial values see Sect. 7.}
2.2 Charge Oriented Modified Nodal Analysis

In this case, the circuit description bases additionally on the charges of capacitors and fluxes of inductors \((y)\). The resulting equation system

\[
Ay' + f(x, t) = 0 \\
y = g(x, t)
\]

contains again (3)

- the KCL (Kirchhoff’s Current Law) equations and
- the constitutive relations of inductors and voltage sources

as well as (4)

- the constitutive relations for capacitors and
- the constitutive relations for inductors.

The matrix \(A\) is usually not quadratically but constant. The equation system (3)–(4) represent so a quasilinear DAE with a constant leading coefficient matrix.

**Remark 1.** We obtain formulation (2) if we substitute the expression (4) for \(y\) into equation (3). The close connection between both formulations allows a transfer of numerous results from one to the other. The charge oriented one is preferred in practice since it guarantees charge and flux conservation automatically (cf. [10]).

2.3 Model Assumptions

We deal with lumped circuits consisting of the following one- and multi-port elements wherein \(j\) describes the branch currents and \(v\) denotes the branch voltages.

- Passive elements and independent sources
  - Capacitors with the constitutive relation \(j = \frac{d}{dt}q_C(v, t)\)
  - Inductors with the constitutive relation \(v = \frac{d}{dt}\phi_L(j, t)\)
  - Resistors with the constitutive relation \(j = r(v, t)\)
  - Independent current sources described by \(j = i(t)\)
  - Independent voltage sources described by \(v = v(t)\)
- Controlled sources
  - Controlled current sources described by \(j = i(j, v, t)\)
  - Controlled voltage sources described by \(v = v(j, v, t)\)

that satisfy certain assumptions (specified in [6]) concerning their location in the network and the kind of controlling variables.

Furthermore, the capacitance matrix \(\frac{\partial q_C(v, t)}{\partial v}\), the inductance matrix \(\frac{\partial \phi_L(j, t)}{\partial j}\), and the conductance matrix \(\frac{\partial r(v, t)}{\partial v}\) are assumed to be positive definite.\(^2\)

\(^2\) For capacitors and inductors with affine characteristics the positive definiteness implies that they are strictly locally passive (cf. [7]).
2.4 Detailed Analysis of the Charge Oriented MNA

Splitting the incidence matrix and considering the special properties of the resulting matrices provides a deeper insight into the structure of the equations. If $A_C$, $A_R$, $A_L$, $A_V$ and $A_I$ denote the incidence matrices of the capacitive, resistive, inductive, voltage source and current source branches, respectively, then the charge oriented MNA leads to a system of the form:

$$
A_C \frac{dq}{dt} + A_R r(A_R^T e, t) + A_L j_L + A_V j_V + A_I i(e, j_L, j_V, t) = 0,
$$

$$
\frac{d\phi}{dt} - A_L^T e = 0,
$$

$$
A_V^T e - v(e, j_L, j_V, t) = 0,
$$

$$
q - q_C(A_C^T e, t) = 0,
$$

$$
\phi - \phi_L(j_L, t) = 0.
$$

Here, $e$ denotes the nodal voltages, $j_L$ the currents of inductive branches and $j_V$ the currents of voltage source branches. We have summarized the characteristics of independent and controlled current sources by $i(\cdot)$. Correspondingly, the independent and controlled voltage sources are represented by $v(\cdot)$.

Due to Kirchhoff's laws, cutsets of current sources and loops of voltage sources are forbidden. This implies for the element related incidence matrices that

- the matrix $(A_C, A_R, A_V, A_L)$ has full row rank and
- the matrix $A_V$ has full column rank.

3 Index Classification

A long experience in circuit simulation has shown that cutsets of inductors and current sources as well as loops of capacitors and voltage sources may lead to difficulties in the transient analysis. It turns out that these network configurations play (as expected) an essential role in the index classification of networks. Before formulating the result, we introduce some projectors that allow a simple mathematical description of such cutsets and loops. Let

- $Q_C$ be a projector onto ker $A_C^T$,
- $Q_{V-C}$ a projector onto ker $Q_C^T A_V$, and
- $Q_{CRV}$ a projector onto ker $(A_C, A_R, A_V)^T$.

Then, the following network characterization is possible.

**Lemma 2.** 1. If the network does not contain an L–I cutset (a cutset of inductors and/or current sources only (cf. Fig. 1)) then the matrix $(A_V, A_R, A_C)^T$ has full column rank and it holds that $Q_{CRV} = 0$. 
2. If the network does not contain a C–V loop (a loop of capacitors and voltage sources (cf. Figure 1)) then the matrix $\tilde{Q}_C^T A_V$ has full column rank and it holds that $\tilde{Q}_{V-C} = 0$.

Theorem 3. If the controlled sources satisfy certain topological conditions (specified in [6]) then the DAE system (3)–(4) has an Index $\leq 2$. The index is equal 2 if and only if the network contains an L–I cutset or a C–V loop.

A detailed proof of this theorem is presented in [6].

![Fig. 1. Example for an L–I cutset (left) and a C–V loop (right)](image)

Remark 4. 1. Controlled sources in L–I cutsets or C–V loops may lead to arbitrarily high index systems (cf. [11]). Our assumptions for controlled sources (specified in [6]) exclude controlled sources in L–I cutsets or C–V loops.

2. The numerically unstable index-2 components consist of:
   - branch voltages of the inductors and current sources of L–I cutsets
   - branch currents through voltage sources of C–V loops.

4 Graph-Theoretical Determination of the Hidden Constraints

For the charge-oriented MNA, the hidden constraints result from

$$\frac{d}{dt} \left( \tilde{Q}_V^T A_V P C e - \tilde{Q}_V^T \tilde{C}_V v(t) \right) = 0,$$  \hspace{1cm} (5)

$$\frac{d}{dt} \left( Q_{CRV}^T A_{CL} j_L + Q_{CRV}^T A_{IL} i_t(t) \right) = 0,$$  \hspace{1cm} (6)

and

$$\frac{d}{dt} (q - q_C(A^T e, t)) = 0,$$

$$\frac{d}{dt} (\phi - \phi_L(j_L, t)) = 0.$$
Here, the functions \( i_t \) and \( v_t \) involve only independent current and voltage sources. We are interested in expressions for (5)–(6) without requiring the computation of the corresponding projectors. In [2] it was shown that these equations can be obtained directly from the network by making use of the following two procedures that analyze its graph. In fact, these procedures precisely determine the linearly independent equations that describe the hidden constraints arising from C–V loops and L–I cutsets, respectively.

Let us first focus on the constraints (5) arising from the C–V loops. Recall that

\[
A_T^{T} e - v(\cdot) = 0 \tag{7}
\]

are the characteristic equations of the voltage sources. Since \( \bar{Q}_{V - C} \) describes the C–V loops, it results that

\[
\bar{Q}_{V - C} A_T^{T} e - \bar{Q}_{V - C} v_t(t) = 0 \tag{8}
\]

corresponds to the sums of the characteristic equations of the voltage sources that form a part of the C–V loops (cf. [2]). More exactly, these equations can be determined by means of the following procedure.

**Procedure 1**

1. Search a C–V loop in the given network graph. If no C–V loop is found, then end.
2. Write the equation resulting from the sum of the derivatives of the characteristic equations of the voltage sources contained in the C–V loop, taking into account the orientation of the loop and the reference direction of the considered branches.
3. Form a new network graph by deleting the branch of one voltage source that forms a part of the loop, leaving the nodes unchanged.
4. Return to 1, considering the new network graph.

Let us now focus on the constraints (6) arising from L–I cutsets. To get an idea of where they arise from, recall that

\[
A_C \frac{dq}{dt} + A_{Rr}(A_R^{T} e, t) + A_{Lj_L} + A_{V j_V} + A_{I} i(\cdot) = 0 \tag{9}
\]

are the nodal equations. That means, these equations arise from KCL for the L–I cutsets. Consequently, the equations corresponding to (6) can be determined by means of the following procedure that starts again considering the original graph (cf. [2]).

**Procedure 2**

1. Search an L–I cutset. If one is found, then select an arbitrary inductor of this cutset. Realize that we can always find such an inductor because cutsets of current sources only are forbidden. If no L–I cutset is found, then end.
2. Write a new equation resulting by differentiation of the cutset equation arising from 1.
3. Delete the chosen inductor from the network contracting its incident nodes.
4. Return to 1, considering the new network graph.

5 **Consistent Initialization**

Using the graph-theoretical description of the hidden constraints, it is possible to calculate consistent initial values at relatively low costs. We obtain a consistent initial value starting up from a possibly inconsistent one \(x^0\) that fulfills the system's equations (for instance, the DC-operating point) in the following way (cf. [3]):

1. Add additional currents that flow through the C–V loops as a consequence of the hidden constraints described by Procedure 1 to the values of the currents through the branches that form a part of C–V loops.
2. Add convenient values to the node potentials to fulfill the additional voltage across the L–I cutsets defined by the hidden constraints described by Procedure 2.

Practical advantage of this approach is that the values that have to be added result by solving a linear system. This system consists of a part of the original DAE together with the equation obtained by Procedure 1 and Procedure 2. The graph-theoretical determination of the relevant equations is very fast with respect to the transient analysis.

6 **Computational Aspects and Practical Results**

The methods suggested were implemented in the simulation package TITAN.\(^3\) For this purpose, a graph oriented algorithm was developed that provides important features for several aspects:

1. Index determination (cf. Theorem 3).
2. Identification of critical variables: the currents through voltage sources that form a part of C–V loops and the branch voltages of branches that form a part of L–I cutsets.
3. Identification of the circuit elements and nodes which form the critical circuit configurations; this provides valuable information for the user how to regularize higher index configurations in case of problems.
4. Specification of smoothness requirements: Additional smoothness has only to be given for the constitutive relations of the current sources and inductors that form a part of L–I cutsets, and for the characteristic equations of the voltage sources and capacitors that form a part of C–V loops.
5. Description of the linear system that provides the values required for the computation of a consistent initialization.

\(^3\) Infineon Technologies AG (formerly SIEMENS AG).
6.1 Index Classification

The code was tested for a variety of artificial test circuits and real designs. For some typical MOS circuits without controlled sources, the results are given in Table 1.

<table>
<thead>
<tr>
<th>circuit</th>
<th>transistors</th>
<th>equations</th>
<th>C–V loops</th>
<th>nodes in C–V loops</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS ringoscillator</td>
<td>134</td>
<td>73</td>
<td>3</td>
<td>71</td>
<td>&lt; 10⁻²</td>
</tr>
<tr>
<td>16 bit adder</td>
<td>544</td>
<td>283</td>
<td>5</td>
<td>279</td>
<td>0.01</td>
</tr>
<tr>
<td>1MBit DRAM</td>
<td>2005</td>
<td>1211</td>
<td>21</td>
<td>1189</td>
<td>0.02</td>
</tr>
<tr>
<td>16MBit DRAM</td>
<td>5208</td>
<td>3500</td>
<td>73</td>
<td>3427</td>
<td>0.11</td>
</tr>
<tr>
<td>ALU</td>
<td>13005</td>
<td>32639</td>
<td>77</td>
<td>29626</td>
<td>6.76</td>
</tr>
</tbody>
</table>

Table 1. Index test for MOS circuits without controlled sources

We observe that all of these circuits are of index 2 due to the existence of C–V loops, and almost all circuit nodes are part of these loops. Furthermore we see that even for large circuits the CPU times (which are measured in sec on a 350 MHz SUN workstation) are very moderate, and the same is true for the circuit examples of Table 2, which contain explicitly given controlled sources.

<table>
<thead>
<tr>
<th>circuit name</th>
<th>equations</th>
<th>controlled sources</th>
<th>index</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ringmo</td>
<td>42</td>
<td>8</td>
<td>= 1</td>
<td>&lt; 10⁻²</td>
</tr>
<tr>
<td>sq3bogner</td>
<td>234</td>
<td>3</td>
<td>= 2</td>
<td>&lt; 10⁻²</td>
</tr>
<tr>
<td>fischer</td>
<td>2494</td>
<td>10</td>
<td>?</td>
<td>0.04</td>
</tr>
<tr>
<td>xsection</td>
<td>13465</td>
<td>958</td>
<td>≥ 2</td>
<td>0.88</td>
</tr>
<tr>
<td>clara</td>
<td>35979</td>
<td>2</td>
<td>= 2</td>
<td>1.85</td>
</tr>
<tr>
<td>teethmi</td>
<td>174881</td>
<td>39</td>
<td>≥ 2</td>
<td>167.01</td>
</tr>
</tbody>
</table>

Table 2. Index test for circuits including controlled sources

Note that for the latter class of circuits any index is possible. For some of the circuit examples the index diagnosis is not sharp or even fails. This is due to one of the following reasons:

- The catalogue of conditions specified in [6] is not yet fully implemented.
- The conditions given in [6] are sufficient conditions for the index to be ≤ 2; the particular circuit configuration considered here is not (yet) included there.
In this circuit configuration, the index depends on actual parameter values; in this case structural index diagnosis cannot be successful.

We should mention that controlled sources are implicitly included in any semiconductor device model. Fortunately, the conditions given in [6] state that these controlled sources are not “dangerous” at all, provided that the positive definiteness of the element stamps mentioned before is satisfied. This can be watched during analysis with local numerical checks.

6.2 Consistent Initialization

Computational aspects played an important role for the particular development of the methods:

1. Special care was taken to use only those variables which are available in the circuit simulator anyway. The only quantities which had to be extra coded were the time derivatives of the input source functions \( i(t), v(t) \).
2. The algorithm was implemented as an add-on in the simulation flow. In case of index-2 structures the consistent initial values are computed in an extra step from the standard initial values \( x^0 \) (“DC-operating point”), which are computed anyway before transient analysis is started.
3. Since the structure of the equations of the resulting linear system is similar to the structure of the original system, part of them can be solved as a structural subset of the original system, taking advantage of the sparse handling. This turns out to be necessary especially in case of C–V loops, which usually cover almost all circuit nodes, see Table 1. Note that this may have an impact on the choice of pivot elements for the linear sparse solver, which are usually selected statically in a preprocessing step in circuit simulation.

As an example, we look at the circuit given in Fig. 1 which contains a C–V loop and hence is of index 2. With \( j_v \) being the branch current of the voltage source flowing from node 1 to node 2, the MNA equations are:

\[
\begin{align*}
    j_v + C_1 \cdot e'_1 &= 0, \\
    -j_v + C_2 \cdot e'_2 + 1/R \cdot e_2 &= 0, \\
    e_1 - e_2 &= v(t).
\end{align*}
\]

The DC-operation point defined by \( e'_1 = e'_2 = 0 \) \( \Rightarrow \) \( e_1 = v(0), \ e_2 = 0, \ j_v = 0 \) satisfies these equations, but may violate the hidden constraint

\[
e'_1 - e'_2 = v'(0).
\]
For computing a consistent initial solution we assume an additional current \( \dot{j}_v \) to flow in the C–V loop, and get

\[
\begin{align*}
\dot{j}_v + C_1 \cdot e'_1 &= 0, \\
-\dot{j}_v + C_2 \cdot e'_2 &= 0, \\
e'_1 - e'_2 &= v'(0).
\end{align*}
\]

Adding the solution to the standard DC-solution, we obtain the following consistent initial values:

\[
\begin{align*}
e'_1 &= \frac{C_2}{C_1 + C_2} v'(0), \\
e'_2 &= -\frac{C_1}{C_1 + C_2} v'(0), \\
e_1 &= v(0), \\
e_2 &= 0, \\
\dot{j}_v &= -\frac{C_1 C_2}{C_1 + C_2} v'(0)
\end{align*}
\]

In practice, the computation of a consistent initialization has been carried out with regard to the following aspects:

1. To start the integration, i.e., in general, the DC operating point is corrected in order to obtain a consistent starting point.
2. To obtain consistent values after discontinuities of the derivatives of the input functions, i.e., at the breakpoints.
3. For a clean handling of user given initial conditions by calculating an appropriate \( x^0 \) (cf. the approach presented in [3]).

As a practical example shows Fig. 2 the waveforms of two index-2 variables of an MOS circuit (word line booster with 124 MOS transistors and 108 equations).

Conventional (dashed) and consistent (solid) waveform differ significantly at the beginning of transient analysis and at the breakpoints, where the slope of input signals changes abruptly. The additional expense for consistent initialization is equivalent to one additional Newton iteration at the beginning and at each breakpoint.

7 Mathematical Background

In this section we summarize new results for specially structured DAEs as described below. The particular properties of the MNA equations guarantee that DAEs arising from circuit simulation have this special structure.

Consider quasilinear DAEs

\[
A(x, t)x' + b(x, t) = 0, \tag{10}
\]
Fig. 2. Conventional (dashed) and consistent (solid) waveforms for index-2 variables of an MOS circuit

fulfilling that $\ker A(x,t)$ and $\text{im} A(x,t)$ are constant. Then, we find constant projectors

$$Q \text{ onto } \ker A(x,t), \quad P := I - Q, \quad \text{and } W_0 \text{ along } \text{im} A(x,t).$$

Observe that each solution $x(t)$ of (10) belongs to

$$M_0(t) := \{ z : W_0 b(z, t) = 0 \}, \quad (11)$$

that implies that the choice of initial values for the integration of (10) is restricted. According to ODE theory, we define for DAEs:

**Definition 5.** A vector $x_0 \in \mathbb{R}^n$ is a consistent initial value of (10) if there exists a solution of (10) that fulfills $x(t_0) = x_0$.

In practice, we are also interested in the corresponding values of the derivatives appearing in the DAE. The following definition will characterize these values properly.

**Definition 6.** A vector $(x_0, P y_0)$ is a consistent initialization of (10), if $x_0$ is a consistent initial value and $(x_0, P y_0)$ fulfills the equation

$$A(x_0, t_0) P y_0 + b(x_0, t_0) = 0. \quad (12)$$

For the index-1 case, there exists always a solution through a point $x_0 \in M_0(t_0)$ [9], i.e., $M_0$ describes exactly the set of consistent initial values. Note that the subspace $\text{im} P$ may be considered to be a practical substitute of the tangent space to $M_0(t_0)$. For suitably given $x^0$, condition

$$P x_0 = P x^0 \quad (13)$$
fixes the free integration constants while $r = \text{rank} A(x, t) = \text{rank} P$ is the dynamical degree of freedom in the index-1 case.

In the higher-index cases, the set of consistent initial values is a proper subset of $M_0$ that is defined by so-called hidden constraints. Let us demonstrate this in detail for index-2 DAEs. Each consistent initial value has to satisfy also the so-called hidden constraint, say equation

$$h(x_0, t_0) = 0. \quad (14)$$

On the other hand, the dynamical freedom decreases in comparison with the index-1 case. By means of an appropriate projector $\Pi = \Pi(x_0, t_0)$ the subspace $\text{im} \, \Pi$ may be considered to be a practical substitute of the tangent space to

$$M_1(t_0) := \{ z \in M_0(t_0), \quad h(z, t_0) = 0 \}. \quad (15)$$

This time, for suitably given $x^0$ the free integration constants are fixed by the condition

$$\Pi(x_0, t_0)x_0 = \Pi(x_0, t_0)x^0. \quad (16)$$

The resulting system (12), (14), (16) to provide a consistent initialization (see e.g. [5]) is somehow difficult to solve. A nice idea ([4]) is now, to use an easily available projector $U$ and to replace the practically very complicated condition (16) by the simple one

$$Ux_0 = Ux^0. \quad (17)$$

The new system (12), (14), (17) is in fact an over-determined extension of (12), (14), (16), supposed $x^0 \in M_0(t_0)$. Then, in the linear case (12), (14), (17) is a consistent linear full rank system and can be solved by least squares.

Let us consider in detail the systems (12), (14), (17) that result if we suppose that the special structure of the MNA equations is given. Suppose that $\text{ker}[A(x, t) + W_0 b'_x(x, t)]$ is constant. Note that this nullspace is identical to $\{0\}$ in case of index-1 DAEs. For index-2 DAEs, this nullspace describes the space of index-2 variables, i.e., the sensitive components with respect to perturbations due to the inherent differentiation problem. We want to assume this space not to be time dependent and even more not to be solution dependent that seems to be natural. For a convenient description we choose constant projectors

$$T \quad \text{onto} \quad \text{ker}[A(x, t) + W_0 b'_x(x, t)] \quad \text{and} \quad U := I - T.$$ 

Furthermore, we define a projector

$$W_1(x', x, t) \quad \text{along} \quad A(x, t) + ([A(x, t)x']_x + b'_x(x, t))Q.$$ 

As a consequence, it results

$$W_1(x', x, t) = W_1(Ux, t) \quad \text{and} \quad W_0 b(x, t) = W_0 b(Ux, t).$$
In other words, the projector $W_1$ as well as the derivative free equations $W_0 b = 0$ are independent of the index-2 components $T x$.

Using the above projectors, our investigations have shown [4] that the equations obtained by means of MNA are quasilinear DAEs even of the form

$$A(Ux, t)x' + \tilde{b}(Ux, t) + BTx = 0$$

(18)

for a constant matrix $B$ if the controlled sources of a network satisfy certain conditions specified (cf. [3],[4]), i.e., the index-2 components appear only linear in the DAE system. Obviously, all solutions $x(t)$ of (18) have to satisfy the algebraic relations

$$W_0 \tilde{b}(Ux(t), t) = 0$$

and additionally its derivative$^4$

$$\frac{d}{dt} W_0 \tilde{b}(Ux(t), t) = 0.$$  

(19)

It results for the specific structure that the hidden constraints arise only from the $W_1$-part of (19), i.e., from

$$W_1(Ux, t)(W_0 \tilde{b})'_x(Ux, t)Px' + W_1(Ux, t)(W_0 \tilde{b})'_t(Ux, t) = 0.$$  

Correspondingly, the set of consistent initial values for DAE systems (18) is given by

$$M_1(t) := \left\{ z : \exists y \quad A(Uz, t)y + \tilde{b}(Uz, t) + BTz = 0, \right.$$ 

$$\left. \left[ (W_0 \tilde{b})'_x(Uz, t)y + (W_0 \tilde{b})'_t(Uz, t) \right] = 0 \right\}.$$  

As a consequence, consistent initial values can be computed as follows:

**Theorem 7.** [4] Suppose that some values $(x^0, y^0)$ fulfilling

$$A(Ux^0, t_0)y^0 + \tilde{b}(Ux^0, t_0) + BTx^0 = 0$$

are given. We obtain a consistent initialization $(x_0, y_0)$ starting up from $(x^0, y^0)$ setting $Ux_0 := Ux^0$, computing the unique solution $(\dot{x}_0, y_0)$ of the linear system

$$A(Ux_0, t_0)\dot{y}_0 + BT\dot{x}_0 = 0,$$

$$U\dot{x}_0 = 0,$$

$$W_1(Ux_0, t_0)(W_0 \tilde{b})'_x(Ux_0, t_0)P[y^0 + \dot{y}_0]$$

$$+ W_1(Ux_0, t_0)(W_0 \tilde{b})'_t(Ux_0, t_0) = 0,$$

$^4$ Actually, smoothness is not necessary for complete $W_0 \tilde{b}$. A detailed analysis of smoothness requirements can be found in [4].
426 R. März et al.

and setting

\[ x_0 = x_0^0 + \hat{x}_0, \]
\[ P y_0 = P y_0^0 + P \hat{y}_0. \]

This corresponds to the approach described in Sect. 5. The realization was achieved making use of the fact that for the equations of the MNA the projectors \( Q, W_0, T, W_1 \) can be expressed in terms of the projectors \( Q_C, \bar{Q}_{V-C}, Q_{CRV} \) introduced in Sect. 3.

8 Conclusion

Exploitation of the special structure of network equations is beneficial for applying recent results of DAE theory to circuit simulation problems of industrial relevance and complexity. Since the relevant DAE subspaces of the circuit equations can directly be constructed from properly coloured network graphs, it is possible to develop very efficient and reliable methods for

- calculating the DAE index of even very large circuits;
- identifying critical circuit configurations and providing suggestions for their regularization;
- computing consistent initial values;
- implementing a clean handling of user given initial conditions.

A key issue of the approaches presented here is to combine global topological criteria – like the existence of C–V loops – with local numerical criteria – like the positive definiteness of the device stamps – thus combining the speed of graph oriented methods with the generality and reliability of numerical checks.

The algorithms were successfully implemented into an industrial circuit simulator and have proven to be practical for circuits with more than \( 10^5 \) equations. Essential outcomes of this work are, that industry

- has learned how to construct future device and circuit models in order to avoid numerical problems due to too high DAE index as far as possible;
- for the first time, has means to really assess the practical relevance of circuit problems with DAE index > 2.

Actual work is concerned with some desirable extensions of the set of conditions for checking the DAE index, and with industrial efforts to drive the new algorithms into practical use. Furthermore, the methods serve for improvements in the cooperative BMBF projects “Modifizierte ROW-Methoden in der elektrischen Schaltungssimulation” at the TU Karlsruhe, see paper of M. Günther et al. in this issue (⇒ consistent initialization), and “Analyse gemischt analog-digitaler hochoszillierender Schaltungen” at the TU München, see paper of A. Schwarz et al. in this issue (⇒ identification of index-2 configurations, fast construction of projectors onto the relevant subspaces).
Desirable topics for future work in this field are a generalization of the results presented here from topological to structural aspects, the development of practical stability criteria, the structural analysis of circuits including extended semiconductor models and the extension to stochastic DAEs such that noise effects can be efficiently included into transient circuit simulation.

References

VII. Tomography, Image Analysis and Visualisation

Reconstructing Crystalline Structures from Few Images under High Resolution Transmission Electron Microscopy
   P. Gritzmann, S. de Vries

Measurement of Paint Layer Thickness with Photothermal Infrared Radiometry
   A. K. Louis, P. Dörr, C. Gruss, H. Petry

Spatio-Temporal Current Density Reconstruction from EEG-/MEG-Data

Signal Correction in NMR Spectroscopy
   H.-O. Peitgen, T. Boskamp, P. Singer

On Scattering of Ultrasonic Waves
   P. Mathé, J. H. Zacharias-Langhans

Smoothing of Tomographic Data and Hybrid Volume-Surface Visualisation
   W. Jäger, C. Dărțu

Video Coding with Adaptive Vector Quantization and Rate Distortion Optimization
   D. Saupe, M. Wagner
CHORAL – a Charge-Oriented Algorithm for the Numerical Integration of Electrical Circuits

P. Rentrop¹, M. Günther¹, M. Hoschek² *, and U. Feldmann³

¹ Universität Karlsruhe (TH), Fachbereich Mathematik, Institut für Wissenschaftliches Rechnen und Mathematische Modellbildung (IWRMM), Engesserstr. 6, 76128 Karlsruhe, Germany
² IBM Unternehmensberatung, Lyoner Strasse 13, 60528 Frankfurt, Germany
³ Infineon Technologies, Otto-Hahn-Ring 6, 81739 München, Germany

Abstract. Circuit simulation packages generate the network equations automatically. In time domain analysis this results in a system of differential-algebraic equations that is solved numerically by BDF schemes and/or the trapezoidal rule. CHORAL, a charge-oriented Rosenbrock–Wanner method, has been developed as an alternative approach for digital circuits. By its successful implementation into TITAN, the circuit simulator of Infineon Technologies, a second integration scheme is available for the first time. Results for benchmarks and industrial circuits show that CHORAL is competitive with the standard ansatz and possesses advantages for oscillatory circuits.

1 Numerical Circuit Analysis in Industrial Applications

Progress in microelectronics is highlighted by milestones in chip technology, i.e. microprocessor and memory chips. This ongoing increase in performance and density – accompanied by decreasing prices – was possible by technological progress and by extensive use of CAD tools, especially circuit simulation.

In circuit simulators the time domain analysis mode calculates the time-dependent (transient) behavior of electrical signals responding to time varying input signals. The numerical circuit analysis is based on the network approach, see Hoschek et al. [9], which has proven to be a powerful tool in computer-aided analysis of various technical systems. A typical data flow of a circuit simulator like TITAN, the circuit simulator of Infineon Technologies, is outlined in Fig. 1:

- A network description of the circuit is generated automatically in computer-aided electronics-design systems (CAE) from designer's drafts or fabrication data files. To end up with a network model of the circuit, semiconductor devices like MOS transistors are replaced by network companion models that are built up of basic elements [5], as shown in Fig. 2. To

* Markus Hoschek expresses his sincere thanks to Siemens AG for awarding him a Ph.D. grant.
Fig. 1. Data flow of circuit simulation

give an example, we use the NAND-gate circuit shown in top of Fig. 3 that performs the logical Not-AND operation of two input signals $V_{in1}$ and $V_{in2}$. The network description of this system is shown in bottom of Fig. 3, together with the corresponding SPICE [10] syntax for the used elements.

- Using Modified Nodal Analysis (MNA), an input processor translates this network description into a data format reflecting the mathematical model of the system. The mathematical network equations are based on the application of basic physical laws like Kirchhoff’s equations onto the network and insertion of the characteristic equations for the network elements.

Fig. 2. Cross section of an MOS-transistor with associated companion network model
- The network equations are solved in the analysis kernel by numerical integration schemes adapted to the type of the network equations.
- Finally, the simulation results are written into an output file, which can be analyzed by a graphical postprocessor.

The core of any circuit simulator is its analysis kernel which consists of the generation and numerical solution of the network equations, and model routines describing the characteristic relations of the basic elements.

![Circuit Diagram](image)

Fig. 3. NAND-gate: Circuit (top) and network description in a SPICE like language (bottom)

<table>
<thead>
<tr>
<th>network element</th>
<th>notation</th>
<th>nodes</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>capacitor</td>
<td>Cxxxxxxx</td>
<td>N+ N-</td>
<td>value in Farad</td>
</tr>
<tr>
<td>indep. voltage source</td>
<td>Vxxxxxxx</td>
<td>N+ N-</td>
<td>value in Volt</td>
</tr>
<tr>
<td>MOS-transistor</td>
<td>Mxxxxxxx</td>
<td>ND NG NS NB</td>
<td>model reference, length, width in meter</td>
</tr>
</tbody>
</table>
2 The Analysis Kernel of a Circuit Simulator

As an alternative to conventional schemes, which describe dynamic effects by capacitors and inductors, a charge/flux oriented formulation has become a standard to build up the network equations due to its property to guarantee charge conservation during simulation. The automatic modelling approach preserves the topological structure of the network and therefore does not aim at state systems with a minimal set of unknowns. Hence an initial-value problem of differential-algebraic equations (DAEs) is generated, which reads

$$\mathbf{F}(\dot{\mathbf{q}}, \mathbf{x}, t) = A \cdot \begin{pmatrix} \dot{q}_1(x, t) \\ \dot{q}_2(x, t) \end{pmatrix} + \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} + \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix} = 0 \quad \text{on} \quad t \in [0, T] \quad (1)$$

with consistent initial values $\mathbf{x}(0) = x_0$. The system (1) can be described as follows: the vector $\mathbf{x} \in \mathbb{R}^n$ of unknowns consists of all node potentials and branch currents through voltage-controlling elements. $A$ denotes an incidence matrix, which describes the circuit topology; so its entries are only 0, +1 and −1. In general, $A$ has not full row rank. The terminal charges $q_1$ and branch-fluxes $q_2$ of all energy-storing elements are nonlinear functions of branch voltages and currents. $f_1$ collects the currents, $f_2$ describes the voltage drop of the voltage-controlling elements. $s_1$ ($s_2$) contains the contributions of the independent current (voltage) sources.

The properties of the network equations depend not only on the type of the circuit and its parameters, but on its modelling assumptions as well [5,6]. The network equations include characteristic time constants of several orders of magnitude (stiff equations) and suffer from poor smoothness properties of transistor model equations.

Since SPICE2 [10], most circuit simulators solve the network equations either with the trapezoidal rule or with backward differentiation formulas (BDF), which are shortly described here: for a time step $h$ from $t_{k-1}$ to $t_k = t_{k-1} + h$ the derivative $\dot{\mathbf{q}}(t_k)$ in (1) is replaced by a backward difference (BDF) operator $\rho q_k$, which is defined by

$$\rho q_k := \frac{\alpha_0}{h} q(x_k) + \frac{1}{h} \sum_{i=1}^{\rho} \alpha_i q(x_{k-i})$$

with real coefficients $\alpha_i$. The numerical solution of the DAE system (1) is thus reduced to the solution of a system of nonlinear equations

$$\mathbf{F}(\rho q_k, \mathbf{x}_k, t_k) = 0.$$ 

Due to the structure of the nonlinear equations the Jacobian (iteration matrix) for Newton’s scheme is

$$\frac{\alpha_0}{h} \cdot \mathbf{F}_{\dot{x}} + \mathbf{F}_x \quad \text{with} \quad \mathbf{F}_{\dot{x}} = A \cdot \frac{\partial \mathbf{q}(x_n)}{\partial x_n}, \quad \mathbf{F}_x = \frac{\partial f(x_n)}{\partial x_n}.$$
This system yields a unique solution if \( \{ \mathcal{F}_x, \mathcal{F}_{2x} \} \) is a regular matrix pencil. For this case, we get even a unique solution for any index, at least theoretically.

BDF methods have been successfully adapted to charge-oriented network problems up to index-2 systems [14]. A number of numerical tricks is used to make BDF more powerful. Some of them are

- A warning for numerical index problems of the step size control can be given by an index monitor. It controls the step size predictions caused by index anomalies [12]. Usually, the implementation of such a monitor yields no difficult problems.
- Additional capacitors are used to regularise the circuit. This may lead to oscillations of high frequency [2].
- The TR-BDF approach combines the advantages of both methods: large timesteps and no loss of energy of the trapezoidal rule (TR) combined with the (moderate) damping properties of BDF [1].
- To gain efficiency, modified Newton methods are used: the Jacobian evaluation and its LR decomposition (which may be expensive for large circuits) are reused for subsequent iterations in the same timestep.
- Parts of a circuit are latent during some steps; this behavior is exploited by bypass strategies or even by multi rate methods. Only the active part of a circuit is integrated accurately using small step sizes. The latent parts can be treated with larger steps.
- Due to the design, circuits possess a natural partition. Different areas of the chip are weakly coupled. Therefore, the large systems may be handled in a multilevel process by solving smaller subsystems with different schemes and/or in parallel on different CPUs.

In professional packages the conventional methods have achieved a high degree of maturity, and have proven to be efficient and very robust in an extremely large variety of applications. Nevertheless there is some motivation to look at alternative schemes also from an industrial point of view:

- The BDF methods are applicable to much more general classes of nonlinear DAEs; can methods be superior, which are definitely constructed for the special linear-implicit nonlinear form (1) of the circuit equations?
- In the charge/flux oriented form of conventional codes, timestep control is difficult, since charge/flux tolerances are not of interest for the user, and extra effort is necessary to derive charge/flux tolerances from the desirable user given node voltage or current tolerances.
- Are there methods with a more natural embedding of timestep control even in charge oriented formulation?
- What about semi-implicit methods in circuit simulation?

3 Implementation of CHORAL in TITAN

Based on the class of Rosenbrock–Wanner schemes, which have been used successfully for solving classical network equations [11], charge-oriented sche-
mes [3,4] have been developed for charge-oriented network equations of type (1) up to index two. The numerical approximation after one time step is given by

\[ x_1 = x_0 + \sum_{i=1}^{s} d_i \kappa_i, \]

with the increments \( \kappa_i \) computed by the linear system

\[
\left( \mathcal{F}_x + h \gamma \mathcal{F}_x \right) |_{x_0} \kappa_i = A \left( q(x_0) - q(a_i) \right) - h \sum_{j=1}^{i} \beta_{ij} [f(a_j) + s(t_0 + h \alpha_j)] - h \sum_{j=1}^{i-1} \beta_{ij} \frac{\partial f}{\partial x}(x_0) \kappa_j - h^2 \tau_i \dot{s}(t_0).
\]

The coefficients of the method are free to fulfill order conditions and guarantee A- and L-stability, resp. See [4] for more details. On account of the low smoothness properties of transistor models, as well as of usually moderate accuracy demands required by the user, an embedded method of order (2)3 seems to be most suitable. The corresponding scheme, CHORAL, has four stages and only three function evaluations. To avoid a constant term in the error estimate due to inconsistent initial values, both methods are chosen to be stiffly accurate [7]. The corresponding coefficient set of CHORAL is given in Table 1.

Since the usual error estimate (see, e.g., [12]) for stiffly-accurate embedded ROW methods is used, a reliable error control and step size selection are offered that are based on node potentials. No additional techniques are necessary to recompute errors from charges into node potentials.

The method is compatible with circuit simulation packages: model evaluation, direct sparse matrix solvers and parallel simulation already implemented in the numerical integration kernel can be used efficiently. Its implementation into the analysis kernel of TITAN is outlined in Fig. 4. A detailed discussion can be found in [8].

---

| \( \gamma \) | 0.5728160624821349 |
| \( \beta_{21} \) | -2.0302139317498051 |
| \( d_1 = \tilde{d}_1 \) | \( \sigma_{21} = \sigma_{31} = \sigma_{41} = 1/\gamma \) |
| \( \beta_{31} \) | 0.2707896390839690 |
| \( d_2 = \tilde{d}_2 \) | \( \sigma_{32} = \sigma_{42} = 0.0 \) |
| \( \beta_{32} \) | 0.1563942984338961 |
| \( d_3 = \tilde{d}_3 \) | \( \sigma_{43} = 1.0 \) |
| \( \beta_{41} \) | 2/3 |
| \( d_4 = 1.0 \) | \( \beta_{42} = 0.08757666432971973 \) |
| \( \alpha_2 = 1.0 \) | \( \beta_{43} = -0.3270593934785213 \) |
| \( \gamma_1 = \gamma \) | \( \gamma_1 = \gamma \) |
| \( \gamma_2 = -2.45739780 \) | \( \gamma_3 = 0 \) |
| \( \gamma_4 = 0 \) | \( \gamma_4 = 0 \) |

**Table 1.** Coefficients for CHORAL
Fig. 4. CHORAL in TITAN

4 Numerical Results

By the successful implementation of CHORAL into TITAN, a second integration scheme is available in an industrial circuit simulation package for the first time. This opens the possibility to gain experience with numerous real life problems. CHORAL has been tested against the BDF approach in TITAN using a bunch of industrial applications: from standard benchmark examples in circuit simulation (LC oscillator and MOS ring oscillator for oscillatory circuits on the one hand, and a 16 bit adder for logical units on the other hand) to real up-to-date industrial applications: critical paths of dynamic memories (DRAMs) of 1 and 16 MBit, and an arithmetic logical unit, the core of a CPU. In these circuits we had to face with 0 up to 13000 transistors, and with 3 up to more than 30000 network equations.

Although the BDF-approach of TITAN had been tuned for more than ten years, CHORAL is comparable to BDF already after a first test implementation, both w.r.t. computational costs in CPU time (see the corresponding results in Table 2 for an HP work station C200 with RISC-processor), and accuracy. One reason for the efficiency of CHORAL seems to be the step size and error control that results in large step sizes and few failures of the step size predictions, see Fig. 5.

In addition, CHORAL preserves well the properties of oscillatory circuits: physical oscillations are preserved, but artificial oscillations of high frequency caused by numerical noise are damped out very rapidly. Compared with BDF methods of lower order, both amplitude and phase error are negligible for
<table>
<thead>
<tr>
<th>Circuit</th>
<th># transistors</th>
<th># equations</th>
<th>CPU time</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CHORAL</td>
<td>BDF</td>
</tr>
<tr>
<td>LC oscillator</td>
<td>0</td>
<td>3</td>
<td>0.57s</td>
<td>0.33s</td>
<td></td>
</tr>
<tr>
<td>MOS ring oscillator</td>
<td>134</td>
<td>73</td>
<td>30.13s</td>
<td>27.61s</td>
<td></td>
</tr>
<tr>
<td>16 Bit adder</td>
<td>544</td>
<td>283</td>
<td>2m41.32s</td>
<td>2m30.1s</td>
<td></td>
</tr>
<tr>
<td>1 MBit DRAM</td>
<td>2005</td>
<td>1211</td>
<td>10m16.18s</td>
<td>8m29.15s</td>
<td></td>
</tr>
<tr>
<td>16 Mbit DRAM</td>
<td>5208</td>
<td>3500</td>
<td>23m37.18s</td>
<td>12m5.11s</td>
<td></td>
</tr>
<tr>
<td>arithmetic logical unit</td>
<td>13005</td>
<td>32639</td>
<td>97m31.64</td>
<td>82m21.03s</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Numerical results: CHORAL versus BDF within TITAN

Fig. 5. One output nodal voltage for the 16 Bit adder: Integration steps of CHORAL (+) vs. BDF (•)

oscillatory circuits, as can be seen in Fig. 6 for an LC oscillator. On the other hand, the better stability behavior of CHORAL compared with the trapezoidal rule is shown in Fig. 7. With this respect, CHORAL behaves like TR-BDF, what can be explained by the properties of its stability function evaluated along the imaginary axis following the lines of [13].

5 Conclusion

Charge oriented ROW methods like CHORAL are promising alternatives for conventional multistep methods in industrial circuit simulation packages. Starting from an existing test implementation in TITAN, it is now the task of
industry to improve efficiency and robustness, and to gain knowledge about the circuit classes for which the method can be recommended, and for which classes improvements are necessary. This may lead to the development of a bunch of different integration schemes tailored to different classes of circuits, and its implementation in circuit simulation packages is a desirable aim for future research.
References

1 Introduction

The present paper is motivated by the demand from material sciences to reconstruct crystalline structures given through their images under high resolution transmission electron microscopy (HRTEM) in a certain limited number of directions. In particular, [31] and [22] show how a quantitative analysis of images from high resolution transmission electron microscopy can be used to determine the number of atoms on atomic columns in certain directions; see Sects. 2 and 3. Mathematically, this leads to the problem of reconstructing finite lattice sets from certain of their marginal sums; see Sect. 4.

A full 3D-reconstruction of interfaces at the atomic scale may in particular be used to fine-tune the production processes in semiconductor industry. In fact, as silicon wafers are coated with an amorphous layer of oxidized silicon, methods based on external surface scanning – like scanning tunnel microscopy – cannot be used.

There are many other applications and potential applications of the mathematical techniques developed in this project in image processing, graph theory, scheduling, statistical data security, game theory, to name just a few areas (see e.g. [32], [30], [8], [21], [19], [11], [12]). In fact, discrete tomography can be regarded as a prime paradigm for a whole new field dealing with (typically ill-posed) discrete inverse problems.

Acknowledgement: The project was carried out in cooperation with the Institute for Semiconductor Physics, Frankfurt/Oder and Tietz Video & Image Processing Systems GmbH, Gauting. The authors are deeply indebted to Peter Schwander for his constant support, particularly for helping with the physics and the simulation of HRTEM and QUANTITEM.

2 High Resolution Transmission Electron Microscopy

In the early 1930's conventional light-microscopy was pushed to its limits to provide a resolution of about 0.5 μm. In 1931/32, Ernst Ruska (Nobel Prize in Physics in 1986) build the first imaging device utilizing electron beams. The
first transmission electron microscopes (TEM) were mainly used to study biological specimens. A TEM works essentially like a light microscope except that the light beam is replaced by an electron beam and the optical elements are replaced by electron-optical analogues (like electromagnetic or electrostatic lenses). Images are produced by measuring the electron intensities after the beam passed through the specimen; regions with heavier atoms or greater thickness diffract more electrons away from the detector and can hence be detected.

High resolution transmission electron microscopy further improved the resolution possible for very thin specimen. Since for thin samples only marginal absorption occurs within the HRTEM imaging with high acceleration potential of about 200 keV the imaging process is driven by other principles. As there is no loss of energy the resulting wave function $\Psi$ of the interaction between the electron beam and the electrostatic potential of the crystal is governed by the time independent Schrödinger equation

$$\nabla^2 \Psi(r) + \frac{8\pi^2 me}{\hbar^2} [E + V(r)] \Psi(r) = 0,$$

where $e$ denotes the electronic charge, $E$ is the acceleration potential of the microscope, $\hbar$ is Planck’s constant, $m$ is the mass of the electron, and $V(r)$ denotes the crystal potential at position $r$. The interaction creates an electron wave emanating from the ‘bottom side’ of the crystal (if the electron beam hits the specimen from ‘above’). It is this wave that is magnified by electrostatic lenses to obtain the final image. Unfortunately, at this high magnification the lens distorts the image rather strongly; to worsen matters, the distortion depends very sensitively on imaging conditions that cannot be fully assessed. So traditionally, given an HRTEM image, the art is to do simulations of different objects under different imaging conditions until a simulated phantom is created that matches the obtained image. In the next subsection we will sketch one particular simulation technique, the Multislice Method due to Cowley and Moodie [5].

3 Simulation and Analysis of HRTEM Images

We will sketch how the electron wave at the exit face of a specimen can be calculated from the specimen’s crystal potential and the imaging conditions by using the multislice method. This allows us to simulate HRTEM and compute realistic (though small) phantoms under full control of the imaging process. Then we describe a vector analysis technique that can be applied to recover height information. It is called QUANTITEM (quantitative analysis of the information provided by transmission electron microscopy) and was developed by Schwander, Kisielowski, Baumann, Seibt, Kim, and Ourmazd [31], [22].

The main issue with the simulation is to solve the Schrödinger equation in order to obtain a description of the wave at the exit face of the specimen.
For an appropriate simulation it is important to take into account that the involved energies are of quite different orders of magnitude: on the one hand the low energy potential field of the crystal and on the other hand the electron beam of about 200 keV. This difference in scale leads to an almost complete absence of backscattering, i.e. most electrons of the beam are scattered only forward, with the scattering occurring only within a very small angular range. Hence the specific conditions allow it to describe the interaction in terms of a wave propagation, permitting a slice-by-slice simulation with slices orthogonal to the direction of the beam. The full potential of each slice is projected onto the plane of the slice closer to the source of the electron beam (entry face). It turns out that the interaction of the beam with the projected potential on the entry face can be described approximately by a phase shift for the beam’s wave. The propagation of the resulting wave from one entry face to the next is then done by using Fresnel approximation.

The leaving wave is modeled as the sum of many spherical (actually, in this approximation, paraboloidal) waves. Their joint effect on the next entry plane is modeled by a convolution integral. Computationally, the convolution is performed in Fourier space; for details see [5]. The magnifying electrostatic lens is modeled by a similar procedure.

For the simulation of HRTEM we utilized the package EMS by Stadtmann [33]. As specimen we always used a wedge of pure silicon with some cells at the top of atomic columns (viewed in direction (001)) removed at random. The heights of the final probe varied between 0 and 14 atoms; see Fig. 1a for the depiction of such a height-field (the tiny numbers give the number of atoms in the corresponding column).

Using EMS we then computed the simulated image for the resolutions $2048 \times 2048$ subdivided into 30 slices orthogonal to the beam. (Higher resolutions were used to verify that the simulation stabilized.) The simulated image is given in Fig. 1b.

It turns out that for important specimen examples like silicon (under viewing direction (100)) or Germanium (under viewing directions (100) or (110)) the wave function can be approximated very well at the exit face as the superposition of only two Bloch waves. The vectors whose coordinates are the intensities of all fundamental solutions of the PDE that satisfy the necessary boundary and initial value conditions all lie on an ellipse. Furthermore, the angle (to a fixed reference point on the ellipse) corresponds to the height of the sample.

For a view in a main lattice direction an atomic column ‘influences’ only a small region, in effect, the height information is localized. It is observed in [22] that after being magnified and distorted by the lens the images can still be approximated well by two basic images. In [22] this is utilized by segmenting the picture into small rectangles that are influenced essentially only by a single atomic column. Each of these rectangles is considered an image vector on which a principal component analysis is based.
(a) Height of the wedge.

(b) HRTEM Simulation.

(c) Coordinates of the image-cells with respect to the largest and second largest eigenimage.

(d) Sizes of the largest eigenvalues.

Fig. 1. HRTEM-simulation of a $15 \times 15 \times 15$-wedge at resolution $4096 \times 4096$. The tiny(!) numbers in (a) show the heights of the corresponding columns.

In our experiments we used the cells of the Voronoi diagram of the projection of the object to determine the (in this case: rhombic) cells. Then the pixels within a cell were used to construct the image vectors. We computed the principal components for the resolution $2048 \times 2048$; their projections onto the two eigenimages with the largest eigenvalues are plotted in Fig. 1c.

The quality of the simulation and analysis can be inferred from the values of the largest eigenvalues. Theory says, that there should be two significant eigenvalues and all others should be smaller than 0.05; see Fig. 1d. For a resolution of $2048 \times 2048$ this turned out to be true; resolutions smaller than $2048 \times 2048$ did not suffice to show this behavior. Actually, one would expect the largest two eigenvalues to be similar in magnitude; but as the wedges
we studied are very thin we see only a small piece of the ellipse which looks locally like a line; so one eigenvalue dominates the other.

The projection onto the main eigenimages in Fig. 1c reveals a lot about the atomic structure under consideration. Note that from left to right the 15 clusters correspond to columns of height 0 through 14 with only heights 0 and 1 not being separated. The band structure of the clusters is partly due to errors at the boundary of the sample and to the fact that the sample is still very thin compared to the extinction length that lies somewhere between 150 and 200 atoms. But it may also exhibit some additional information related to the relative position of the atoms that may be utilized in the subsequent reconstruction phase.

Since there was no substantial change between resolutions 2048 × 2048 and 4096 × 4096 the simulation seems to be stable. This is also supported by the striking similarities between the two eigenimages at medium resolution (Figs. 2b, e) to those at higher resolution (Figs. 2c, f), respectively. Even though the second most important eigenimages (Figs. 2e, f) are of very little influence here, the slight asymmetry between their top and bottom part can be nicely explained by the fact that the bottom-top direction is the direction along which the wedge rises.

Fig. 2. Comparison of the most (top) and second most (bottom) dominating eigenimages for HRTEM-simulations of a 15 × 15 × 15-wedge at resolution 512 × 512, 2048 × 2048, and 4096 × 4096 (from left to right)
4 The Mathematical Model

In principle, the imaging process described in the previous section essentially provides us with the information how many atoms of a given object interact with given ‘sharply focussed electron beams’ in a certain small number of viewing directions. In a simple but already highly relevant model suggested by Peter Schwander and Larry Shepp the atoms are identified with lattice points in 3-space while the electron beams are modeled as lines parallel to a given direction. Due to the crystalline structure of the samples and the affine invariance of the basic problem it turns out that it is mostly enough to consider subsets of the integer lattice.

Since in practice, one degree of freedom for moving the imaging device is used to control the position of the object, the view directions for which data are provided all lie in one plane (i.e., virtually, the microscope is rotated around one axis). Hence the 3D-problem lends itself to a 2D-slice-by-slice reconstruction.

Now, let us be more precise. Let $F$ be a finite subset of $\mathbb{Z}^n$, let $S$ be a line through the origin, and let $A(S)$ denote the set of all lines of Euclidean $n$-space $\mathbb{E}^n$ that are parallel to $S$. Then the (discrete) X-ray of $F$ parallel to $S$ is the function $X_S F : A(S) \to \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ defined by

$$X_S F(T) = |F \cap T| = \sum_{x \in T} \chi_F(x), \quad \text{for } T \in A(S),$$

where $\chi_F$ denotes the characteristic function of $F$. Note that the integral in the definition of the standard X-ray transform of some function in computerized tomography reduces to a finite sum here; see [27] for an exposition of the mathematics of computerized tomography. Of course, for the practical applications outlined before only the cases $n = 2$ and $n = 3$ are relevant. However, many of the following results hold also in higher dimensions, and for some of the other applications of discrete tomography mentioned in the introduction higher dimensions are important.

In the following let $\mathcal{F}^n = \{ F : F \subset \mathbb{Z}^n \land F \text{ finite} \}$ and $\mathcal{S}^n = \{ \text{lin} \{ u \} : u \in \mathbb{E}^n \setminus \{0\} \}$. The elements of $\mathcal{F}^n$ are called lattice sets. Quite typically, we have some additional a priori information available. This is modeled by considering a suitable subset $\mathcal{G}$ of $\mathcal{F}^n$. For instance, $\mathcal{G}$ may incorporate some contiguity condition that reflects that the crystalline structures that are to be reconstructed do not consist of ‘scattered’ atoms but are highly connected. In most cases it will also be necessary to consider (highly restricted) subsets $\mathcal{T}$ of $\mathcal{S}^n$ since electron microscopic images of high enough resolution can only be obtained in certain directions. An important subset of $\mathcal{S}^n$ is the set $\mathcal{L}^n = \{ \text{lin} \{ u \} : u \in \mathbb{Z}^n \setminus \{0\} \}$ of lattice lines.

Suppose now that $S_1, \ldots, S_m \in \mathcal{L}^n$ are $m$ different lines specified beforehand. Typically, in practice, $m$ is at most 5 since the high energy needed to produce the images leads to distortions already after a few images are taken.
In the inverse problem \( \text{RECONSTRUCTION}_G(S_1, \ldots, S_m) \), we are given \textit{data} functions

\[
  f_i : A(S_i) \to \mathbb{N}_0, \quad i = 1, \ldots, m
\]

with finite support and the task is to find a set \( F \in G \) whose X-rays \( X_{S_i}F \) coincide with \( f_i \) for \( i = 1, \ldots, m \). Of course, due to data errors, in practice it is in general only reasonable to ask for approximate solutions.

To exclude trivial cases we will in the following always assume that \( m, n \in \mathbb{N} \) with \( m, n \geq 2 \).

It should be mentioned that the whole model can be rephrased in a purely combinatorial form that relies only on incidences. Hence many of our results hold for query sets much more general than lines. Further, it is straightforward to extend the model to ‘polyatomic’ structures so as to deal with the reconstruction of compounds.

The following sections describe the main results some of which are based on the additional assumption of exact data. Of course, this assumption is not realistic for our prime application. It is all right though for some other applications indicated above. Also, the uniqueness theorems, results on the computational complexity of the problem, and the development of exact algorithms under this additional assumption lay the ground for the subsequent parts that deal with the treatment of data errors in practice.

We pointed out already that simulations were performed. For applications in practice, visualization of the results is another important issue. In fact, since typically, samples consist of \( 10^8 \) to \( 10^9 \) atoms, visualization tools are needed that can cope with many orders of magnitude, viewing the object from some distance in order to detect ‘interesting parts’ on which has to be zoomed in then.

The main emphasis in the following will be on structural and algorithmic aspects of the problem. Part of what is following is based on [13]. Figure 3 shows a (small) 3D-sample that has been reconstructed from three of its X-rays by methods outlined in Sect. 8.

5 Uniqueness Results

A most fundamental question is that of whether the given information (measurements plus a priori knowledge) determines the underlying object uniquely. Of course, in practice data errors may corrupt solutions hence one has to resort to approximate solutions. But how sensitive are solutions under data errors? Clearly, this question, while looking quite innocent, touches all basic practical problems related to the ill-posedness of the task.

In the present section we are dealing with the issue of uniqueness under the assumption of exact data. This will give a first glimpse of the underlying difficulty of the problem; see also [10].

We say that a subset \( G \) of the set \( F^n \) of all finite lattice sets of \( \mathbb{R}^n \) is \textit{determined by} \( m \) X-rays parallel to the lines in a subset \( T \) of \( S^n \) if there
exist $S_1, \ldots, S_m \in \mathcal{T}$ such that the following holds: When $F_1, F_2 \in \mathcal{G}$ and $X_{S_j} F_1 = X_{S_j} F_2$ for $j = 1, \ldots, m$, then $F_1 = F_2$.

Here is an utterly trivial result.

*With respect to $\mathcal{S}^n$, the class $\mathcal{F}^n$ is determined by one X-ray.*

Of course, a line in a non-lattice direction either misses $\mathbb{Z}^n$ or, if it contains a lattice point, this lattice point is the unique lattice point on this line. As simple as it is, this result indicates already the fundamental difference between discrete and continuous tomography. In fact, it is well-known that compact sets are not determined by their (continuous) X-rays in finitely many directions. Hence discrete tomography is not just the discretization of continuous tomography in the sense that the latter comes from a limiting process of the former using lattice refinements.

Since sufficiently high resolution in the practical imaging process can only be achieved in certain main directions of the lattice, for all practical purposes we have to restrict ourselves to $\mathcal{L}^n$, actually to small subsets thereof. Then the situation is less promising.

*Let $\mathcal{L}$ be a finite subset of $\mathcal{L}^n$. Then there are sets in $\mathcal{F}^n$ that are not determined by the their X-rays parallel to the lines in $\mathcal{L}$.***
There are examples that resemble crystals with just ‘a few impurities’, see [10]. Hence this negative result captures already some of the problems that can actually be seen in practice. Hence, at best, one can only hope for somewhat weaker uniqueness results. Here is a simple one due to [28] and [18].

Let $\mathcal{F}^n(m)$ be the class of sets in $\mathcal{F}^n$ of cardinality less than or equal to $m$. Let $\mathcal{L} \subset \mathcal{L}^n$ with $|\mathcal{L}| \geq m + 1$. Then the sets in $\mathcal{F}^n(m)$ are determined by their X-rays parallel to the lines in $\mathcal{L}$.

The problem with this result in practice is that the typical atomic structures that have to be reconstructed comprise about $10^8$ to $10^9$ atoms. Hence an extremely large number of X-ray images would be needed to guarantee uniqueness. Since the energy needed for the imaging process is very high, after just very few images the object is, however, corrupted by the radiation.

Here is a uniqueness result of [9] for the restricted class $\mathcal{C}^n$ of convex lattice sets, i.e., finite subsets $F$ of $\mathbb{Z}^n$ such that $F = \mathbb{Z}^n \cap \text{conv}(F)$, see [6] for an extension.

Let $\mathcal{L} \subset \mathcal{L}^n$ with $|\mathcal{L}| \geq 7$ and all lines in $\mathcal{L}$ being coplanar. Then the sets in $\mathcal{C}^n$ are determined by X-rays parallel to the lines in $\mathcal{L}$. Further, there is a set $\mathcal{L} \subset \mathcal{L}^n$ with $|\mathcal{L}| = 4$ such that sets in $\mathcal{C}^n$ are determined by X-rays parallel to the lines in $\mathcal{L}$.

This result says that the class $\mathcal{C}^n$ is determined by X-rays parallel to suitable 4 or any 7 different coplanar lattice lines. As pointed out before, in practice the coplanarity assumption is satisfied since in essence the microscope is only rotated about one axis.

While this result is quite reassuring, it is only partly practical. There may be some applications, for instance in colloid physics, but the main demand for mathematical methods for solving the inverse problems of discrete tomography comes from applications that involve the reconstruction of nonconvex objects. In particular, quality control in certain stages of chip production involves the detection of 'bumps' on the interfacial surface of silicon chips, hence convexity is not an appropriate condition in this situation.

Another possibility to outsmart the general nonuniquess result is to change the 'experimental environment'. The above concept of unique determination was based on an a priori choice of the lines for taking X-rays. What if we take the first X-ray in an arbitrary direction but then use the information gained from analyzing the image in order to determine the next line for taking an X-ray? For the third direction, then use the complete information given by the first two X-rays and so on. This approach of successive determination leads to strong uniqueness results even for higher dimensional X-rays and even for sets more general than lattice sets, see [9]. In particular

$\mathcal{F}^n$ can be successively determined by 2 X-rays.
Again, while seeming to be satisfactory, this result is not practical at all. In essence it resembles the first uniqueness result that utilized irrational slopes. In fact, the second X-ray has to be so ‘skew’ that one cannot produce HRTEM images of high enough resolution in this direction.

The lack of practically satisfactory uniqueness results even in the absence of noise indicates already that we need to settle for less. We may ask for the ‘core’ of all solutions, the set of all invariant points that must belong to all solutions. Equally, we are interested in as large as possible sets of points that do not belong to any solution. We might also wish to determine a ‘typical’ solution. Of course, such weaker concepts of ‘solution’ do make immediate sense even in more realistic models that allow data errors.

6 Computational Complexity

Another realistic relaxation of the concept of uniqueness in the previous section is to ask only for a certificate of uniqueness for a given concrete instance of the problem. I.e., rather than requiring general a priori uniqueness results we want to check uniqueness of a solution for each instance separately. Hence, in effect we are asking for an efficient procedure for checking uniqueness algorithmically. This brings up the algorithmic aspect of discrete tomography in the context of uniqueness. But it is certainly clear that efficient procedures for reconstruction are needed anyway.

In this section we begin the discussion of algorithmic features of discrete tomography by stating results on the computational complexity of the questions of checking consistency of X-ray data, of determining uniqueness of given solutions and, of course, of finally reconstructing the objects. In the following we will focus on the full family $\mathcal{F}^n$, and the results will only be stated for that case. However, most of the results hold for a great variety of other subclasses $\mathcal{G}$ as well, [10].

Suppose that $S_1, \ldots, S_m \in \mathcal{L}^n$ are $m \geq 2$ lines specified beforehand. The inverse problem $\text{RECONSTRUCTION}(S_1, \ldots, S_m)$ was already defined in Sect. 4. (Here and in the sequel we will omit the subscript $\mathcal{F}^n$.) In the realm of computational complexity theory decision problems are more appropriate than reconstruction problems. Hence we consider also the problem $\text{CONSISTENCY}(S_1, \ldots, S_m)$ whose instances are just the same, i.e, consist of given data functions

$$f_i : \mathcal{A}(S_i) \to \mathbb{N}_0, \quad i = 1, \ldots, m$$

with finite support, but whose task is restricted to the decision whether a solution exists (without being obliged to produce one). Similarly, the problem $\text{UNIQUENESS}(S_1, \ldots, S_m)$ asks whether, given a solution $F$, there exists another one.

Here are some basic tractability and intractability results. (Of course, one needs to specify the data structures for the problems more precisely. But this is easily done; see [11].)
Let $S_1, \ldots, S_m \in \mathcal{L}^n$ be $m$ different lines. If $m = 2$, the problems CONSISTENCY$(S_1, S_2)$, UNIQUENESS$(S_1, S_2)$ and RECONSTRUCTION$(S_1, S_2)$ can be solved in polynomial time. For $m \geq 3$, CONSISTENCY$(S_1, \ldots, S_m)$ and UNIQUENESS$(S_1, \ldots, S_m)$ are \textsc{NP}-complete in the strong sense while RECONSTRUCTION$(S_1, \ldots, S_m)$ is \textsc{NP}-hard.

Proofs of the tractability part of the above result, i.e. for the statement for $m = n = 2$, can be found in [4], [32], [29], [30] or [1] while the intractability result is due to [11]. The particular case of the \textsc{NP}-completeness of CONSISTENCY$(S_1, S_2, S_3)$ when $S_1, S_2, S_3$ are the coordinate axes in $\mathbb{E}^3$ was previously dealt with in [21, Section 4.1] in the context of statistical contingency tables.

Let us emphasize that the constructions exhibit intractability already for solid crystals with 'just a few impurities', i.e. for physically reasonable objects, whence explaining the algorithmic difficulties observed in practice. Some extensions focusing on nonapproximability results are given in [16]. Some related complexity results can be found in [2], [25] and [10].

7 Integer Programming, LP-Relaxation and Randomized Rounding

Various approaches have been suggested for dealing with the basic \textsc{NP}-hard problems in discrete tomography; see [16] for an account of the success and failure of many of those techniques. In the following we will outline some basic standard approaches for RECONSTRUCTION$(S_1, \ldots, S_m)$ that are based on an integer programming formulation of this problem whose variables correspond to the possible positions of elements of a solution. The grid $G$ of a given instance of the problem consists of all (finitely many) lattice points that arise as points of intersection of $m$ lines parallel to $S_1, \ldots, S_m$, respectively, whose data function value is nonzero, i.e.,

$$G = \mathbb{Z}^n \cap \bigcap_{i=1}^{m} \bigcup_{T \in T_i} T,$$

where $T_1, \ldots, T_m$ denote the supports of the given data functions $f_1, \ldots, f_m$, respectively. The incidences of $G$ and $T_i$ can be encoded by an incidence matrix $A_i$. If $G$ consist of, say, $N$ points, $M_i = |T_i|$ for $i = 1, \ldots, m$, and $M = M_1 + \cdots + M_m$, then the incidence matrices $A_i$ are in $\{0,1\}^{M_i \times N}$, and can be joined together to form a matrix $A \in \{0,1\}^{M \times N}$. Identifying a subset $F$ of $G$ with its characteristic vector $x_F \in \{0,1\}^N$, the reconstruction problem amounts to solving the integer linear feasibility problem

$$Ax = b, \quad \text{s.t. } x \in \{0,1\}^N,$$

where $b^T = (b_1^T, \ldots, b_m^T)$ contains the corresponding values of the data functions $f_1, \ldots, f_m$ as the right hand sides of $A_1, \ldots, A_m$, respectively. Of course,
the equality constraints model the situation of exact data. In practice, this problem will be infeasible and should be replaced by some relaxation, e.g.

$$\max c^T x \quad \text{s.t.} \quad Ax \leq b, \ x \in \{0, 1\}^N,$$

where the objective function $x \mapsto c^T x$ models some specific goal. The all-ones vector $\mathbf{1} = (1, \ldots, 1)^T$, for instance, is used to maximize the number of points that can be placed under the data constraints, but other objectives like the ‘closeness’ to some template etc. can also be facilitated. Approximation algorithms will be dealt with in Sect. 9.

Since linear programming problems can be solved in polynomial time the first natural approach is to consider the LP-relaxation

$$\max c^T x \quad \text{s.t.} \quad Ax \leq b, \ 0 \leq x \leq \mathbf{1};$$

see [7]. Since linear programming codes are available for solving these problems very efficiently for all sizes of crystalline structures that are relevant in practice, computation time is not so much of an issue for this heuristic. However, the solution is usually far from being integer. N. Young studied a randomized rounding strategy where in some stochastic experiment an atom is placed at some lattice point with the probability coming from the fractional solution produced by the LP-solver. This way an approximative solution is produced. Compared to the heuristics described in Sect. 9 known bounds for such solutions are, however, in general rather weak.

8 Polytopes in Discrete Tomography

Of course, one cannot expect too tight general a priori error bounds for polynomial-time approximations in general. For most practical purposes, however, good bounds generated in the course of the computation for a given specific instance are good enough. One can run an improvement algorithm on the given data until the gap between upper and lower bounds is small enough (or until the allotted time elapsed) and then terminate with an approximate solution including a performance guarantee for exactly that solution. This is the underlying idea of branch-and-cut strategies. These are branch-and-bound methods augmented by cutting plane techniques that we are going to sketch now.

The ultimate object for studying cutting planes is of course the convex hull

$$P(A, b) = \text{conv} \{ x \in \{0, 1\}^N : Ax = b \}$$

of all solutions of the problem. In general, even computing the dimension of the polytope $P(A, b)$ is $\mathsf{NP}$-hard. Also, $P(A, b)$ is only relevant in case of consistency whence typically irrelevant in the presence of noise in the imaging process. Hence the submissive

$$T(A, b) = \text{conv} \{ x \in \{0, 1\}^N : Ax \leq b \}$$
is much more appropriate and we study the facial structure of that polytope. 
As we know that the reconstruction problem is \textbf{NP}-hard, we cannot aim at 
a ‘concise’ description of $T(A, b)$ by means of a system of linear inequalities. 
However, appropriate local information is all that is needed; see [3] for a 
nontechnical introduction to polyhedral combinatorics.

The algorithm begins by maximizing a suitable objective function, say 
$x \mapsto \mathbb{1}^T x$, over the polytope 
$$T^{LP}(A, b) = \text{conv} \{ x \in [0, 1]^N : Ax \leq b \}.$$ 
This is a task of linear programming that can be handled efficiently. If the 
produced vertex $x^*$ at which the optimum is attained is integral then the 
problem is solved. Otherwise, we try to improve the relaxation $T^{LP}(A, b)$ 
of $T(A, b)$ by adding a suitable inequality that separates $x^*$ from $T(A, b)$.
Geometrically we intersect $T^{LP}(A, b)$ with an appropriate closed halfspace 
that contains $T(A, b)$ but not $x^*$. Such a halfspace is called a \textit{cut} and the 
approach of successively solving the current LP-relaxation, and adding a cut 
if still necessary is called a \textit{cutting-plane} algorithm. In general, it is not 
clear how to find reasonably deep cuts efficiently, and typically some lurking 
\textbf{NP}-hardness is connected to this. (Of course, if the original problem was 
\textbf{NP}-hard, its intractability cannot magically disappear.) Hence it may be 
possible that the current LP-optimum is not integer, yet we do not know 
any cut that can be added. Then we need to resort to a branching strategy, 
splitting the problem into various others by forbidding certain properties for 
some and forcing them for other subproblems. For instance, we may split the 
tomographic reconstruction task into two, one where a certain point must be 
present, the other, where this point must not be present.

This is the \textit{branch-and-cut} approach, performed within the branch-and-
bound framework, with cutting planes used for improving the upper bound 
at each node of the branching tree.

Of course, it is similarly important to obtain good lower bounds, i.e. 
approximate solutions close to the optimum. They can be achieved by running 
heuristics in each branching node. We report about theoretical and practical 
results for such approximation algorithms in Sect. 9. Primal improvement 
strategies based on tests sets were studied in [35], where a general algebraic 
framework for improvement strategies in discrete tomography was developed.

The \textit{tomography polytopes} $T(A, b)$ have some quite special structure. In 
particular, all submatrices of $A$ corresponding to just two directions are \textit{totally unimodular}. Since all tomography polytopes are 0-1-polytopes it follows from 
[26] that the combinatorial \textit{diameter} of a tomography polytope is at most $N$. 
This means that, in principle, an edge-path could be found leading from 0 to 
a solution of the problem that is rather short.

Further, by means of various preprocessing techniques, the practically 
relevant dimensions of these polytopes can be reduced to about $10^4$. Judged 
on the base of the sizes of successfully solved instances of the traveling salesman 
problem this seems encouragingly small. While the traveling salesman
polytopes correspond to the complete graph on the given number of ‘cities’ and are hence universal, tomography polytopes depend, on the other hand, on the right-hand side \(b\). So the most important goal of polytopal investigations is to find large systems of valid inequalities that are facet-defining under weak conditions on the right-hand side. In [34] various classes of facet-defining inequalities are determined under weak assumptions on \(b\). Usually it is only necessary to require that all components of \(b\) are at least 2 or 3 and that no X-ray line is completely filled with atoms.

9 Approximation Algorithms

In view of the computational complexity of \(\text{RECONSTRUCTION}(S_1, \ldots, S_m)\) for \(m \geq 3\) and in view of the presence of noise in the data \(b\) it is reasonable to study approximation algorithms for this problem. It turns out that in spite of the underlying \(\text{NP}\)-hardness some simple heuristics already yield good a priori approximation guarantees and behave surprisingly excellent in practice.

The problem \(\text{BEST-INNER-FIT}(S_1, \ldots, S_m)\) [BIF] accepts as input data functions \(f_1, \ldots, f_m\); its task is to find a set \(F \in \mathcal{F}^n\) of maximal cardinality such that

\[
X_{S_i} F(T) \leq f_i(T) \quad \text{for all } T \in \mathcal{T}_i \text{ and } i = 1, \ldots, m.
\]

This problem and its outer counterpart were studied in [15]. In particular, various greedy and improvement strategies were fully analyzed. The main results will now be stated.

The greedy algorithm considers the positions of the grid \(G\) of candidate points in some order and successively fills in points as long as this is possible without violating the constraints. A simple observation shows that any such algorithm produces a solution \(V\) such that

\[
|V|/|F| \geq 1/m,
\]

where \(F\) is an optimal solution; cf. [24]. The bound is sharp and reflects the fact that the more data are given, the harder it is for a greedy strategy to satisfy them. In the experiments it turns out, that \(|V|/|F|\) is typically greater than 0.9 and for large instances greater than 0.96 even for \(m = 5\).

The greedy strategy is very flexible and allows various specifications for breaking the ties between different choices for points to be placed next. For example, the X-ray data can be used in a way that is very similar to back-projection techniques to express preferences. Also connectivity of the solution (in a sense that is justified by the physical structure of the analyzed material) can be rewarded. Similarly, information of neighboring layers can be taken into account in a layer-wise reconstruction of a 3D-object.

In the practical experiments, an algorithm ‘GreedyC’ was particularly good. Here the insertion order is based on weights that are dynamically
assigned to the candidate points and represent the ‘changing importance’ of a point to be included in a solution. The computational part of [15] shows that GreedyC actually produces solutions with small absolute error. The average for GreedyC when applied to instances with 250000 variables and point density of 50% were 21.62, 64.13 and 111.88 missing atoms for 3, 4, 5 directions, respectively.

In order to further improve the performance of such iterative insertion algorithms, one can apply r-improvements for $r \in \mathbb{N}_0$ that replace an $r$-point subset of a current feasible solution $F \subset G$ by $r + 1$ points of $(G \setminus F)$ while maintaining feasibility. (Of course, 0-improvements are just insertions.) A feasible set $F \subset G$ is called t-optimal for the given instance of [BIF] if no r-improvement is possible for any $r \leq t$.

The power of r-improvements can be seen in the next result; (asymptotically) the a priori performance bound is improved by another factor of 2.

Let $t \in \mathbb{N}$, let $F$ be a solution of a given instance of [BIF] and let $V \subset G$ be t-optimal. Then

$$\frac{|V|}{|F|} \geq \frac{2}{m} - \epsilon_m(t),$$

where $\epsilon_m(t)$ approaches 0 exponentially fast.

In [15] $\epsilon_m(t)$ is given explicitly and examples are constructed that show that the bounds are tight. Computationally, it turns out that performing 1-improvements after GreedyC yields substantial improvements. In fact, in our computational study the absolute errors go down to 1.07, 23.28, 64.58 for 3, 4, 5 directions, respectively; see [15].

Another approximation algorithm is based on LP-relaxation. It turns out that by rounding down all fractional components of an optimal vertex of $T^{LP}$ we obtain a solution of [BIF] with objective value at most $M$ worse than the optimum. This permits to devise polynomial approximation schemes for certain classes of instances of [BIF]; see [14] for details.

10 Dealing with Ill-Posedness

The results of Sect. 9 are extremely encouraging in that they show that the ‘combinatorial optimization part’ of the reconstruction problem of discrete tomography can be handled very efficiently in spite of its computational complexity. However, there is more to be taken care of. In fact, the relevant measure for the quality of an approximation to a binary image would of course be the deviation from this image. Hence in order to devise the most appropriate objective function one would have to know the underlying solution of the given inverse problem. However, the whole point is of course to find this unknown solution. Hence one can only consider objective functions
for measuring the quality of approximation that are based on the given input data. While a good approximation in this sense is close to a solution in that its X-ray images parallel to the given lines are close to those of the original set, the approximating set itself may be off quite substantially. In fact, the inverse discrete problem is ill-posed and it is precisely this property that causes additional difficulties. In particular, if the input data do not uniquely determine the image even a ‘perfect’ solution that is completely consistent with all given data may be quite different from the unknown real object.

There are various approaches for dealing with nonuniqueness. If we use the fact, that in our prime application the lines parallel to which the X-ray images are taken are coplanar, then the objects can be reconstructed in a layer-by-layer wise process. Suppose that the first layer has been reconstructed and that it is uniquely determined by the given information (or known beforehand). Then it may be reasonable to assume, that the second layer does not vary too much from the first. As pointed out already this can be modeled easily by an objective function for the second layer that is given through the incidence vector of the solution for the first layer. In case of just two X-rays such an approach was suggested by [32].

In [7] an interior point heuristic is proposed for identifying positions that are uniquely determined by the given data. While the problem of detecting whether a given subset of $G$ belongs to all solutions for a given instance is again $\mathsf{NP}$-hard, [11], [16], it is demonstrated in [7] that for certain phantoms such an approach produces quite a large number of fixed variables. Note that the problem of deciding whether a subset of a possible solution can actually be extended to a full solution is again $\mathsf{NP}$-complete, see [11].

Various other techniques for reducing ambiguity are proposed in the study [17]. In any case, it is still important to utilize additional physical knowledge and experience in order to be able to produce solutions that are close to the actual physical objects.

The main problem, however, is that of instability under changes of the data $b$. Of course, it is easy to see that the cardinality of a maximum solution of [BIF] is stable. Also, if $b_1$ and $b_2$ are close and $F_1$ is a solution for the measurement $b_1$, then there is always an approximate solution $F_2$ for $b_2$ that is close to $F_1$ in terms of their symmetric difference. However, one can use constructions of [23] to produce for any given $m \geq 3$ and $k \in \mathbb{N}$ two vectors $b_1$ and $b_2$ with $\|b_1 - b_2\|_2 \leq 2(m-1)$ such that there are unique but disjoint sets $F_1, F_2$ of cardinality at least $k$ whose X-rays coincide with $b_1, b_2$, respectively. Hence the data differs only by a constant, while the symmetric difference of the optimal solutions can be arbitrarily large. This shows the ultimate limitations of general theoretic stability results. However, the examples are very specific and in practice the results are typically quite reasonable in that they exhibit the main features of the original object. In any case, a general theory of discrete inverse problem is still to be developed.
11 Software

The software is build around a module that encapsulates all data-structures that are specific to questions of discrete tomography. In particular, they permit to treat the case of arbitrary three-dimensional lattices.

We implemented all the approximation algorithms mentioned in Sect. 9 and several others. Further, we included a branch-and-bound module based on CPLEX [20].

On top of these modules with data-structures and solver algorithms resides the user-interface. Besides being the command center for the solvers it provides means for file-input/output and different methods for generating instances of various types. One such feature is line-convexity in certain directions, used in a first experimental study of large instances that mimic the etching process applied in silicon wafer production.

![Screenshots of the user interface for a small example.](image)

**Fig. 4.** Screenshots of the user interface for a small example. In the upper left the underlying phantom is depicted. The lower left shows a reconstruction from images parallel to the lines spanned by $(1, 0), (0, 1),$ and $(1, -2)$. The symmetric difference between the phantom (black) and its reconstruction (grey) is printed in the upper right corner. The main menu of the tool is shown in the lower right.
References


Measurement of Paint Layer Thickness with Photothermal Infrared Radiometry

A. K. Louis¹, P. Dörr¹, C. Gruss², and H. Petry²

¹ Institut für Angewandte Mathematik, Universität des Saarlandes, 66041 Saarbrücken, Germany, http://www.uni-sb.de
² Phototherm Dr. Petry GmbH, Altenkessler Strasse 17, 66115 Saarbrücken, Germany, http://www.phototherm.de

Abstract. Photothermal infrared radiometry has been used for the measurement of thermo-physical, optical and geometrical properties of multi-layered samples of paint on a metallic substrate. A special data-normalization is applied to reduce the number of sensitive parameters which makes the identification task for the remaining parameters easier. The normalization stabilizes the evaluation of the photothermal signal and makes the infrared radiometry more attractive for applications in the industrial environment. It is shown that modelling and multi-parameter-fitting can be applied successfully to the normalized data for the determination of layer thicknesses. A second approach is presented to verify the adaptability of reconstruction algorithms for thickness measurements. An algorithm which uses the affinity of the thermal waves to the acoustic waves and the inverse scattering problem demonstrates the applicability in general.

1 Introduction

The determination of coating thickness is an important requirement for the industry [1,2]. As input for remote control systems, thickness information has to be supplied in realtime. Data acquisition has to be performed in a non-contact manner. Not only non-destructive and remote control are required, but also methods allowing to determine the composition of a multilayer-coating. Furthermore the calibration should be as easy as possible. In conventional modulated photothermal infrared radiometry, paint-film-thickness is measured using lock-in amplifiers. The phase-difference between modulated laser-excitation and detected infrared signal that is emitted by the sample [3,4] is sensitive to changes of the paint-layer thickness and therefore used as calibration variable. The phase- and amplitude data to a series of excitation frequencies is called a frequency-scan and can be used to identify thermophysical parameters of the sample [5,6]. It is represented by a function of complex value with the excitation frequency as argument. If – as in our instrumentation – a mechanical chopper is used to generate a periodical heat source on, respectively within, the sample then we waste a lot of information if only the first harmonic part of the infrared signal is used: Since the chopped laser beam can be approximated by a square wave, we have significant amplitudes at odd multiples of the chopper frequency. The heat equation, which
is the mathematical model of the heat diffusion in the sample, is a linear differential equation in the case of temperature independent thermo-physical properties. Furthermore the radiation depends approximately linearly on the temperature in the case of small temperature variations. Therefore we will have significant amplitudes of the periodical radiation-intensity-function of the sample only at odd multiples of the chopper frequency. In this paper we describe a general concept how this information can be extracted from the measured signal and used in infrared radiometry, we call it *simultaneous calibration*. In a first step we derive a mathematical model of the heat flow in \( N \)-layer samples. Several approaches to multi-layer-systems can be found in literature [3,7,8]. It is assumed that the first layers are light-absorbing and that the last is a metallic substrate, and the layered sample is surrounded by air, which is assumed to be a thermal isolator, see Fig. 1.

![Diagram of experimental set-up](image)

**Fig. 1.** Schematic illustration of the experimental set-up

In the next section we present the general method that we use to extract higher harmonic information from the time signal. This is done theoretically as well as experimentally. The inverse problem of parameter identification is presented with some remarks on the ill-posedness of the problem. Parameter studies performed on measured data sets are used to verify our modelling and normalization procedure.

A second approach makes no assumption on the special geometrical structure of the sample. The diffusivity profile of the sample is assumed to be a function of depth. The actual instrumentation does not allow to make frequency scans in real time. But assuming this would be possible, we were able to show that reconstruction of the diffusivity- respectively effusivity-profiles can be used to determine the paint layer thickness. All experiments
were performed by the commercial system PS-Online of the Phototherm Dr. Petry GmbH. An illustration of the experimental setup is shown in Fig. 2.

2 Mathematical Model I

We consider a sample composed of $N$ layers; see Fig. 1. The laser spot is assumed to be very large compared to the sample thickness and to the detection spot. Since in addition the detection spot is assumed to be large compared to the thermal diffusion length $\mu = \sqrt{2\alpha/\omega}$, our problem can be reduced to a one-dimensional one [3]. $\alpha$ is the thermal diffusivity, $\omega$ the angular frequency. The sample will be heated by a square wave heat-source function $q(x, t)$, $0 \leq x \leq x_N$, $0 < t$ induced by a time periodic chopper at frequency $f_0$

$$q(x, t) = q_0 \beta \exp(-\beta x) \text{square}(f_0, t),$$  \hspace{1cm} (1)

where $\beta$ is the infrared absorption coefficient of the first layer and $\text{square}(f_0, t)$ the square wave function at frequency $f_0$. For the sake of simplicity we assume that scattering and reflection of the induced laser beam can be ignored and that all layers have the same absorption coefficient $\beta$. Then $q$ is also time periodic with frequency $f_0$ and the temperature $T$ in the sample is the solution of the heat equation, which can be written in the frequency domain as [9]:

$$\frac{\partial^2}{\partial x^2} T(x, \omega) - \sigma^2(\omega) T(x, \omega) = -q(x, \omega),$$  \hspace{1cm} (2)

where $\sigma = (1 + i)\sqrt{\omega/2\alpha}$ is the thermal wave number, $\alpha$ the thermal diffusivity and $q_0$ the maximum induced power. The heat equation is transformed into the Helmholtz equation which shows the similarity between thermal and acoustic waves. Here the change of variables $t$ and $\omega$ denotes the transition to frequency domain. Since $q$ is a square wave function we know that $q(x, \omega)$ has amplitudes different from zero only for frequencies which are odd multiples of the chopper frequency $\omega_0 = 2\pi f_0$. From now on we write $q(x, \omega)$ and $T(x, \omega)$
and \( \omega \) is assumed to be an odd multiple of the chopper frequency \( \omega_0 \). We assume the sample layers to have constant physical properties. In each layer \( j \) the solution of (2) has the form

\[
T_j(x) = A_j e^{-\sigma_j(x-x_j)} + B_j e^{\sigma_j(x-x_j)} + C_j e^{-\beta(x-x_j)},
\]

where \( x_0 = 0, T_j(x), 0 \leq x \leq L_j \) and \( q_j := q_0 e^{-\beta x_j} \). We are searching for a smooth solution, where \( T_j \) satisfies the following boundary conditions:

\[
\frac{\partial}{\partial x} T_0(0) = \frac{\partial}{\partial x} T_{N-1}(x_N) = 0
\]

\[
T_j(x_{j+1}) = T_{j+1}(x_{j+1})
\]

\[
k_j \frac{\partial}{\partial x} T_j(x_{j+1}) = k_{j+1} \frac{\partial}{\partial x} T_{j+1}(x_{j+1})
\]

for \( j = 0, \ldots, N - 2 \). This linear system of equations (4–6) for the coefficients \( A_j \) and \( B_j \) can be solved analytically. The surface temperature is then \( T(0) = A_0 + B_0 + C_0 \). The value \( C_0 \) is determined by equation (2). More differentiated models can be derived in the same way as presented here. In practical application the absorbed power \( q_0 \) is difficult to quantify and dependent on environment conditions as well as apparatus components. We remove the influence of \( q_0 \) using the fact that all coefficients \( A_j, B_j \) contain the factor \( C_j \). So \( q_0 \) can be factored out and is then cancelled through normalization. This is the subject of the next section.

3 Simultaneous Calibration

In the case of a one dimensional model the radiometric signal is [10]

\[
S(\omega) = K \beta_{IR} \int_0^L \exp(-\beta_{IR} x) T(x, \omega) dx,
\]

where \( K \) is an instrumental constant depending on geometrical factors, the emissivity averaged over the spectral bandwidth of the detector, the Stefan Boltzmann constant, and the ambient temperature. Additionally we have the influence of electronic devices such as amplifiers in the parameter \( K \) so that \( K = K(\omega) \) may depend on the frequency and has complex value. We suggest a new way to normalize the measured data which is based on a multi-harmonic thermal excitation (which in turn produces a multi-harmonic infrared signal). This allows to use one harmonic part as reference and one or all the other harmonic parts as information carrier [11]. In the case of a square wave excitation we apply the following transformation to the measured time data \( S(t) \):

\[
S(t) \rightarrow \hat{S} = \left( \frac{S(\omega_0)}{3\sqrt{3}S(3\omega_0)}, \frac{S(\omega_0)}{5\sqrt{5}S(5\omega_0)}, \ldots, \frac{S(\omega_0)}{(2n+1)^{3/2}S((2n+1)\omega_0)} \right)^T
\]
We call the mapping $S(t) \rightarrow \hat{S}$ \textit{simultaneous calibration} \cite{11} because information retrieval and calibration are carried out in parallel. The factors $(2n+1)^{3/2}$ are chosen so that a semi-infinite sample with surface absorption has a normalized phase of zero and a normalized amplitude of one in all entries of $\hat{S}$. Compared to traditional normalization there are some serious advantages of this method: first the gathered information is not affected by changes of the induced maximum power, which may result from changes of the ambient terms or replacement of apparatus components. Second, we get information about different thermal depth from only one measurement.

4 \hspace{1em} \textbf{Multi-Parameter-Fitting}

It is well known that the inversion problem of parameter identification for thermal wave data is an ill-posed problem \cite{8}. One consequence of this effect is that the excitation frequency has to be chosen well so that data is sensitive to changes of sample parameters. Moreover, ill-posedness limits the amount of information that can be retrieved from noisy data. Let us consider the case of a two-layer system paint on metallic substrate. Let $T(x, \beta)$ be the temperature distribution to the absorption coefficient $\beta$ in the thermally excited sample. Then it can be shown using the mathematical model derived in Sect. 2 that

$$\lim_{\beta \rightarrow \infty} \int_{0}^{L} e^{-\beta_{IR}x}T(x, \beta)dx = \lim_{\beta_{IR} \rightarrow \infty} \int_{0}^{L} e^{-\beta_{IR}x}T(x, \beta)dx = T(0, \beta).$$

when we insert the same value for $\beta$ and $\beta_{IR}$. This means that the influence of absorption and emission is interchangeable in the common case of small optical penetration depths. For that reason we omit one of the parameters to stabilize the inversion procedure. Since the remaining parameter $\beta$ – in our examples we set $\beta_{IR} = \infty$ – compensates the influence of the absorption and emission effect, it cannot be seen as quantitative variable. But since we are interested in layer-thickness this is not a disadvantage. With the introduction of the compensation-parameter we decrease the number of free parameters from three to two per layer.

5 \hspace{1em} \textbf{Experimental Results: Measured Data}

The experimental set-up is presented in Fig. 2: we have a CO$_2$ laser whose output beam is first modulated by a chopper, then focused by a lens of ZnSe, so that the beam can pass a small hole in a mirror. Futhermore this lense controls the diameter of the laser-spot on the sample. The function of the mirror is to deflect the emitted radiation in direction of the IR-detector. Another lense, a CAFI lens focus the radiation on the InSb detector which is cooled by a Stirling cooling system. Digitized infrared time data were sampled and
analyzed. We present the results for a three layer sample: two different coating layers of paint on a steel substrate. The sum of the coating thicknesses and substrate thickness (430 μm) were measured by other methods and therefore approximately known. They were used to verify our mathematical model and normalization approach. Time data acquisition was done on six samples. Paint layer thickness varied between 10 and 20 μm, ten chopper frequencies between 2 and 75 Hz were applied. The second layer was a ground coat layer. It was thin compared to the first layer. The substrate layer thickness and the thickness of the ground coat layer were supposed to be the same for all samples. Digitized time series of the pre-amplified detector signal were sampled at a rate of 200 MHz. No Lock-In amplifier was used.

First of all it has to be mentioned that we were not able to differentiate the two paint layers with this data in a stable and reproducible way. Fitting procedure with the three-layer model converges to a two-layer sample. For that reason we used the two-layer model for all parameter estimations.

We want to verify the approximation quality of our mathematical model. The verification method is as follows: Using the knowledge of layer thicknesses we can replace the model parameter \( \text{thermal thickness } L_0/\sqrt{\alpha_0} \) of the first layer with the diffusivity \( \alpha_0 \) because \( L_0 \) is known as mentioned before. Now we fit the frequency scan data to all samples simultaneously. We use the normalized data \( S_1, \ldots, S_4 \) (8). In Figs. 3, 4 we see the respective plots for normalized amplitudes and phases. The test presented in Table 1 provides information about the stability of our approach as a single-sample-calibration method. One calibration sample per row with known thicknesses was used to determine the thermophysical parameters of the sample. For the other samples solely the thickness \( L_0 \) remains as fitting parameter. It can be seen

![Fig. 3. Multi-Parameter-Fit: Normalized amplitudes \(|\hat{S}_1|\). Six samples of different and known paint layer thickness were used to fit the thermophysical parameters of the samples](image-url)
**Fig. 4.** Multi-Parameter-Fit: Normalized phases \( \arg(\hat{S}_1) \)

<table>
<thead>
<tr>
<th>calibration with</th>
<th>10 ( \mu )m</th>
<th>12 ( \mu )m</th>
<th>14 ( \mu )m</th>
<th>16 ( \mu )m</th>
<th>18 ( \mu )m</th>
<th>20 ( \mu )m</th>
</tr>
</thead>
<tbody>
<tr>
<td>all samples:</td>
<td>10.0</td>
<td>12.1</td>
<td>14.2</td>
<td>16.1</td>
<td>18.1</td>
<td>19.4</td>
</tr>
<tr>
<td>10 ( \mu )m sample:</td>
<td>10.0</td>
<td>11.9</td>
<td>13.9</td>
<td>15.5</td>
<td>17.3</td>
<td>18.4</td>
</tr>
<tr>
<td>12 ( \mu )m sample:</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>15.7</td>
<td>17.5</td>
<td>18.7</td>
</tr>
<tr>
<td>14 ( \mu )m sample:</td>
<td>9.7</td>
<td>11.9</td>
<td>14</td>
<td>15.8</td>
<td>17.8</td>
<td>19</td>
</tr>
<tr>
<td>16 ( \mu )m sample:</td>
<td>9.5</td>
<td>11.7</td>
<td>14.0</td>
<td>16</td>
<td>18.0</td>
<td>19.3</td>
</tr>
<tr>
<td>18 ( \mu )m sample:</td>
<td>9.4</td>
<td>11.6</td>
<td>14</td>
<td>16.0</td>
<td>18</td>
<td>19.3</td>
</tr>
<tr>
<td>20 ( \mu )m sample:</td>
<td>8.7</td>
<td>11.2</td>
<td>14</td>
<td>16.3</td>
<td>18.6</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table 1.** Computed thicknesses of the first layer for all samples. Thermo-physical properties were computed with the calibration data. All parameters but the first layer thickness were frozen. One parameter fitting for each sample gives the computed quantitative layer thickness. Given thickness was measured with a non-contact-free ultrasound method

that all samples deliver stable parameters, solely the 20 \( \mu \)m sample leads to slightly different results. This is probably caused by less accurate thickness information. But we find our method to be stable, at least in the sense that the computed thicknesses are monotonically increasing when thickness increases.

**6 Mathematical Model II**

The first approach uses the layered structure of the samples. The concept presented here does not make such assumptions. We reconstruct a profile of some thermophysical properties of the sample. Thermal diffusivity is very different for metal and for paint, so that our aim will be to find the position of the interface between the layers. Some methods can be found in literature
concerning the reconstruction of conductivity or effusivity profiles of hardened steel. We compare one of these methods [12] with a new one which uses the analogy with the acoustic inverse scattering problem. Starting from the Psdeudo Helmholtz equation [13]

$$\left( \frac{d^2}{dx^2} - \sigma_0^2 \right) T(x) = F(x)T(x) \tag{9}$$

where

$$F(x) = \begin{cases} \sigma_0^2(n(x)^2 - 1) : 0 \leq x \\ 0 : x < 0 \end{cases}$$

and the thermal refraction index $n(x) = \sqrt{\alpha_0/\alpha(x)}$ Then $T$ is a solution of (2) and can be written as $T(x) = T_i(x) + T_s(x)$

$$T_s(x) = -\alpha_0 \int_0^\infty G(x, y)F(y)T(y)dy \tag{10}$$

$$T_i(x) = \frac{I}{\sigma_0 k_0} e^{-\sigma_0 x} \tag{11}$$

the sum of incident and scattered thermal wave. $G$ is the Greens function of equation (9)

$$G(x, y) = \frac{1}{2\sigma\alpha} \left( e^{-\alpha|y-x|} + e^{-\alpha|x+y|} \right)$$

We assume the temperature values at the surface were available to different excitation frequencies. With these data we have to reconstruct the profile $F(x)$. Using the Born-Approximation

$$T_s(0, \omega) \approx -\alpha_0 \int_0^\infty \frac{1}{2\alpha_0\sigma_0} 2 e^{-\sigma_0 y} T_i(y, \omega)F(y)dy$$

$$= -\frac{\alpha_0}{2k_0} I \int_0^\infty e^{-\sigma_0 y} \left( \frac{1}{\alpha(y/2)} - \frac{1}{\alpha_0} \right) dy, \quad \text{:=} f(y) \tag{13}$$

this problem becomes linear. We see that our data can be approximated by the Laplace transform of the object $f$, which is directly connected to the original object function $F(x, \omega) = i\omega f(x)$. The concept called approximative inverse [14] has been applied for the numerical inversion of the the Laplace transform. This problem is extremely ill posed, and the algorithm for the numerical solution very complex, so that we cannot present further details here. One important fact is, that we need a frequency scan, that includes the sensitive frequencies. If not – there will be a lack of information: Since the excitation frequency triggers the penetration depth of the thermal wave, we cannot reconstruct a profile above a certain maximum depth which depends not only on the frequency but also on the diffusivity of the object itself.
7 Experimental Results: Simulated Data

A chopped laser beam which is used for the thermal excitation of the sample does not provide the frequency-scan data with more than 50 excitation frequencies in real time which is necessary for profile reconstruction. Real time measurements of such a frequency-scan cannot be realized with the given apparatus. For that reason we use synthetically data which was generated using the (nonlinear) layered model presented in Sect. 2. Our aim is to show what could be possible, if the measurement technology is able to produce real-time frequency-scans. As a benchmark we examine five samples of paint on a metallic (semi-infinite) substrate. Again we were not able to distinguish multiple paint layers when the data is disturbed by (artificially introduced) noise. But the location of the interface between paint and metal was successful. A comparison between the born-approximation solution and the results of the state-of-the-art algorithm developed by Bertolotti et al. shows, that both algorithms can be used for thickness computation. We have to mention, that the bertolotti algorithm was originally designed for the reconstruction of smooth effusivity profiles of hardened steel. We showed, that it can be adopted for the determination of layer thickness. In Figs. 5, 6 we compare the reconstruction quality of the two algorithms. While Bertolottis algorithm (Fig. 5) is very good in the region of constant effusivity we see better results in the reconstruction of the constant profile of the substrate

![Graph](image)

**Fig. 5.** Diffusivity profile reconstruction using born approximation. 100 Frequencies between 0.3 and 400 Hz were used. Regularization with truncated singular value decomposition and smoothing with gauss-kernel
Fig. 6. Effusivity profile reconstruction using bertolotti algorithm. 100 Frequencies between 0.3 and 400 Hz were used. Regularization with truncated singular value decomposition and smoothing with gauss-kernel

for the born-approximation algorithm (Fig. 5). Examples have shown, that the localization of the interface is more stable under the influence of noise for the born-approximation. Both algorithm were able to demonstrate, that linearized models are quite good approximations for the nonlinear problem of profile reconstruction – at least if we are only interested in the location of interfaces within the sample.

<table>
<thead>
<tr>
<th>exact thickness</th>
<th>computed thickness Bertolotti algorithm</th>
<th>computed thickness Born-approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 μm</td>
<td>38.8</td>
<td>40.1</td>
</tr>
<tr>
<td>42 μm</td>
<td>40.1</td>
<td>42.1</td>
</tr>
<tr>
<td>44 μm</td>
<td>42.8</td>
<td>44.5</td>
</tr>
<tr>
<td>46 μm</td>
<td>44.2</td>
<td>46.3</td>
</tr>
<tr>
<td>48 μm</td>
<td>48.4</td>
<td>48.4</td>
</tr>
<tr>
<td>50 μm</td>
<td>48.8</td>
<td>50.1</td>
</tr>
</tbody>
</table>

Table 2. Comparision of reconstruction algorithms. Synthetically generated data for excitation frequencies between 0.3 and 400 Hz, Gaussian noise of 1%. Bertolotti algorithm uses maximum of effusivity contrast $e'(x)/e(x)$ to identify interface location between paint and metal. Born-approximation algorithm uses maximum of differentiated diffusifity profile $\alpha(x)'$
8 Conclusion

Photothermal radiometry has been used to identify samples of a paint layer on a metallic substrate. A very robust method consisting of a signal analysis – and modelling module has been presented. It makes it possible to identify parameters with very few data, acquired in a rough industrial environment.

We were able to show that multi-parameter fitting can be a valuable tool in the measurement technology if we reduce the free parameters to those with the highest sensitivity. The development of suitable data transforms – like simultaneous calibration – in combination with reduced mathematical models leads to measurement methods with some kind of intelligence. Compared to a neural network approach, the modelling-multi-parameter-fitting process has significant advantages: only one frequency scan to only one sample is enough training for our method. After the parameter-identification is performed, the thickness of any other sample – consisting of the same materials – can be identified. We want to emphasize particularly the normalization scheme that we presented. It is the key part of the procedure since it reduces the number of parameters with a strong influence on the (normalized) photothermal signal to four free parameters for a two-layer sample. Moreover the photothermal information – the transformed frequency scan data – is independent of the instrumentation; replacing instrument components does not induce the need of new calibration or multi-parameter-fitting.

Furthermore we were able to show that profile reconstruction as a tool for thickness determination could also be successful, if necessary modifications of the measurement apparatus are technically feasible. But linear models as presented here are not sensitive enough to distinguish different paint layers, a case which is important in industrial applications. Therefore future work has to deal with nonlinear models, and nonlinear profile reconstruction.

References

1. C. Gruss, Grant report for the foundation Stiftung Industrieforschung, Köln, Germany, (1999)


Spatio-Temporal Current Density Reconstruction from EEG-/MEG-Data

Alfred K. Louis¹, Uwe Schmitt¹, Felix Darvas², Helmut Buchner³, and Manfred Fuchs⁴

¹ Institut für Angewandte Mathematik, Universität des Saarlandes, 66041 Saarbrücken, Germany, URL: http://www.num.uni-sb.de
² Neurologie, RWTH Aachen, 52057 Aachen, Germany, E-mail: fdarvas@post.klinikum.rwth-aachen.de
³ Klinik für Neurologie und klinische Neurophysiologie, 45667 Recklinghausen, Germany, E-mail: hbuchner@kk-recklinghausen.de
⁴ Neuroscan Labs, 20255 Hamburg, Germany, E-mail: mfuchs@neuroscan.com

Abstract. The determination of the sources of electric activity inside the brain from electric measurements on the surface of the head is known to be an ill-posed problem. In this paper a new algorithm which takes temporal a-priori information into account is described and compared to existing algorithms as Tikhonov–Phillips. There are further applications in medical and technical fields as the determination of electrical sources in the living heart and the determination of acoustic sources.

1 Introduction

Inverse source reconstruction has many applications in clinical and theoretical medicine. Examples are the noninvasive localization of focal epileptogenic discharges [1], [2] and the study of somato sensoric evoked potentials (SEP’s) [3]. An overview of the existing techniques is given in [4]. The same technique can be applied to the human heart e.g. in order to localize the origin of premature and extrasystolic beats [5].

Measurements of the electromagnetic activity of the brain with electroencephalography (EEG) or magnetoencephalography (MEG) provide an excellent temporal resolution, compared to positron-emission-tomography (PET) or functional magnetic-resonance-imaging (fMRI). However, since both methods, EEG and MEG, measure activity only on the surface of the head, they have a considerable lack of spatial resolution. Furthermore EEG and MEG measurements are strongly affected by noise. As we will see, the problem of determining electrical sources inside the scalp from EEG/MEG-measurements is ill-posed. Thus we need additional spatial and temporal information about the brain activity in order to tackle this problem.

There are two different strategies for the solution of the inverse problem:

– By optimization techniques a small number of dipoles and their parameters (location, orientation and magnitude) are calculated ensuring best fit between measured data and forward calculated potentials [6], [7]. See
Fig. 1 left picture. This is a nonlinear problem and regularization is achieved by the small number of parameters.

- By current density reconstruction (CDR) together with additional constraints. See Fig. 1, right picture. One possible constraint is the minimum norm criterion \[8\], \[9\]. Current CDR procedures reconstruct sources separately for each time slice. We present a new CDR method by introducing a temporal constraint, the \textit{smooth temporal activation model}. This constraint also gives insight into the time-dynamics of the sources. CDR techniques lead to linear problems. For a survey of different CDR techniques see [10].

2 The EEG/MEG Forward Model

The fundamental equation governing the interaction of electrical sources \(j\) and the electric field \(\Phi\) is the Poisson equation in connection with a Neumann boundary condition

\[
\begin{align*}
\text{div}(\sigma \nabla \Phi) &= \text{div} j \quad \text{in} \quad \Omega \\
\langle \sigma \nabla \Phi, n \rangle &= 0 \quad \text{at} \quad \Gamma = \partial \Omega. 
\end{align*}
\]

(1)

\(\sigma\) is the conductivity tensor and the open and bounded set \(\Omega\) describes the geometry of the head.

If we put \(\Gamma_0 \subset \Gamma\), the EEG forward model can be specified as follows:

\[
A_E : (H_0^1(\Omega))^3 \xrightarrow{\text{div}} L^2(\Omega) \xrightarrow{S} H^1(\Omega) \xrightarrow{\gamma} L^2(\Gamma_0) \quad j \quad \mapsto \quad J = \text{div} j \quad \mapsto \quad \Phi \quad \mapsto \quad \Phi|_{\Gamma_0}.
\]

(2)

Here \(S\) is the weak solution operator of equation (1), \(\gamma\) is the trace operator [11] from \(\Omega\) to \(\Gamma_0\).
We define

\[(Tf)(x) = \int_{\Omega} f(y) \times \frac{x-y}{|x-y|^3} \, dy\]  

(3)

According to the Biot–Savart law the MEG-measurements are modelled by

\[B(x) = (T(j + \sigma \nabla \Phi)(x), n_x)\]  

(4)

with \(x \in \Gamma_1, \Gamma_1 \cap \bar{\Omega} = \emptyset\) and \(n_x\) is the normalvector in \(x\). Thus the MEG forward model can be sketched as

\[A_M : (H^1_0(\Omega))^3 \xrightarrow{\text{div}} H^1(\Omega) \xrightarrow{j+\sigma \nabla} (L^2(\Omega))^3 \xrightarrow{(T \cdot n_x)} L^2(\Gamma_1)\]  

(5)

**Theorem 1.** The operators

\[A_E : (H^1_0(\Omega))^3 \rightarrow L^2(\Gamma_0)\]  

and

\[A_M : (H^1_0(\Omega))^3 \rightarrow L^2(\Gamma_1)\]  

are compact and have nontrivial nullspaces.

**Proof.** The trace operator \(\gamma\) is compact [11]. \(T\) is compact, as the kernel in (3) is continuous, and \(\Gamma_1 \cap \bar{\Omega} = \emptyset\). The remaining operators in (2) and (5) are continuous [11]. Thus \(A_M\) and \(A_E\) are compact. As the div-operator has a nontrivial nullspace, the same is valid for \(A_E\) and \(A_M\).

**Corollary 2.** The problems

\[A_E j = \Phi_0 \quad \text{respectively} \quad A_M j = B_0\]  

and the combined problem

\[A_E j = \Phi_0 \land A_M j = B_0\]  

are ill-posed.

### 3 The Leadfield Matrix

We discretize \(\Omega\) to \(N\) points \(p_i\). The set \(\{p_i\}\) is called influence space. It is assumed that measurements are given at \(n\) points on \(\Gamma_0 \cup \Gamma_1\). The so-called leadfield matrix \(L \in \mathbb{R}^{n \times 3N}\) contains in the \(i\)-th column the data vector \(m_i\) of size \(n\) belonging to a dipole with moment \((1,0,0)^\top\) located at influence point \(p_i\). Columns \(i+N\) and \(i+2N\) contain data belonging to dipoles with moments \((0,1,0)^\top\) and \((0,0,1)^\top\) respectively.
If we discretize $j$ by $j = \sum_{i=1}^{N} \sum_{k=0}^{2} c_{ik} e_{ik}$, we get

$$Lj = \sum_{i=1}^{N} \sum_{k=0}^{2} c_{ik} L e_{ik}$$

$$= \sum_{i=1}^{N} \sum_{k=0}^{2} c_{ik} (L)_{,i+3Nk}$$

Thus the product $Lj$ yields the data $m$ belonging to the discretized current distribution $j$. In practice $\Omega$ and $\sigma$ can be obtained from MRI pictures. The columns $(L)_{,i}$ of $L$ are computed by the FE method [12]. The elements of the FE-Modell are achieved by CT-pictures of the head. See Fig. 2 for a segmentation based on CT-pictures.

For isotropic $\sigma$ the BE method can be used too [13]. Analytical formulas are only available for simplified geometries [14], [15].

4 Conventional CDR Methods

Source reconstruction methods are classified as dipole reconstruction methods and current reconstruction methods. The first one tries to determine $q$ dipoles at locations $p_i$ with moment $M_i$ such that this configuration explains the measured data. This leads to a nonlinear optimization problem. Current reconstruction methods (CDR methods) try to determine $j$ from data $m$. This is a linear problem.

In practice the number of influence points $N$ is much larger than the number of measurements $n$. Thus the system $Lj = m$ is underdetermined. Due to the ill-posedness of the problem, the matrix $L$ is ill-conditioned. Thus we need additional a-priori information to achieve a unique and stable solution. One possibility is Tikhonov-Phillips regularization [12], [16], [17]:

$$j = \min \arg \left\{ \|Lj - m\|_p^p + \lambda \|Bj\|_p^p \right\} \quad (6)$$
$B$ is called spatial model operator. $p = 2$, $B = I$ yields minimum norm least squares solutions, $p = 2$, $B = \Delta$ is in this field of application the so called LORETA method [18].

This approach doesn’t take temporal information into consideration. Calculations only use one timeslice of data.

5 Spatio-temporal CDR: The Smooth Activation Model

In this section we assume the data $m$ are given as a function of time $m(t)$. Hence the solution $j(t)$ of (1) is also a function of time. Notice that the model is assumed to be stationary, i.e. $\sigma$ is not a function of time. So we can apply the methods discussed in section 4 for each time instance.

A coupling of different time instances is achieved by the assumption that the electrical sources do not change too much. Hence we make the physiologically motivated assumption

$$\left\| \frac{d}{dt} j(t) \right\| \to \min$$

(7)

This is coupled with the approach (6) in the following way. Assume that the data $m_i$ are given for $T$ timeslices $t_i$. The current at time $t_i$ is termed as $j_i$.

5.1 The Smooth Activation Model with Zero Order Spatial Smoothness

Since in most measurements the signal is assumed to be zero before the stimulus onset and likewise after the stimulus, forcing $j_1$ and $j_T$ to be very small is a possible additional condition. This and (7) can be considered as follows:

$$j = \min \arg \{ \| L_T j - m \|^2_2 + \lambda \| Bj \|^2_2 + \mu \| K j \|^2_2 \}$$

(8)

Here

$$j = (j_1, \ldots, j_T)$$

(9)

$$m = (m_1, \ldots, m_T)$$

(10)

$$L_T = L = \text{blockdiag}(L, \ldots, L)$$

(11)

$$B = \text{blockdiag}(I_{3N}, \ldots, 0, I_{3N})$$

(12)

$$K = \begin{bmatrix}
    I_{3N} & -I_{3N} & 0 & \ldots & 0 \\
    0 & I_{3N} & -I_{3N} & \ldots & 0 \\
    \vdots & \ddots & \ddots & \ddots & \vdots \\
    0 & \ldots & 0 & I_{3N} & -I_{3N}
\end{bmatrix}$$

(13)
Thus (8) is equivalent to

\[
\mathbf{j} = \min_{\mathbf{j}} \left\{ \sum_{i=1}^{T} \| L \mathbf{j}_i - \mathbf{m}_i \|_2^2 + \lambda \left( \| \mathbf{j}_1 \|_2^2 + \| \mathbf{j}_T \|_2^2 \right) + \mu \sum_{i=1}^{T-1} \| \mathbf{j}_{i+1} - \mathbf{j}_i \|_2^2 \right\}
\]

(14)

Due to \( \nabla \| A \mathbf{x} - \mathbf{b} \|_2^2 = 2( A^\top A \mathbf{x} - A^\top \mathbf{b}) \), (8) amounts to the regularized normal equation

\[
(L_T^\top L_T + \lambda B^\top B + \mu K^\top K) \mathbf{j} = L_T^\top \mathbf{m}
\]

(15)

The matrix above is positive definite, so an efficient way to solve the problem is the cg-algorithm.

5.2 The Smooth Activation Model with 2nd Order Spatial Smoothness

The second model uses the Laplacian of \( \mathbf{j} \) as spatial model constraint. On a regular grid we consider

\[
\mathbf{j} = \min_{\mathbf{j}} \left\{ \sum_{i=1}^{T} \| L \mathbf{j}_i - \mathbf{m}_i \|_2^2 + \lambda \sum_{i} \| \Delta \mathbf{j}_i \|_2^2 + \mu \sum_{i=1}^{T-1} \| \mathbf{j}_{i+1} - \mathbf{j}_i \|_2^2 \right\}
\]

(16)

Thus we set

\[
B = \text{blockdiag}(\Delta, \ldots, \Delta)
\]

(17)

instead of (12). Again the minimization leads to an equation of the form (15). This is the temporal analog to the LORETA method [18]. LORETA yields spatially stable reconstruction results.

6 Simulations

6.1 The Simple Volume Conductor Model

For testing the properties of the different methods of CDR a simple volume conductor model was chosen.

The simple model setup consisted of an influence space consisting of a 10 by 10 grid with a length of 10 arbitrary units per side centered at (5.5, 5.5, 0.0). Nine sensors were placed in a square planar array with center at (5.5, 5.5, 2) above the grid. See Fig. 3.
We use constant conductivity $\sigma$ inside and outside the head. Thus the Leadfield matrix is obtained by

$$
L_{i,j} = \frac{r_{i,1} - r_{j,1}}{|r_i - r_j|^3}
$$

$$
L_{i,j+N} = \frac{r_{i,2} - r_{j,2}}{|r_i - r_j|^3}
$$

$$
L_{i,j+2N} = \frac{r_{i,3} - r_{j,3}}{|r_i - r_j|^3}
$$

Here $r_i \in \mathbb{R}^3$ is the position of the i-th sensor, $r_j \in \mathbb{R}^3$ is the position of the j-th gridpoint.

### 6.2 Inverse Calculations

Two point sources were placed on the grid and a forward solution using (18) was calculated. To each point a gaussian activation was assigned with a width of $w$ and center at $t_p$:

$$
q(t) = q_0 \exp \left\{ -\frac{(t - t_p)^2}{w^2} \right\}
$$

The simulation was calculated for 16 timeslices.

Two equally oriented dipoles with moment $(0,0,1)^T$ at $x=3$, $x=8$ and both at $y=5$ were placed on the ten by ten grid. A gaussian dipole-strength time-series was assigned to each dipole with peaks at timeslice 5 (dipole 1) and 9 (dipole 2) and a width of $w = 2.5$. See Fig. 4.

The regularization parameter $\lambda$ was chosen by the L-curve criterion [19], [20].

Figs. 5, 6, 7 and 8 show the results of one simulation run. Noise with 30% of the maximal signal amplitude was added to the data. The results show the contour-plots of $\|j(x,y,t)\|$. The white squares in the contour-plots indicate the original source-positions.
Fig. 4. The activation curves of the dipoles

Fig. 5. Temporal reconstruction: 2nd order spatial smoothness (LORETA) and temporal coupling

Fig. 6. Temporal reconstruction: 2nd order spatial smoothness (LORETA) without temporal coupling

Fig. 7. Temporal reconstruction: zero order spatial smoothness and temporal coupling

Fig. 8. Temporal reconstruction: zero order spatial smoothness without temporal coupling
The coupled solutions are significantly less affected by the noise. In the time-range from timeslice 4 to timeslice 10, the coupled solutions also reconstruct the correct number of sources, whereas the non-coupled solutions reconstructs only one source.

In order to get a systematic comparison of the coupled and non-coupled method data sets with different noise-levels were generated. As you can see in Fig. 9 the coupled-solutions achieve in general a better localization error.
7 Summary

It could be shown, that the introduction of temporal constraints to existing CDRs leads to significant improvements in spatial and temporal accuracy. This general advantage of spatio-temporal CDR was shown to be most prominent in the case of noisy data, but also in the ability to separate sources.

References

Signal Correction in NMR Spectroscopy

Heinz-Otto Peitgen\textsuperscript{1}, Tobias Boskamp\textsuperscript{*1}, and Peter Singer\textsuperscript{2}

\textsuperscript{1} CeVis – Center for Complex Systems and Visualization, University of Bremen, Universitätsallee 29, 28359 Bremen, Germany
\textsuperscript{2} Fachhochschule Ingolstadt, Dept. of Engineering Sciences, Esplanade 10, 85049 Ingolstadt, Germany

Abstract. The correction of baseline and phase distortions is an important problem in magnetic resonance spectroscopy. In this work, fast and automatic correction methods based on the adaptive construction of Wiener filters are presented. The proposed methods consist of estimating and classifying the model parameters of the measured spectrum and of approximating correlation operators for the measured and the corrected signal. Results of numerical simulations and applications to real data are given.

1 Introduction

Magnetic resonance spectroscopy (MRS) is a powerful analysis tool with a variety of applications in chemistry, biology and medical diagnostics. In recent years, it has been used as a medical imaging technique in order to obtain spectral decompositions for each examined volume element. This application of MRS could – in principle – provide detailed maps of the concentration of various chemical bonds in e.g. a human brain, an information that cannot be obtained by other methods such as the related techniques of magnetic resonance imaging (MRI) and computed tomography (CT).

Unfortunately, signal distortions like underlying baselines, phase distortions and noise severely complicate the quantitative analysis of MR spectra. Most methods used to compensate these distortions are either not suitable for in vivo spectra, or require user interaction and thus cannot be used for automatic correction of large data sets.

The goal of this project is to develop fast, automatic and accurate correction algorithms of such distortions, with a special focus on in vivo spectra and on medical applications. The project has been carried out in cooperation with Bruker Daltonik GmbH at Bremen (formerly Bruker-Franzen Analytik GmbH) and the research group of Prof. Leibfritz, Institute for Organic Chemistry at the University of Bremen.

1.1 Magnetic Resonance Spectroscopy

The magnetic resonance phenomenon, discovered in 1946, is a quantum mechanical effect that affects atomic nuclei with an angular momentum (or

\textsuperscript{*} E-Mail: tboskamp@cevis.uni-bremen.de

W. Jäger et al. (eds.), Mathematics - Key Technology for the Future
© Springer-Verlag Berlin Heidelberg 2003
spin) placed in an external magnetic field. If such nuclei are exposed to a high frequency electromagnetic excitation pulse, they respond by emitting a signal known as the free induction decay (FID), with a strength proportional to the number of responding nuclei. The frequency of the response signal (the resonance frequency) depends on the type of the nuclei and the magnitude of the magnetic field. Since the latter is also influenced by the molecular surrounding of the nuclei, the resonance frequency is not the same for all nuclei of a given type. Instead, nuclei in different chemical bonds emit signals with different resonance frequencies. This so called chemical shift forms the basis of magnetic resonance spectroscopy.\footnote{For a more detailed introduction to the principles of MRS and some of its applications, see e.g. [1].}

The time-domain FID signal \( f(t) \) obtained by a single MRS experiment is commonly described by the Lorentz model as a sum of \( K \) exponentially decaying sinusoids in complex form,

\[
f(t) = \sum_{k=1}^{K} c_k e^{(-\alpha_k + i\omega_k)t} + n(t) ,
\]

with complex amplitudes \( c_k \), positive damping factors \( \alpha_k \), resonance frequencies \( \omega_k \) and additive noise \( n(t) \) (\( i \) denotes the imaginary unit \( \sqrt{-1} \)). Applying the Fourier transform, we obtain the spectrum

\[
s(\omega) = \hat{f}(\omega) = \sum_{k=1}^{K} \frac{c_k}{-z_k + i\omega} + \hat{n}(\omega) ,
\]

where \( z_k = -\alpha_k + i\omega_k \) denotes the complex evolution factor of the \( k \)-th component.

The Lorentz model is often only a crude approximation to the real signal, especially for in vivo or imaging measurements. For this reason, other signal models, such as the Gaussian model, where the decay factor in (1) is replaced by \( \exp(-\beta_k t^2) \), or the Voigt model with a decay term \( \exp(-\alpha_k t - \beta_k t^2) \), are also considered [2].

For in vivo spectroscopy, the individual peaks in the spectrum usually overlap, so that smaller peaks are distorted by underlying broad components with large \( \alpha_k \), or by the tails of dominant components with large absolute amplitudes \( |c_k| \). In addition to this baseline (Fig. 1), the amplitudes \( c_k \) have different complex phases, which disturb the measurement of peak heights and areas in the spectrum (Fig. 2).

To obtain a representation of the true corrected spectrum, we assume that the measured spectrum is given by (2) and that we can classify the individual signal components as part of either the baseline or the net signal, which we want to evaluate. The characterization of the net signal and baseline has to be guided by a-priori knowledge of the experimental setup and by the signal...
Fig. 1. The baseline distortion superimposes relevant peaks in the spectrum

Fig. 2. Real parts of a pure peak (left) and a phase distorted peak (right)

components to be examined. Usually, this is done by specifying a frequency range or a damping range (or both), outside of which a component is viewed as part of the baseline. If we let $SIG \subset \{1, \ldots, K\}$ denote the set of indices belonging to the net signal, the true corrected spectrum can be written as

$$s_{\text{true}}(\omega) = \sum_{k \in SIG} \frac{|c_k|}{-z_k + i\omega}.$$  \hspace{1cm} (3)

2 Baseline Correction

The baseline correction problem for a measured spectrum $s$ consists of finding the net signal $x$ and the baseline $y$, such that $x + y = s$. From a stochastic (or Bayesian) point of view, $x$, $y$ and $s$ may be regarded as samples from stochastic processes $X$, $Y$ and $S = X + Y$. The task then is to find "good" estimates for $x$ based on the measurement $s$ and some information about the processes $X$ and $Y$.

Problems of this type are the subject of filter theory, where one typically asks for the best linear pointwise estimator $\hat{x}(t)$, based on previously observed data $s(\tau), \tau \leq t$. For stationary signals, this leads to the classical Wiener filter, expressed in terms of the correlations of $X$ and $S$ (see e.g. [3]).

In the present application, the signal $s$ must not be considered as a series of distinct, correlated events, but as a single event. Hence, we are not interested in filtering with respect to a fixed time $t$, but rather in the estimation of the net signal $x$ as a whole. An appropriate framework for this problem is the Wiener filter theory for stochastic processes over Hilbert spaces, which we briefly outline in the following, omitting some of the technicalities (see [4,5] for details).

2.1 The Wiener Filter

Let $H$ be a complex Hilbert space and denote the inner product of two elements $u$ and $v$ of $H$ by $v^* u$. An element of a stochastic process $W$ over $H$
is given by a linear functional \( w \),
\[
w : H \rightarrow \mathbb{C}, \quad u \mapsto w(u).
\] (4)

Since the elements of \( W \) are not required to be bounded, they may not be representable by elements of \( H \). For convenience, we use the notations \( w^*u = w(u) \) and \( u^*w = \overline{w^*u} \). The autocorrelation of \( W \) is given by the linear, self-adjoint and positive semidefinite operator \( R_{WW} \), defined by
\[
u^* R_{WW} v = E(u^*W)(W^*v) \quad u, v \in H,
\] (5)

where \( E \) denotes the expected value. Similarly, the cross-correlation \( R_{WQ} \) of two processes \( W \) and \( Q \) is given by \( u^* R_{WQ} v = E(u^*W)(Q^*v) \).

Using these definitions, a best (in the mean square sense) linear estimate \( \hat{x} \) for the net signal \( x \) can be obtained by
\[
\hat{x} = R_{XS} R_{SS}^+ s,
\] (6)

where \( R_{SS}^+ \) denotes the Moore–Penrose inverse (or pseudo-inverse) of \( R_{SS} \). In the case where \( R_{SS} \) is invertible, \( R_{SS}^+ \) is equal to the inverse, \( R_{SS}^{-1} \), and the solution (6) is unique. Otherwise, there exist other best estimators beside the solution \( \hat{x} \).

2.2 The Estimated Wiener Baseline Filter

To make use of the Wiener filter approach described in the previous section, we need to know the correlation operators \( R_{SS} \) and \( R_{XS} \). However, we are given only a single spectrum \( s \), hence no statistical properties of \( S \), let alone \( X \), can be directly measured. Instead, we have to develop models for the stochastic processes \( S \) and \( X \) to obtain estimates for their correlation operators. Based on these considerations, we propose the estimated Wiener filter method, which basically consists of three steps:

- Estimate system order \( K \) and parameters \( z_k \) and \( c_k \) of the measured signal \( s \),
- construct models for the stochastic processes \( X \) and \( S \), and
- use the correlation operators to compute the Wiener filter.

In the following, we briefly describe each of these steps (see Fig. 3). For details, we refer to [5].

Signal Parameter Estimation. For the first step in this process, the parameter estimation, we use the HSDV method [6], which takes advantage of the special structure of the Lorentz model. Motivated by a state-space approach, the HSDV method computes the singular value decomposition (SVD) of a Hankel matrix formed by the discretized time-domain signal \( f \).
The number of significant singular values equals the system order $K$, and the evolution factors $z_k$ can be obtained from the corresponding left singular vectors. The amplitudes $c_k$ are finally computed by solving a linear least squares problem.

The most time consuming part of this method is the SVD computation. To obtain good results, the matrix size should be at least $100 \times 100$, a larger size increases the accuracy of the estimated parameters. On the other hand, only the first $K$ (typically, $K < 20$) singular values and left singular vectors are needed. For this reason, we can speed up the computation by about a factor of 10 by using the iterative Lanczos SVD algorithm, which can be terminated as soon as enough singular values and vectors are calculated [7,8].

As a result, we obtain estimated amplitudes $c_k$, evolution factors $z_k$ and an approximation $s_0$ to the measured spectrum $s$, given by

$$s_0(\omega) = \sum_{k=1}^{K} \frac{c_k}{-z_k + i\omega}.$$  \hspace{1cm} (7)
**Stochastic Model Construction.** Even if the measured spectrum strictly matches the Lorentz model (2), the estimated parameters $z_k$ and $c_k$ are likely to differ from the true values, due to the presence of noise. We therefore assume that the estimated parameters are samples from independent, complex valued random variables $Z_k$ and $C_k$, with real($Z_k$) < 0 with probability 1 and the following expected values existing:

$$
E(C_k) = c_k, \quad E|C_k - c_k|^2 = \sigma^2_{c_k}, \quad E(Z_k) = z_k, \quad E|Z_k - z_k|^2 = \sigma^2_{z_k}.
$$

This provides a model for the stochastic process $S$ in the frequency domain,

$$
S(\omega) = \sum_{k=1}^{K} \frac{C_k}{-Z_k + i\omega}.
$$

An empirical study of the HSVD method reveals a relation between the parameter variances $\sigma^2_{z_k}$ and $\sigma^2_{c_k}$, the estimations $z_k$ and $c_k$ and the noise level of the signal. This way, we can obtain approximations to the parameter variances by measuring the noise level of the signal.

To obtain a model for the stochastic net signal process $X$, we assign the estimated parameters to either baseline or net signal, as indicated in Sect. 1.1. This classification is governed by a-priori rules (e.g. “large $\alpha$ is baseline, small $\alpha$ is net signal”), as well as an adaptive mechanism that adjusts the parameter threshold based on the estimated parameter values. As a result, we can represent $X$ by

$$
X(\omega) = \sum_{k \in J_x} \frac{C_k}{-Z_k + i\omega},
$$

where $J_x$ denotes the set of indices assigned to the net signal. Moreover, we let $x_0$ denote the net signal approximation described by the $z_k, c_k$ with $k \in J_x$.

**Wiener Filter Computation.** Approximate expressions for the correlation operators $R_{SS}$ and $R_{XS}$ can be derived by making first order Taylor expansions of (9) and (10). The resulting expressions contain the signal approximations $s_0$ and $x_0$, as well as the parameter variances $\sigma^2_{z_k}$ and $\sigma^2_{c_k}$. However, in the computation of the Wiener filter according to (6), the parameter variances cancel out. Hence, the resulting filter operator is independent of the parameter variances.

In addition, the null space of the filter operator is orthogonal to the space spanned by the partial derivatives of $S$ with respect to the $Z_k$ and $C_k$. As a consequence, the above filtering approach eliminates those parts of the signal that are not covered by the signal model underlying the representations (9) and (10). This may be desirable in cases where the signal model is accurate,
but it is certainly not acceptable if the model is known to be only a crude approximation, as, for instance, in in vivo MRS. For this reason, we propose a different approach, introducing the concept of diagonal filter operators.

The Diagonal Wiener Filter. Let $L : H \rightarrow \widetilde{H}$ be a linear bijection into some function space $\widetilde{H}$. An operator $G : H \rightarrow H$ is called diagonal with respect to $L$, if there is an element $g \in \widetilde{H}$, such that

$$Gv = L^{-1}(g \cdot Lv) \quad \text{for all} \quad v \in H,$$

where "·" denotes pointwise multiplication. In many typical applications of the Wiener filter theory, the resulting filters are of this type (most often with $\widetilde{H} = H$ and $L$ being the Fourier transform), and diagonal filters also have been proven useful for baseline correction [9].

In the following, we let $L$ be a fixed, linear isometry and additionally assume that the samples of $S$ and $X$ are elements of $H$. We restrict the optimization problem of Sect. 2.1 to operators that are diagonal with respect to $L$, i.e., we want to find a function $g \in \widetilde{H}$, such that

$$E\|LX - (g \cdot LS)\|^2 \rightarrow \min .$$

The solution to this optimization problem is the diagonal Wiener filter, where the minimizing function $g$ is given by

$$g = \frac{E(LX \cdot LS)}{E|LS|^2} .$$

As before, the choice of $L$ must be governed by a-priori knowledge about $S$ and $X$. In our numerical experiments, we chose $L$ to be the Fourier transform, in which case the filter function $g$ is given by

$$g(t) = \frac{\hat{x}_0(t) \cdot \hat{s}_0(t) + \sum_{k \in J_x} (\sigma_{z_k}^2 |c_k|^2 t^2 + \sigma_{c_k}^2) e^{2\alpha_k t}}{|\hat{s}_0|^2 + \sum_{k=1}^K (\sigma_{z_k}^2 |c_k|^2 t^2 + \sigma_{c_k}^2) e^{2\alpha_k t}} \quad (t \leq 0).$$

2.3 Results

To test the quality of the described baseline correction methods, we performed a series of numerical simulations with test signals generated from randomly chosen parameters. In order to test the effect of deviations from the ideal signal model, we generated mixtures of signals constructed according to a number of different models, including the Lorentz, Gauss, Voigt and linear model. The model distance, i.e. the relative $L_2$-distance between the generated signal and the corresponding ideal Lorentz signal, has been varied from 0 to

\[ \text{By this choice of } L \text{ the filtering can be performed in the time domain, prior to computing the spectrum.} \]
0.5. The simulated signals have been disturbed by white Gaussian noise of a prescribed noise level. For each combination of noise level and model distance, 3000 test signals have been generated.

We compared our results to the estimated optimal filter, which we proposed earlier [9], and to the common method of baseline removal [6], where the estimated baseline $s_0 - x_0$ is subtracted from the measured signal $s$. This method is known to give good results when the signal model is accurate.

The results for two different noise levels are shown in Fig. 4. As can be seen, the general Wiener filter method yields poor results for medium to large model distances since it is not able to handle possible deviations from the signal model. The diagonal Wiener filter, on the other hand, does account for these deviations, although they have not explicitly been included in the underlying stochastic models. The estimated optimal filter, although constructed by heuristic considerations, shows good performance for the case of large model distances.

An example for the baseline correction of real data$^3$ is given in Fig. 5. In a first step, the dominant water resonance has been suppressed by performing a parameter estimation and subtracting components with a resonance frequency near zero. The remaining spectrum has been corrected with the diagonal Wiener filter as described above.

### 3 Phase Correction

The problem of correcting the phase distortion of a spectrum is of special importance in spectroscopic imaging, where computationally expensive methods of peak quantification are not feasible because of the large number of measurements. Instead, the peak height of the real spectrum $\text{real}(s)$ can be used

$^3$ In vivo spectrum from a human brain, measured using the PRESS sequence and an echo time of 30 ms. With kind permission by Dr. Stefan Roell, Siemens AG.
to determine the concentration of a chemical bond, if the phase distortion is properly corrected.

One of the methods proposed for automatic phase correction is the ZOE algorithm [10], which consists of the selection of two peaks in the spectrum and of applying a linear phase correction such that the imaginary parts of the integrals over each of the two peaks vanish.

We have extended this method to incorporate more than one peak, which makes it more reliable, and were able to replace the iterative determination of the linear phase coefficients by a direct computation, speeding up the method by a factor of about two.

However, this method is not well applicable for in vivo spectra with strongly overlapping peaks and high noise level. Fig. 6 shows a comparison between manual correction (top spectrum), where linear phase coefficients were adjusted manually for maximum peak symmetry, and phase correction by our enhanced ZOE method (middle spectrum). The weaker peak symmetry in the ZOE corrected spectrum indicates that the phase distortion is not completely removed. This is due to the signal noise, which disturbs the proper selection of integration areas and thus affects the integral approximations.

As an alternative, we investigated the application of the Wiener filter methods developed in the previous section to the phase correction problem. A similar approach, where we have replaced the stochastic model (10) for the net signal, $X$, by

$$X(\omega) = \sum_{k=1}^{K} \frac{|C_k|}{-Z_k + i\omega}$$

and applied the diagonal Wiener filter in the frequency domain, yields results competitive to the manual phase correction, i.e. the peaks in the real part of the spectrum are adjusted to the same high symmetry (Fig. 6, bottom).
4 Conclusion

The proposed methods of estimated Wiener filtering allow fast and automatic correction of baseline and phase distortions of MR spectra. The results of our numerical simulations and experiments indicate that the proposed methods outperform other automatic correction methods found in the literature, especially in the case of medium to large model distance.

To allow the experimental evaluation under realistic conditions, we integrated the baseline correction methods into the software packages WinNMR and WinMRI, two MR data analysis tools offered by our cooperation partner Bruker Daltonik GmbH. In addition, these methods are currently evaluated by the group of Prof. Leibfritz. First results of these evaluations corroborate the results presented above.

References

On Scattering of Ultrasonic Waves

Peter Mathé and Johann Hinrich Zacharias-Langhans

Weierstrass-Institute for Applied Analysis and Stochastics
Mohrenstraße 39, 10117 Berlin, Germany
E-mail: mathe@wias-berlin.de, langhans@wias-berlin.de
URL: www.wias-berlin.de

Abstract. We report investigations regarding the possibility for improving ultrasonic imaging by an analysis of pulse deformations. Starting with a stochastic solution of the lossy wave equation, we present a general approach for perturbing stochastic processes on Lie groups by a Poisson process. Applications to radiation transport are indicated.

1 Introduction

Many applications of ultrasonic technology aim at determining certain parameters from the interior of some medium, which cannot be measured directly, we mention areas like medical diagnostics and nondestructive testing. Often the device works analogous to a radar. Short pulses, usually of high frequency and a few wavelengths, are sent through the sample, are reflected at interior inhomogeneities and registered by a receiver, often identical with the transmitter. In this way one recovers essentially reflection properties of the medium, whereas parameters like absorption or velocity are usually measured by transmission. The latter is technically more involved and it is difficult to distinguish between losses caused by absorption or scattering.

The underlying hypothesis of this project claims, that shape deformations occur due to scattering and absorption in the sample material, thus allowing to obtain inference on such material parameters, also. This information should be extracted by partitioning the signal and investigating separate parts. Then, taking into account this additional information, one could design electronic devices, providing ultrasonic images with more details. Our investigations were initiated by practical experiences in the area of laser tomography, see R. Willenbrock [12] and acoustical experiments as carried out by G. Bayer, H. Kuttruff und T. Niedertränk, see [1]. A mathematical problem in this direction can be formulated as follows: Can we use telegrapher’s equation to describe propagation of ultrasonic waves in highly scattering media?

Below, we will sketch some physical background and describe details of the experiments in Sects. 2 and 3. In Sect. 4 we outline some results concerning random motion for solving telegrapher’s equation, originally due to M. Kac. We emphasise intrinsic difficulties when switching from one spatial dimension to higher dimensions. It turns out, that in general such processes are governed
by transport type equations, which differ from wave type ones, except in the one-dimensional case.

2 Background

In pulsed laser technology the front and back parts of received (after transmission) signals are evaluated in order to determine the parameters of scattering and absorption, respectively, see [12]. This approach is based on the model of diffusion. Such a diffusion model could also be validated for ultrasonic wave propagation in the presence of many centers of scattering, see Bayer et al. [1]. In either case a point source is assumed. Measurements by Weiland, Giese and Hundertmark [11] with time resolved pulses proved deformation of the pulse form. These deformations were interpreted as a low pass effect of the medium. Nevertheless these authors use maxima and area of the envelopes to determine material parameters.

Presumably such a situation can be explained in terms of stochastic models for the media. Actually, for certain instances such models are intensively investigated and related to certain phenomenological equations. In particular, the telegrapher's equation, the O’Daherty and Anstey equations and equations of radiation transport can be derived as appropriate scaling limits.

The theory of O’Daherty and Anstey is capable to predict pulse deformation in the form of a Gaussian low pass filter and may support the experiments by Weiland et al. Moreover, this theory implies a time delay in addition to the sound velocity of the medium, which should be taken into account. If scattering intensity increases, the equation of radiation transport may well approximate the propagation behavior. But it cannot explain other recently discovered effects, as there are phase and phenomena of interference, see [3]. Below we shall study a generalization of the equation of radiation transport which occurs in a natural way on a larger phase space. This allows to incorporate more information than just the wave number vector. At present it is not clear, whether this additional information can be related to the phase in a proper way.

3 Experiments

We are going to shortly explain the equipment used for measurements, a typical A-scan, see Fig. 1 below.

Cylindrical sample media Plexiglas (PG) or Polyethylene (PE) of different size were placed on a device, used to fix the transducer at the sample, which in turn was connected to an oscilloscope and a controllable source of voltage. Signals were directed through an analog–digital converter to a computer and saved for further investigation. For this purpose we developed specific software, which enables visualization and various analysis of the measured or simulated signals in both time and frequency domain.
It turned out, that a direct comparison between transmitted and received signals is useless unless having regard to hardly predictable factors. In particular the following has significant influence on shape deformations:

- The frequency range of the receiver (in fact identical with the transducer) is band-limited.
- There is no canonical coupling between transducer and sample: this causes severe intensity fluctuations.
- There may occur field effects. In our case, the signal stabilized, in dependence of frequency and size of the sender, after 4–8 cm at a frequency of 1 MHz and 2 cmØ or 4 MHz and 1 cmØ, respectively.
- Geometry of the sample.

After modifying the apparatus and introducing reference measurements in water these influences could be limited.\footnote{There are attempts to principally alter the apparatus. In cooperation with G & W Instruments, Berlin equipment for under-water measurements was ordered. Due to technical difficulties this is still not working, such that there are no new measurements available at the moment.}

Let us briefly describe a series of measurements, where the signals travelled first through an area of water, wide enough for the pulses to stabilize. We realized, that the form of the signals was not influenced in a significant way by the different materials. Actually, this is consistent with the telegrapher's equation, which predicts a diminution for signals with total mean zero, but nearly no deformation. This means, we cannot observe effects of multiple scattering for these materials. When working with other materials, like aluminium, the scattering centers could always resolved, which implied, that we were still in the regime of single scattering, as is shown in Fig. 2 below. Finally, we tried organic samples. But here, problems in particular that of coupling, prevented stable and reproducible measurements, so that these experiments must be postponed until better equipment is available.
In many applications it is not possible to work with this fore-run, and the pulses are subject to deformations, depending on the equipment. In this case, a splitting of the signals may be useful. It turned out, that the front part of the received signals underlies much weaker fluctuations than the back-part, or even than the position of the maximum. For example, in determining absorption coefficients, the evaluation of the leading edge gave up to three times lower deviations, than the evaluation of the whole signal. A splitting of the signals and the comparison to specific reference signals of water or other samples stabilize the fluctuations of the pulse and thus may improve image quality.

4 Monte Carlo Methods for Solving the Telegrapher’s Equation

Below we shall study some attempts made to solve the following equation

\[
\frac{\partial^2}{\partial t^2} u(t, x) + 2\lambda \frac{\partial}{\partial t} u(t, x) = \Delta u(t, x), \quad x \in \mathbb{R}^d, \ t \geq 0
\]

\[
u(0, x) = u_0(x) \quad \text{and} \quad \frac{\partial u}{\partial t} (0, x) = 0,
\]

which is known as telegrapher’s equation (with unit speed).

4.1 The One-Dimensional Case

First we exhibit the one-dimensional case in some detail to explain the proposed approach. Let \( \lambda \) be the scattering intensity from (1). This gives rise to
a homogeneous Poisson process \( \mathcal{N}_t \) with intensity \( \lambda \), see [8], i.e.,

\[
P(\mathcal{N}_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.
\]

The corresponding processes

\[
T_t = \int_0^t (-1)^{\mathcal{N}_s} ds \quad \text{and} \quad v_t = (-1)^{\mathcal{N}_t}, \quad t \geq 0,
\]

(2)
can be interpreted as location and (velocity) direction of a particle moving with constant velocity along the real axis, which reverses its speed at random moments \( t_k = \inf\{t : \mathcal{N}_t \geq k\} \).

In [4] M. Kac proved, that (1) is solved by the function

\[
\phi(x, t) := \frac{1}{2} \mathbb{E}\{f(x - T_t) + f(x + T_t)\}.
\]

In higher dimensions there is a respective relationship, \( \phi(x, t) = \mathbb{E}\psi(x, T_t) \), between the solution \( \phi \) of the telegrapher’s equation and the solution \( \psi \) of the corresponding wave equation, in which \( \lambda \) is simply set to 0. We refer to M. Kac [4], S. Kaplan [5], and A. Z. Veselovskaya [10] for more details.

For our further reasoning it is important to note, that in one space dimension the telegrapher’s equation is (up to regularity assumptions) equivalent to the following hyperbolic system

\[
\begin{align*}
\partial_t u_1 &= -\partial_x u_1 + \lambda (u_2 - u_1) \\
\partial_t u_2 &= \partial_x u_2 + \lambda (u_1 - u_2) \\
\end{align*}
\]

(3)
in the sense, that \( \phi = u_1 + u_2 \) solves the telegrapher’s equation, if \( u_1 \) and \( u_2 \) solve (3).

There is an efficient method to solve (3), based on the following observation due to J. Kisynski [6]. If we endow the phase space \( \mathcal{G} = \mathbb{R} \times \{1, -1\} \) of the particles with the operation \( (s, u) \circ (t, v) = (s + ut, uv) \), then \( \mathcal{G} \) is a non-commutative group, and the compound process \( g_t := (T_t, v_t) \), obtained from (2), acts on \( \mathcal{G} \) as a homogeneous Markov process, having independent identically distributed increments, hence is a Lévy process [2]. We can assign a continuous semi-group \( (U_t)_{t \geq 0} \) of operators with generator \( H \), which composes in our case as \( H = H_0 + \lambda H_1 \), with \( H_0 \) being the generator of the group of translations \( (t, 1)_{t \in \mathbb{R}} \) and \( H_1 \) describing the Poisson reversion of velocity. In detail, with \( g = (T, v) \) we have

\[
H_0 u(g) = \lim_{\varepsilon \to 0} u \ast \frac{1}{\varepsilon} (\delta_{(\varepsilon, 1)} - \delta_{(0, 1)})(g) = v \frac{\partial u(T, v)}{\partial T}.
\]
and
\[ H_1 u(g) = u * (\delta_{(0,-1)} - \delta_{(0,1)})(g). \]

Above, \( \delta \) denotes the Dirac measure and \( * \) the convolution on \( G \). If we now let \( u(t,g) = E u_0(g \circ g_t^{-1}) = U_t u_0(g) = u_0 * \mu_t(g) \), then \( u(t,g) \) solves the following evolution equation
\[ \partial_t u = H u \quad \text{with} \quad u|_{t=0} = u_0(g) = \frac{1}{2} f(T), \]
equivalent to (3). We have thus indicated, that the random motion \( (g_t)_{t \geq 0} \) is governed by the telegrapher’s equation. Moreover, this allows to base a numerical scheme by Fourier transform. Since irreducible representations of \( G \) are two-dimensional and can be parametrized by \( \alpha \in \mathbb{R} \), the solution \( u(t,g) \) can efficiently be computed by Fourier transform, knowing that
\[ \tilde{\mu}_t(\alpha) = \exp t \left[ \begin{pmatrix} -i\alpha & 0 \\ 0 & i\alpha \end{pmatrix} \right] + \lambda \left[ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \right]. \]

When used with standard FFT, this solves the discretized (in \( n \) steps) telegrapher’s equation in \( n \log(n) \) operations, in favor of difference schemes, which typically require a number of operations proportional to \( n^2 \).

4.2 Random Walks in Two Dimensions

Here we shall study the random walk, introduced by E. Orsingher [7], designed to solve the two-dimensional telegrapher’s equation. We aim at proving, that random walks as proposed by E. Orsingher cannot solve equations of telegraph type. Instead they are governed by equations of transport type.

To this end, the walk is first described on a lattice \( h \mathbb{Z}^2 \) with mesh size \( h \). As Orsingher claims, letting \( h \to 0 \), the limiting walk solves the two-dimensional telegrapher’s equation. Here we shall not provide further details. Instead we are looking for the equation which governs the random walk by transferring it to the group \( G := \mathbb{C} \times \Gamma \), where \( \Gamma \) denotes the circle, and \( G \) is equipped with operation
\[ (z, \alpha) \otimes (y, \beta) = (z + e^{i\alpha}y, \alpha + \beta). \]

Actually Orsingher’s walk, say \( g_t^{(h)} \), acts on the discrete sub-group \( h \mathbb{Z}^2 \times \{0, \pi/2, \pi, 3/2\pi\} \).

Let the random variable \( \gamma \) be \( G \)-valued with distribution
\[ \mu_\gamma = \frac{1}{2} \{ (1 - \lambda h)(\delta_{((h,0),0)} + \delta_{((0,h),0)}) + \lambda h(\delta_{((h,0),\pi/2)} + \delta_{((0,h),-\pi/2)}) \}. \]

At discrete times \( t = nh \), \( g_t^{(h)} \) is distributed according to the \( n \)-fold convolution of \( \mu_\gamma \), hence gives rise to a Lévy process on \( G \), again. The generator \( H_h \)
of this process is determined by convolution

$$H_h f = f * \frac{1}{h} (\mu \gamma - \delta_{(0,0)})$$.

Again this decomposes into a transport and scattering part, resulting in a corresponding evolution equation, which is a system of four coupled difference equations. In the limit, as \((h \to 0)\), the processes \(g_t^{(h)}\) converge to a continuous Lévy process, say \(g_t\) on \(\mathcal{G}\) and the generators \(H_h\) converge to some \(H\). If we denote by \(\langle e^{i\alpha}, \nabla \rangle\) the directional derivative in direction \(\alpha\), \(e\) the neutral element of \(\mathcal{G}\) and \(\sigma\) the measure \((\delta_{(0,\pi/2)} + \delta_{(0,-\pi/2)})/2\), then we can represent this limiting operator as

$$Hf(z, \alpha) = -\langle e^{i\alpha + \pi/4}, \nabla \rangle f(z, \alpha) + \lambda \int_{\mathcal{F}} d\sigma(\beta)f(z, \alpha - \beta) - f(z, \alpha)$$

$$= H_0 f(g) + \lambda f * (\sigma - \delta_e)(g).$$

Again, accompanied with initial values, this system can efficiently be solved by Fourier transformation. But, in contrast to the one-dimensional case, it is not equivalent to any telegrapher’s equation, as can readily be seen at the simple instance \(\lambda = 0\). Instead, this is a specific equation of transport type.

### 4.3 Group Structure of Radiation Transport

Both processes from above can be considered as special cases of more general processes on Lie groups \(\mathcal{G}\), determined by an underlying deterministic or stochastic movement, perturbed by Poisson shocks.

More precisely, let \((g_0^0, \Omega_0)\) be a Lévy process on \(\mathcal{G}\) with generator \(H_0\), and denote by \(\mu_0\) the corresponding convolution semi-group. Moreover, we let \((\mathcal{P}, \Omega_\pi)\) be a simple Poisson process on \(\mathcal{S} = \mathbb{R}^+ \times \mathcal{G}\), having intensity measure \(\rho\). The marginal process \(N(\cdot) = \mathcal{P}(\cdot \times \mathcal{G})\) shall also be simple with intensity \(\lambda(\cdot) = r \lambda(\cdot \times \mathcal{G})\), such that with \(\mathcal{N}_t := \mathcal{N}(0, t]\) we have

$$\mathbb{P}(N_t = k) = e^{-\lambda(0,t]} \sum_k \frac{\lambda(0,t]_k}{k!}.$$

Path-wise, i.e., given \(g_0^0\) and \(\mathcal{P} = \sum \delta(t_i, \gamma_i)\) with \(0 < t_1 < t_2 \ldots\), we define \(g_t\) on \(\Omega_0 \times \Omega_\pi\) by

$$g_t = (\prod_{i=1}^{N_t} g_{t_{i-1}, i}^{\gamma_i}) g_{t_{N_t}, t},$$

with the convention \(t_0 := 0\) and \(g_{s,t} := (g_s^0)^{-1} g_t^0 = d g_{t-s}^0\). It can be seen, that \(g_t\) has independent increments which implies for the transitions

$$\mu_{s,t} = \mu_{s,r} * \mu_{r,t} \quad (0 \leq s \leq r \leq t).$$
The intensity $\rho$ admits for measurable sets $B \in \mathbb{R}^+$ the desintegration

$$\rho(B \times \cdot) = \int_B \sigma_t(\cdot)d\lambda(t)$$

for a certain transition kernel $\sigma_t$. If $\lambda$ and $\sigma_t$ are independent of $t$, then the increments of $g_t$ are identically distributed.

The following results show a perturbation expansion for the transition kernels and characterize the generator of the compound process $g_t$.

**Theorem 1.** Let $0 < s < t < \infty$, $S = S_{s,t} = (s, t] \times \mathcal{G}$ and

$$\mu^n_{s,t}(\cdot) = \mathbb{P}(g^{-1}_{s,t} \in \cdot | \mathcal{N}(S) = n)\mathbb{P}(\mathcal{N}(S) = n), \quad (n \in \mathbb{N}_0).$$

Then $\mu_{s,t}$ decomposes as $\mu_{s,t} = \sum_{n=0}^{\infty} \mu^n_{s,t}$. Moreover the following representation

$$\mu^n_{s,t} = \int_s^t \mu^{n-1}_{s,h} \star \sigma_h \star \mu^0_{h,t} d\lambda(h) \quad (n \geq 1)$$

is valid.

**Theorem 2.** Let $g_t$ from above and assume that $\lambda$ has a density, again denoted by $\lambda$. Then there is $D \subset \text{Dom}(H_0)$, dense in $\mathcal{C}_0(\mathcal{G})$, such that for any $f \in D$ the function $u : \mathbb{R}^+ \times \mathcal{G}$, defined as

$$u(t, g) = \mathbb{E}f(g \circ g_t^{-1})$$

solves the following initial value problem.

$$\partial_t u = -H_0 u + \lambda(t) u * (\sigma_t - \delta_e)$$

$$u|_{t=0} = f$$

In one spatial dimension ($\mathcal{G} = \mathbb{R} \times \{1, -1\}$) this covers Kac’ result with time-dependent scattering intensity, first proven by S. Kaplan [5]. In two dimensions and with $\mathcal{G} = \mathbb{R}^2 \times \{0, \pi/2, \pi, 3\pi/2\}$, we reproduce Orsingher’s walk. In higher dimensions, (5) often turns to a radiation transport equation. Let $\mathcal{G} = \mathbb{R}^d \times A$ be the semi-direct product of $\mathbb{R}^d$ and $A \leq \text{GL}_d$, acting on $\mathbb{R}^d$. If the unperturbed process $g_t^0 = (t \cdot e_1, id)$, $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$ consists of translations, then the generator is the directional derivative

$$H_0 f|_{g=(x,A)} = \lim f * \frac{1}{h} \left( \delta_{(h \cdot e_1, id)} - \delta_{(0, id)} \right)(g)$$

$$= \lim \frac{1}{h} \left( f(x - h \cdot A \cdot e_1, A \cdot id) - f(x, A) \right)$$

$$= -\langle A \cdot e_1, \nabla_x \rangle f(x, A).$$

The particle’s velocity in state $g = (x, A)$ is $v = A \cdot e_1$. If the scattering measure $\sigma$ is concentrated on $\{(x, A) \in \mathcal{G} : x = 0\}$, then paths are a.s.
continuous. If, in addition, $\sigma$ is conjugate-invariant, $\sigma(A^{-1}gA) = \sigma(g)$, $g \in \mathcal{G}$, then, turning to conditional expectations

$$\phi(t, x, v) := E_{\theta}(u(t, x, A) \mid A e_1 = v),$$

where $\theta$ is the normalized Haar measure on $\mathcal{A}$ and $K(v, \mathcal{V}) := \sigma(\{B : B^{-1}v \in \mathcal{V}\})$ ($\mathcal{V} \subset \mathbb{R}^d$ measurable), we arrive at

$$\partial_t \phi(t, x, v) = -\langle v, \nabla_x \rangle \phi(t, x, v) + \lambda(t) \left( \int \phi(t, x, v') K(v, dv') - \phi(t, x, v) \right),$$

the more usual form of a transport equation.

5 Conclusions

According to the experiments, reported in Sect. 3, it does not seem to be possible to extract additional valuable information from the shape of the signals. The sample materials PE and PG do not show notable scattering. The observed shape deformations result from effects, caused by the experimental device. Telegrapher's equation is capable to describe ultrasonic propagation in these materials, and does not predict shape deformations.

This equation cannot be solved by Kac's stochastic approach if the parameter $\lambda$ is space dependent. Random processes, as studied above, will not be governed by wave-type equations, unless the spatial dimension is one. Finally, as we have indicated above, the random walk proposed by Orsingher is not governed by telegrapher's equation, but by some specific transport equation.

The analysis of the random walks is naturally carried out on groups, which enlarge the physical phase space. It is open, whether this additional information can be properly related to phenomena of interference in the scattering regime.

References


Smoothing of Tomographic Data and Hybrid Volume-Surface Visualisation

Willi Jäger and Cătălin Dărțu

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany, jaeger@iwr.uni-heidelberg.de

Abstract. Visualisation of higher dimensional data sets is an important tool for their analysis and most important in medical diagnosis and therapy. Starting from application in dental medicine and facial surgery, this project had as an aim to improve the methods of volumetric image processing.

Here, high quality of the visual representation and speed in generating the images on standard hardware are challenging demands. Nonlinear filtering techniques were applied to smooth the data obtained from computer tomography. The paper presents some features and implementation details of an object oriented software specialized in volumetric image processing. It is a report of the results in hybrid visualisation of volumes and different surface types, where the most important progress could be achieved. The volume rendering techniques of this project were developed for medical applications, however, they are in the meantime applied to high dimensional data sets, e.g. to visualise reactive flow and transport.

This BMBF supported project had in medical application the following partners:

Klinik für Mund-, Kiefer- und Gesichtschirurgie, Universität Heidelberg
(Priv. Doz. Dr. Dr. S. Haßfeld)
FRIATEC AG, Mannheim-Friedrichsfeld

1 Problem

In the last years a large number of higher dimensional datasets are being produced as results of physical experiments or by solving mathematical models.

The various kinds of tomographic devices (computer tomography (CT), magnetic resonance (MR), optical scanners (non invasive), laserspectroscopy) offers a large number of 3D scalar datasets for further processing. The medical imaging technologies generate data containing the representation of internal organs by density characteristic and allow diagnosis without invasive surgery.

The volumetric dataset can be viewed like a stack of 2D images and can be visualised in 2D, considering to be a 2D dynamic process, like an animated 2D sequence that can give a certain perception of the inner structure of the scalar field.

Usually the visualisation of such higher dimensional datasets using classical graphic library routines is done by using cutting planes or by making binary decisions in order to construct polygonal representations of certain
structures. For data with high coherence a divide and conquer approach to visualise existing level surfaces is the marching-cubes algorithm.

Often and especially in the case of physical experiments (like combustion measurements) the important regions or structures are not known a priori and a global view of the result using a more natural visualisation is necessary.

Computational results of reactive flows and fluid dynamics models can be used to predict behaviors of real processes and a visual analysis of this simulations can be done by implementing fast volumetric visualisation tools.

An important advantage of the volume rendering techniques is that the data does not need to be thresholded giving a global view of the scalar field, on a 2D image.

Problem arises in higher dimensional visualisation due to large amount of information to be processed in acceptable time combined with high computational and storage resources.

The rendering method has to be fast and to preserve a high quality of the resulting image.

## 2 Mathematical Models

The volumetric database represented by a 3D matrix of scalar values is considered to be an optical medium. Physically a light ray entering this semitransparent matter will interact with the material particles and transport inner informations (colour, light intensity) out from the volume. What we see from the volume, is the amount of coloured light that reaches the eyes after being reflected by the volume.

Given the volumetric datasets, we have to construct a mathematical model simulating a light transport through the optical medium.
By applying projections along the rays we decrease the domain dimension from a volume to a textured surface and make it accessible for 2D visualisation (Fig. 1).

Considering the case of linear transport of light and interaction of light with the participating medium via absorption, emission and scattering. Then the energy balance equation on the light ray direction is given in the form of the standard one-speed particle transport equation or Boltzmann equation: [2]

\[ n \cdot \nabla I(x, n) + \mathcal{A}(x, n) I(x, n) + \frac{1}{4\pi} I(x, n) \int p(n \rightarrow n')d\Omega' = \mathcal{E}(x, n) + \mathcal{J} p(n' \rightarrow n) I(x, n')d\Omega' \]  

(1)

with the following notation:

- \( n \) – ray direction
- \( I(x, n) \) – ray intensity at point \( x \) in the direction \( n \)
- \( \mathcal{A}(x, n) \) – absorption coefficient at point \( x \) in the direction \( n \)
- \( \mathcal{E}(x, n) \) – emission coefficient at point \( x \) in the direction \( n \)
- \( p(v_1 \rightarrow v_2) \) – probability of scattering from direction \( v_1 \) to direction \( v_2 \).

The normalized integral is taken over the solid angle.

The two integrals represent the outscattering and inscattering components of the model.

Depending on the illumination model we compute the factor determining the material color at a given point in the framework of the particle shading process.

The amount of incident light computed in one voxel in concordance with the local gradient using the Phong model is:

\[ I = I_a k_a C_a + f_{att} I_{ray}[k_d C_{diff} (\nabla F \cdot \overline{L}) + k_s C_s (\overline{R} \cdot \overline{V})^n] \]  

(2)

with the following notation:

- \( I_a \) – intensity of ambient light
- \( k_a C_a \) – color amount from ambient reflection
- \( I_{ray} \) – intensity of the light rays
- \( k_d C_{diff} \) – color amount from diffuse reflection
- \( \overline{L} \) – rays direction
- \( \overline{R} \) – reflection direction
- \( \overline{V} \) – viewer direction
- \( n \) – specular reflection exponent
- \( F \) – the discret function in \( \mathbb{R}^3 \) (the volume data)

Computationally it is better to use the halfway direction \( \overline{H} = \frac{\overline{L}+\overline{V}}{|\overline{L}+\overline{V}|} \) to compute specular reflection coefficients instead of \( \overline{R} \). The illumination equation is evaluated over a three component color model (RGB). The vector \( \overline{H} \)
is constant over the illuminated object and the reflection vector $\vec{R}$ is variable depending on the normals to the isosurfaces in the volume. The new specular reflection term can be expressed as $(\vec{N} \cdot \vec{H})^n$. [8]

The ray integration brings also transparency information of the integrated interval. The transparency information is used as blending parameter of the $RGBA$ boundary texture.

3 Filtering

The raw volumetric datasets arising from physical experiments are often noisy due to electronic devices sampling technology. In such cases low-pass filters and reconstructions filters need to be convoluted with the raw volumetric data in order to “clean up the data”. If the noise of the volume data is sufficiently random, then filtering, which computes averages of adjacent voxels, should average out the noise, or at least filter its high-frequency components.

These local area operators (filters) perform operations on the local area surrounding a voxel, implemented as a sliding box centred at the voxel and moved across the data set.

The segmentation part is directly related to the filtering of the raw data. Reconstruction filters with edge enhancement (surfaces in volume data), facilitate the data segmentation that affect directly the quality of the final projection. An efficient type of nonlinear filtering algorithms is represented by median filters with good edge enhancement and low-pass filtering properties. We use this type of filtering for the volumetric raw data.

We choose a 3x3x3 filter kernel to compute in each voxel the filtered value. The value is defined to be the middle value of the sorted list that contains the 27 kernel values.

These kind of filters can be very quickly convoluted with the volume data, but more accurate results can be obtained by using anisotropic nonlinear diffusion filters. They are more flexible and combine smoothing properties with edge enhancement qualities [3], [4]. The main problem for more complex filters remains the low speed of convolution with large scalar fields.

4 Data Representation

The most often used sampling grid for a volumetric database is a regular rectilinear grid. This topological distribution of the scalar information permits a fast access to the database but in the case of mathematical models adaptive numerical solver generates unstructured volumetric grids. In this case a conversion to regular grid is imposed.

We consider the situation that a finite number of point light sources send rays through the volume dataset. The most used technique in 3D sampling by light transport is the image ordered ray-tracing, in which the sampling rate is determined by the viewer plane. In this case the intensity of the light exiting
the volume will be computed in any point of the 2D raster that represent the
viewer plane. By ray integration we try to reconstruct the information from
the discret grid.

Using a backward projection the image ordered algorithms have to com-
pute the intersection points of the ray with the volume boundary before
starting the ray integration.

In order to get a better control over the data we will use a different
approach considering an object-ordered ray-casting to compute the intensity
of light rays at the volume boundary. The volume datasets are represented by
a function in $\mathbb{R}^3$ such that the space needed to represent such functions is at
least $\mathbb{R}^4$. Classical graphic libraries can manipulate functions in $\mathbb{R}^2$ (surfaces)
and using a light ray integration we create a texture of the volume’s boundary,
that contains material color and transparency informations.

An object-ordered algorithm ensures that every volume element (voxel)
will contribute to the final image without the need of adaptive supersampling.
In such a way we can take benefit of spatial coherences existing in the volume
data and also of ray-path template [5] that remains consistent for every ray
in orthogonal projection (Fig. 2).

The voxel represents an element (cube) of the volume partition delimited
by eight grid node that represent the voxel corners. Other implementations
use voxels centered on the grids nodes.

Spatial coherence can improve the early ray termination and accelerate
the integration along the current light ray by using run-length encoding of
the 3D scalar matrix.

In the case of a rectilinear regular grid the volume represents a cube,
so at most three faces will become accessible to the viewer at a given time.

Fig. 2.Intensity evaluation points Ray-path template
The sampling rays will create three textures that represent projections of the volume on the three visible faces in the viewing direction.

Using the object-ordered tracing we eliminate the step of computing intersection points of light rays with the volume. Aliasing artefacts arising from different sampling rates are strongly diminished.

The projective method used to bring interior information to the boundary of the volume has to be consistent with the projective method used by the graphic library to represent the 3D textured object on the viewer plane (display) (Fig. 3).

For a fast numerical solution of the light transfer equation we implement a simplified model of equation (1).

In order to simplify the transfer equation we consider the single scattering case that reduces the model to an emission–absorption model. The differential form of the equation becomes:

\[ n \cdot \nabla I(x, n) = -\tau(x)I(x, n) + \xi(x, n). \]

By defining the optical depth between two points we get

\[
\begin{align*}
x_1 &= p + s_1 n \\
x_2 &= p + s_2 n \\
\tau(x_1, x_2) &= \int_{s_1}^{s_2} \tau(p + s'n) ds'.
\end{align*}
\]

Equation (1) can be reduced to:

\[
\frac{\partial I(x)}{\partial s} = -\tau(x)I(x) + \xi(x)
\]

and solved using the integrating factor \(e^{\tau(x_0,x)}\)

\[
\frac{\partial}{\partial s} (I(x)e^{\tau(x_0,x)}) = \xi(x)e^{\tau(x_0,x)}.
\]

Fig. 3. Viewing direction  Orthogonal to Z axis  Orthogonal on the XZ plane
We get a solution of the transfer equation:

\[
I(x) = I(x_0)e^{-\tau(x_0,x)} + \int_{s_0}^{s} \xi(x')e^{-\tau(x',x)}dx'.
\] (3)

At this stage we have the intensity of light at point x along the ray, given by the attenuated incoming intensity, summed with the contributions of every emittent particle along the ray so far, with the initial point \(x_0\).

The revealing of inner structures is improved by controlling the distribution of the light intensity and color with a defined light model. We take also in consideration the possibility of visualising volumetric datasets with embedding opaque surfaces generated by a classical graphic library (GL) and enhancement of the inner surfaces structures.

Textured cube in: (see Fig. 3)

In the case of unstructured 3D grid we use an octree-based space partitioning to order the unstructured cells [7] in concordance with octree nodes intersection. Each octree node has a list of cells that intersect the node (Fig. 4).

5 Rendering Pipeline

A volume visualisation process includes four steps:

**Data compression** – by eliminating transparent voxels and by taking benefit of spatial coherence with a run-length encoding approach.

**Segmentation** – a material index is associated to every voxel. A low pass filtering with edge enhancement has to be applied to the raw data.
Classification – that associates an opacity (transfer function) to every segment.

Rendering – that converts the raw scalar data into a final projection.

The manipulation of large amount of information is requesting a good problem decomposition, in order to get a full control over informational flux.

To visualise volumetric datasets it is necessary to provide a module controlling the ray propagation through this participating medium. The Ray-Tracer represents the last node (module) in the informational graph and achieves the graphical representation of the input scalar field. The output of this module represents the texture of the volume’s boundary and will be passed to the interface modules (viewers, editors, manipulators).

The information within the preprocessing modules is independent under rotation and has to be updated only in the case of inner parameters or upper modules modifications in the informational flux graph.

Fig. 5. Modules
The informational flux diagram (Fig. 5) presents the principal modules that operate on data, in order to compress the main input information in the format needed by the ray integrator (tracer module).

By a quantification of parameters the only module that depends on the raw data volume is the 3D-texture volume, the other modules need constant internal memory allocation which is direct proportional with the discret refinement.

The answer speed of any module is improved by this large autonomy of modules, the tracer module now is able to process a volume of $256^3$ in a few seconds (5–20 s on R10000 SGI) depending on global transparency and without taking benefit of spatial coherence. Further acceleration modules will be implemented for static volume data later in this paper.

6 Implementation

Each module is represented by at least two C++ classes: the internal class that implements the modules functionalities and the corresponding user interface class.

The Open Inventor graphics library contains objects for graphics applications and orders the 3D scene in a hierarchical structure forming a directed acyclic graph. The volumetric object has to be implemented in such a way that the scene manager recognizes and successfully deals with the new objects inserted in the scene graph.

For the implementation of the application interface we use the Open Inventor's component concept and for the eventual customisation the OSF/Motif application programming interface and the widget concept [1].

The main interface's SoComponent is the scene viewer that has control over the scene graph and enables the communication between other interface components and corresponding internal objects.

Interface components for tools classes are implemented like an inherited SoXtVolLightEditor (directional light editor), SoXtVolMaterialEditor (material properties editor), SoXtVolClipEditor (orthogonal clipping planes editor) (Fig. 6).

The scene viewer controls also the different states of the rendering algorithm like low resolution rendering, bounding box rendering or anaglyph rendering (stereo red-green) (Fig. 7).

The tracer module will be represented by a SoShape class named SoVolume in order to have access to the graphic context via SoState structure. The rendering algorithm of the tracer is implemented in the inherited GL-Render method.

We take into consideration only the path from the particle to the viewer, that implies visualisation of volumetric data without shadowing effects.

The boundary of the volume has to be partitioned in surfaces primitives, in this case squares and the tracer has to loop over visible faces and discrete points on each face:
for (every visible face)
  for (every discret point on face)
      IntegrateRay(ray-direction, color, transparency);

In the preprocessing steps the 3D-Texture module computes the quantified local gradient for each voxel and the ray path template structure that contains the information needed to perform the numeric integration on the light ray path.

```
_template struct
  signed char x, y, z;    // offset of next voxel
  signed char offset1, offset2, offset3, offset4; // offsets of the interpolated voxels
  float w1, w2, w3, w4;  // Interpolation weights (bilinear)
  float iCoeff;          // Integration coefficient
```

In order to reduce the number of parameters we choose only the absorption part of equation (3) and using the rectangle rule for integration we have

\[ I_{out} = I_{in} e^{-R_i \Delta x} \]

Iteratively we have the intensity of light ray

\[ I_{out} = I_{in} e^{-R_i \Delta x} \]
Fig. 7. Scene viewer in stereo mode (red-green)

\[ I_{i+1} = I_i(1 - \alpha_i) \]

and the final color

\[ C_{f_{i+1}} = C_{f_i} + (I_{i+1} - I_i)C_i = C_{f_i} + I_i \alpha_i C_i \]

and we get the final expression of the color at step \( i+1 \),

\[
C_{f_{i+1}} = \alpha_0 C_0 + (1 - \alpha_0) \alpha_1 C_1 + (1 - \alpha_0)(1 - \alpha_1) \alpha_2 C_2 + ... \\
= \sum_{k=0}^{i} c_k \prod_{j=0}^{k-1} (1 - \alpha_j) \\
= c_0 \over\over c_1 \over\over ... \over\over c_i ,
\]

also called the volumetric compositing equation, with

\[ \alpha_i = 1 - e^{-\kappa_i \Delta x}, \quad c_i = \alpha_i C_i, \]
\[ I_0 = 1, \quad C_{f_0} = 0. \]
The quantified gradient over a solid angle is coded on 14 bits (6 bits φ angle 6 bits θ angle and on 2 bits the sign of φ and θ) and represents an index to the gradient-based shaded LUT, the second index in this table will be represented by the material index computed by the segmentation operator. The gradient index is created by codification of the polar coordinates φ and θ on 14 bits. Using a normal-based light model (Phong), the gradient-based module computes for each voxel a shading coefficient in order to get a better perception of the 3D structure of existing isosurfaces.

Equation (2) depends, over the given volume, only on the local gradient and diffuse color. We consider the ambient and specular color to be constant white.

If we define

\[ coef_{Phong1} = f_{att}k_d(\nabla F \cdot \bar{L}) \]
\[ coef_{Phong2} = f_{att}k_sc_s(\nabla F \cdot \bar{H}) + k_aC_a, \]

then the color intensity becomes

\[ I_a = I_{ray} \]
\[ CI = I_{ray}(coef_{Phong1} * C_{diff} + coef_{Phong2}). \]

Both coefficients will be computed for every discretised gradient. Then the final color intensity CI depends on the number of diffuse colors (given by the number of segments) and the number of gradients-index

\[ CI(segments, gradients) \]

will be computed in a preprocessing step in a table with dimension number-of-segments x number-of-gradients.

To make possible the use of the property that the gradient in point \((x_0, y_0, z_0)\), is normal to the level surface passing through this point, we need...
to convolve the raw data with a low pass reconstruction filter in order to improve spatial coherence.

Along the rays the sampling points are chosen to be the intersection points of the light ray and voxel faces, the corresponding value being computed by bilinear interpolation. In the case of gray-level segmentation by multithresholding of the values domain $D \subset \mathbb{R}^+$ we have:

$$\text{MaterialIndex}(x, y, z) = m_i \quad \text{for} \quad F(x, y, z) \in D_i$$

$$\cup D_i = D, \quad D_i \cap D_j = \phi \quad \text{for} \quad i \neq j.$$

The material generator module can perform an interpolation of parameter values inside the threshold domain $D_i$. We assume that data coherence insures that neighborhood segments contain spatially-neighbour voxels.

Two kinds of surfaces can be visualised at the same time with the volumetric projection. One type is represented by surfaces that arise by affecting the optical behavior of the voxel implied in the surface enhancement. This surface type it is visualised by the tracer. Another type is a polygonal surface created by the graphic library, for which collisions with the light ray have to be considered.

For the first case we enhance the topologic boundary of a segmented object from the volume considering a segment to be homogeneous and modifying the absorption coefficient in such a way to preserve a boundary surface of given thickness. The absorption coefficient will be modified with the following formula that can be found in [6]:

$$\alpha = \alpha_{\text{segm}} \begin{cases} 0 & : \quad |\nabla f(x)| = 0 \quad f(x) \in D_{\text{segm}} \\ 1 - \frac{1}{r} \frac{|f_{\text{segm}} - f(x)|}{|\nabla f(x)|} & : \quad |\nabla f(x)| > 0 \quad \text{and} \quad \frac{|f_{\text{segm}} - f(x)|}{|\nabla f(x)|} < r \end{cases}$$

otherwise the absorption coefficient is not modified.

where:

- $r$ — surface thickness
- $f_{\text{segm}}$ — segment parameter value (homogeneous segment)

Fig. 9. Segmented domain $\Omega$ Boundary of thickness $r$ ($\partial \Omega$)
The difference is that we enhance the contact surfaces between segments which are not necessary isosurfaces.

In order to handle classical GL surface visualisation we use the depth-buffer maintained by the graphic library (or hardware for graphic stations) to get access to the current depth information.

On the ray path the depth value (distance from the viewer) of every sample point is compared with the corresponding value in the depth-buffer (depth test). If the last sampling point has a larger depth value, the last integration point will be at the distance given by the depth-buffer.

In orthogonal projection the pixel coordinates affected by a point on the volume boundary need to be computed only one time per boundary point, the displacement on the ray-path does not affect the raster position of the current point, only the depth value changes.

The problem of accurate collision detection arises in the case of perspective projection, due to the nonlinear perspective division implemented for the depth-buffer in GL.

The algorithm does not make a voxelization of the surface-based 3D object and thereby preserves the original rendering resolution of these objects.

In the scene graph, the volumetric object has to be the last graphic object, to ensure that the depth informations in the depth-buffer is correct. In this way any collision will be correctly handled by the tracer.

Another feature of the application is the multichannels rendering by creating a new volume from a serie of volumes in the following way:

$$\rho_{ijk}^{new} = \max\{\rho_{ijk}^c : c \in \{1, ..., nrChannels\}\},$$

with the corresponding colour
\[ C_{ij,k}^{new} = \sum_{c=1}^{nr Channels} \left( \frac{\alpha_{ij,k}^c}{\alpha_{ij,k}^{\text{max}}} \right), \]

where \( \alpha_{ij,k}^c \) is the absorption coefficient (direct proportional to the density \( \rho_{ij,k}^c \)) at position \((i,j,k)\) in the channel \(c\).

7 Acceleration Techniques

The presented method of casting rays into the volume and integrating along the rays remains too slow for large datasets and further acceleration of the integration module needs to be implemented.

One reason why the integration part is too slow is due to missing the cache levels because of random access to integration voxels. By accessing the integration voxel in the order given by the ray direction leads especially on large datasets to the missing of the cache levels and slows down significantly the integration time.

A large amount of computational time is wasted by accessing and integrating the empty voxels that represent in most of the cases a large percent of the volume data and does not affect the final value of the ray integration.

We can improve the integration time by a better use of the cache functionality and by taking benefits of data coherence of the empty voxels inside the volume. To take full benefits of the cache functionality we have to improve the access to the volume elements by accessing them in such an order that sequentially accessed voxels are spatially neighborhoods in memory storage (Fig. 11).

![Fig. 11. Scan lines](image-url)
By using object ordered ray casting and the parallel projection we can change the ray integration order into a scanline ordered integration. The textures on the boundary of the volume are created from parallel scanlines. Every texture scanline is created by blending (integrating) parallel scanlines from the volume that represents the intersection between the plane formed by the texture scanline and the view direction and the faces of the voxels.

We will take benefit of spatial coherence in the volume data by using the linear coherence per voxel scanline. In this way we need only 2 bytes to save the length of compact empty voxel lines, one byte for voxel scanlines parallel with Ox direction and one byte for voxel scanlines parallel with Oy direction. The amount of memory used to encode this type of scanline run-length is much lower than the amount of memory used to encode an octree based run-length.

We call a run of length $l$ a sequence of length $l$ of transparent voxels.

We need at least two scan directions because at most four faces can be constructed from scanlines parallel with a given direction, the remaining two faces need two be constructed from scanlines parallel with one of the remaining two directions.

Only one byte per run is necessary because the run with length greater than 255 is split into several runs with length between 0 and 255. Every run is also split at the end of each voxel scanline, insuring that every voxel scan line begin with a new run. We need the run length only on the positiv axis direction, this order is maintained also for scan lines of the final texture.

For a given direction each voxel has a value that represent the remaining length of the current run (Fig. 12). This value together with the scan direction gives a complete definition of the current run and the access at any position inside the run is consistent.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{run_length.png}
\caption{Run length}
\end{figure}
By using this method of run length encoding, we have a speedup of order 5–10x for a volume with 70–80% transparent voxels, such a volume of $256^3$ is visualised in 1–3sec compared with 12–15sec of the ray casting method that use only early ray termination technique for the same volume. The runs which are independent under geometric transformations of the volume, have to be recomputed only when the classification function changes.

8 Future Work

Further optimizations can be implemented in order to accelerate the time of applying operators by using, if possible, the separation property of linear operators.

Segmentation by anisotropic diffusion filtering will be implemented in order to improve the quality of edges and to automatize the structure detection module. A multiprocessor version of the tracer has been tested on a four R10000 processors board with a speed-up of order three.

References

VIII. Statistical Methods in Medical Applications

The Application of Statistical Methods of Meta-Analysis for Heterogeneity Modelling in Medicine and Pharmacy, Psychology, Quality Control and Assurance


An Application for the Analysis of Human Tremor Time-Series

Video Coding
with Adaptive Vector Quantization
and Rate Distortion Optimization

D. Saupe¹ and M. Wagner²

¹ Institut für Informatik, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany
² Institut für Informatik, Universität Freiburg, Am Flughafen 17, 79110 Freiburg, Germany

Abstract. The goal of the research presented in this paper is the development and evaluation of adaptive image sequence coding. The method, based on adaptive vector quantization, has been combined with several video coding techniques like wavelet transform, quad-trees, and rate distortion optimization. In addition we provide a comparison with a state-of-the-art video codec (H.263) and describe experiments with motion compensation.

1 Introduction

In the project “Adaptive Verfahren zur Bild- und Bildsequenzkodierung¹” of the BMBF research initiative “Mathematische Verfahren zur Lösung von Problemstellungen in Industrie und Wirtschaft”, we investigated and improved encoding approaches for images and image sequences.

For the encoding of images, we improved methods of near-lossless coding. In this approach, the distortion of the original and reconstructed image is measured with the $L_\infty$ norm. Thus, the maximal error of each image component is restricted. Results of our research have been published in [1–3].

For the encoding of image sequences, we evaluated the prospects of a new adaptive image sequence coding scheme for very low bitrates. During the project, we developed and successively improved our coding strategy. A first impression is given by Fig. 1. The results have been published in [4–8].

Further work carried out in this project is the investigation and testing of hierarchical table lookup vector quantization for rapid video encoding and decoding, and the simulation of a hardware implementation of industry standard image sequence coding methods (H.261, H.263 and parts of MPEG 2).

In this paper, we can only give an introduction to our adaptive image sequence coding approach. The codec is based on hierarchical adaptive vector quantization (AVQ) in the wavelet domain and, unlike standard video coding, does not apply motion compensation. The hierarchical organization

¹ This project is supported by our industrial partner Micronas Intermetall GmbH, Freiburg.
Vector quantization (VQ) has been proven a powerful compression scheme for coding of images and image sequences [13,14]. However, in most schemes a static codebook is used. The vector quantizer applies only one fixed codebook neglecting that image sequences typically are not stationary and that different

\begin{figure}
\centering
\begin{tabular}{ccc}
\includegraphics[width=0.3\textwidth]{frame_005.png} & \includegraphics[width=0.3\textwidth]{frame_020.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} \\
\includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} \\
\includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} \\
\includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} & \includegraphics[width=0.3\textwidth]{frame_050.png} \\
\end{tabular}
\caption{Line by line (top down): The 5th, 20th, 50th, and 100th frame of the Salesman sequence. Column by column (left to right): The original sequence, the decoded sequence of our current codec, and the decoded sequence of the real time version at a bitrate of 16 kbit/s.}
\end{figure}

of the wavelet coefficients is made by a special tree data structure called quad-trees[9]. To achieve optimized results, we adopted and extended techniques from rate distortion theory, the mathematical foundation of lossy data compression[10–12].
sequences have different statistical behavior. Thus, adaptive vector quantiza-
tion (AVQ) for image sequences was proposed [15,8]. In order to improve the
AVQ-codec, a rate-distortion (RD) optimization was investigated in [7,16,17].
It was shown in [4–6] that vector quantization in the wavelet domain and
hierarchical organization of the wavelet coefficients leads to additional coding

2 Outline of the Codec

In a preprocessing step a two times octave-band wavelet transform (9/7
biorthogonal filter [18]) is applied to each frame. The coefficients of the
transformed frames are regrouped into macro blocks each corresponding to
a 16 × 16-pixel area in the spatial domain. The organization of the wavelet
coefficients of a macro block is shown in Fig. 2. Each macro block consists of
256 coefficients.

For every macro block there are three possible decomposition levels from
coarse to fine. Level-0 describes the whole macro block with all 256 wavelet
coefficients. Alternatively, the level-0 block can be decomposed into four
level-1 blocks each one corresponding to a spatial 8 × 8-pixel area containing
64 coefficients. Moreover, one can decompose each level-1 block into four level-
2 blocks each one corresponding to 4 × 4-pixel blocks in the spatial domain.
The grouping of the wavelet coefficients for level-1 and level-2 blocks within
a macro block is depicted in Fig. 3. The decomposition of a macro block can
be described by a quad-tree.

In order to create vectors from blocks, the blocks are scanned line by
line, excluding coefficients from several subbands depending on the block
level. Thus, there are three kinds of vectors, called level-0, level-1 and level-2
vectors (for details see [4,5]).

Fig. 2. Regrouping of macro blocks in the wavelet domain. (a) Subband grouping. Coefficients are grouped corresponding to their type of subband membership. (b) Macro block grouping. Coefficients are grouped corresponding to their spatial position
For level-0 vectors there is only one encoding mode. Replenishment is applied, i.e., the content of the same position in the previously decoded transformed frame is restored. For level-1 and level-2 vectors the AVQ-approach is applied. There are three encoding modes:

**Mode 0.** Replenishment mode. This works like the replenishment mode for level-0 blocks, i.e., the content of the same position in the previously decoded frame is restored.

**Mode 1.** VQ-mode. The vector is quantized and encoded by VQ.

**Mode 2.** Update mode. The vector is scalar quantized and encoded.

For level-1 vectors and level-2 vectors different vector quantizers are used since level-1 vectors are 12-dimensional and level-2 vectors are 15-dimensional. In order to maintain an efficient variable length encoding, the codebooks $C_1 = \{x_0^1, \ldots, x_{n_1-1}^1\}$ and $C_2 = \{x_0^2, \ldots, x_{n_2-1}^2\}$ of the two vector quantizers are organized as follows. A frequency count is assigned to every vector. At the beginning of the encoding of a frame, the vectors are sorted with respect to the frequency count. The most frequent vector can be found at the beginning of the codebook, the least frequent vector at the end. The frequency count is incremented by 1 every time a vector is used by mode-1 encoding. When the encoding of a frame is finished, all vectors are sorted again. After that, the vectors encoded in mode 2 have to be inserted in the codebook by a heuristic rule [7]. The new vectors are assigned a frequency count of $f(\left\lfloor \frac{n}{2} \right\rfloor) + 1$ with codebook size $n$ and $f(i)$ describing the frequency count of the $i$th vector in the codebook; i.e. the frequency count of the vector in the middle of the codebook is taken and incremented by 1. After all mode-2-vectors have been inserted, the vectors with the least frequency count are removed in order to maintain a constant codebook size.
For variable length coding, an adaptive arithmetic coder is used. The decision how to decompose the quad-trees and how to choose coding modes is done with an algorithm based on the generalized BFOS algorithm [12]. Thus, the modes and the quad-tree decompositions are selected in a rate-distortion optimal fashion. For the encoding of the chrominance components, about $\frac{1}{10}$th of the given bitrate is used.

3 Rate-Distortion Optimization

In the last section we discussed the encoding options that could be used in order to encode a frame. We didn’t consider so far the issue of how to select these options and which quad-tree decomposition should be used. Apparently, this is a constrained optimization problem. Let a frame $X$ consisting of $M$ macro blocks $(X_0, \ldots, X_{M-1})$ and a previously decoded frame (reference frame) $X^r = (X^r_0, \ldots, X^r_{M-1})$ be given. Then the objective is to construct an approximating frame $\hat{X} = (\hat{X}_0, \ldots, \hat{X}_{M-1})$ by means of encoding options described in Sect. 2 with the smallest possible overall distortion

$$D = d(X, \hat{X}) = \sum_{m=0}^{M-1} d(X_m, \hat{X}_m)$$

subject to $R = r(\hat{X}) = \sum_{m=0}^{M-1} r(\hat{X}_m) \leq R_t$. (1)

Here $d(x, y) = \| x - y \|_2^2$, $r(\hat{X}_m)$ describes the bits needed to encode macro block $\hat{X}_m$, and $R_t$ the target rate.

Having the set of all possible encoding options $\mathcal{I}$ for a macro block, including all encoding modes for all quad-tree decompositions, we can assign to each $i \in \mathcal{I}$ and each macro block $X_m$ an encoding rate and the corresponding distortion $(R^i_m, D^i_m)$. Thus, we can define for each macro block a set of rate distortion points $\mathcal{P}_m = \{(R^i_m, D^i_m) : i \in \mathcal{I}\}$.

The constrained minimization (1) can be transformed into an unconstrained minimization with the well known Lagrangian multiplier approach:

$$\text{minimize } \sum_{m=0}^{M-1} D^i_m + \lambda \cdot \sum_{m=0}^{M-1} R^i_m, \ i_m \in \mathcal{I}$$ (2)

The problem of (2) is that we have not only to determine the optimal encoding options but also a value of $\lambda$ that yields a solution meeting the rate constraint. Thus, we have to iterate the minimization several times to find an optimal $\lambda$. With a fixed value of $\lambda$, the minimization can be performed for each macro block independently. The solution index of (2) for macro block $X_m$ is denoted by $i^*_m$. 
In order to compute the solution more efficiently we make use of the geometric interpretation of the Lagrangian multiplier approach. For this we consider for each macro block $X_m$ the set of RD-points $\mathcal{P}_m$. The solution $(R^{\ast}_m, D^{\ast}_m)$ of (2) for macro block $X_m$ can be constructed by the following procedure. A line with slope $\lambda$ is drawn through the origin. Afterwards the line is shifted towards the convex hull of $\mathcal{P}_m$. The first point met is the point $(R^{\ast}_m, D^{\ast}_m)$. If we consider the lower convex hulls of the sets $\mathcal{P}_m$ as “rate distortion curves” then the solution of (2) can be interpreted as a “constant slope” condition (Fig. 4). This suggests to compute the solution by direct exploration of the lower convex hulls of all sets $\mathcal{P}_m$. One example of this kind of procedure is the algorithm by Westerink et al. [19]. Starting from the minimal possible rate for each $\mathcal{P}_m$, the lower convex hulls are computed for increasing rates until the target rate is reached. This works as follows: First, the smallest achievable bitrate is searched, i.e. for every macro block $X_m$ the encoding option $i_m \in \mathcal{I}$ with the smallest rate $R^{\ast}_m$ is computed. For every macro block $X_m$ the coding option $j_m$ describing the next point (with a higher rate) on the lower convex hull (LCH) is determined. Let $\lambda_m$ be the slope between the RD-points specified by $i_m$ and $j_m$. The macro block $n$ with the “steepest”, i.e. minimal, $\lambda_n$ is selected. The coding option for $X_n$ is changed $i_n \leftarrow j_n$. Then the next point on the LCH of $\mathcal{P}_m$ is searched and the value of $j_n$ is updated accordingly. After that, again the “steepest” $\lambda_m$ is searched and so forth.

The BFOS algorithm [12] generalizes this approach in order to make it applicable for tree structured coding. However, for this approach a monotonicity condition is assumed, i.e. it must be assured that a larger tree leads to a larger encoding cost and to a smaller distortion. In addition it is assumed that there is only one encoding option for a node. Both assumptions are not met in our hierarchically organized AVQ-codec. Thus, we developed a more general approach without these restrictions.

Such optimization techniques can be summarized as incremental computation of the convex hull [5].

![Fig. 4. Geometrical interpretation of the Lagrangian multiplier](image)
4 Results

In this section, we provide some coding results of our current codec and compare it with older versions to show the successive improvement during the project. In addition we present comparisons to the state of the art video coding standard for very low bitrates (H.263). The experiments are made with QCIF frames (176×144) and codebook sizes of $n_1 = 64$ and $n_2 = 512$. The codec uses a frame rate of 8.3 frames/s. As usual, the peak signal to noise ratios (PSNR) are computed using only the luminance component.

Figure 5 shows the mean PSNR for several bitrates of our AVQ-codec and the previous versions without quad-trees. It can be seen that the application of quad-trees leads to a coding gain of up to 0.5 dB PSNR compared to the best of the previous codecs.

Figure 5 also shows the development of our codec during the project. The codec with the least performance is purely based on AVQ in the spatial domain and is able to encode in real-time using current PC technology. The third best codec applies RD-optimization to the latter one. The second best one additionally uses a wavelet transform and applies the AVQ scheme in the transform domain. Together with the improvement by using quad-trees the performance was improved up to 3.5 dB PSNR during the development.

Unfortunately only a small number of results have been reported in the literature for VQ coding of video sequences that are suitable for a comparison, none of which can provide a fair comparison with other VQ-approaches. To give the reader something at hand to assess the performance of our codec, we provide a comparison with the tmn codec (H.263). The tmn codec (version 3.0) is used with syntax based arithmetic coding at a bitrate of 8000 bits/s and a frame rate of 8.3 frames/s. The target bitrate is achieved by the off-line rate control method of the tmn codec, i.e., the codec tries to adjust an average bitrate of $8000 \over 8.3$ bits/frame over the whole sequence. The result is depicted in Fig. 6. Figure 6a shows the Mother & Daughter sequence. One can see that at the beginning the performance gap between these two encoding schemes is
Fig. 6. PSNR course for (a) Mother & Daughter and (b) Salesman for the tmn and the AVQ-codec at 8000 bits/s

very large. The gap is decreasing during the encoding of the first 100 frames. After that an average difference of about 1.2 dB PSNR between the tmn and the quad-tree codec is maintained over the rest of the sequence. After subsequent 100 frames the average performance gap is about 0.9 dB PSNR. A similar observation can be made for Salesman in Fig. 6b. This shows the adaptability of our approach that needs about 100 frames to evolve. Note that we are not using motion compensation.

5 Conclusion and Future Work

As a part of this project we have developed a new image sequence encoding scheme for very low bitrates. This scheme is based on AVQ and quad-tree coding in the wavelet domain. The codec uses an efficient RD-optimization mechanism.

The performance of the codec was successively improved during this project. The comparison with standard transform coding still shows a perfor-
mance gap of about 1 dB for some test sequences. We conclude that motion compensation is essential also for codecs based on AVQ. In a given image sequence the encoding history implicitly contained in the adaptive VQ codebook is not sufficient to replace traditional motion compensation techniques. Therefore, in order to further improve AVQ video coding we combine it with motion compensation. In fact, experiments combining the AVQ scheme with motion compensation show promising results. First results, given in Fig. 7, confirm our conclusion, and, thus, for the remaining part of the research project, we will further investigate this issue in more detail.

**Acknowledgments.** During the project from August 1997 to present, many people, apart from the authors, made substantial contributions. The authors would like to thank the senior researchers Raouf Hamzaoui, Hannes Hartenstein, and Huiming Ding. Furthermore, the authors are very grateful for the help of the student researchers Erik Behrends, Bernd Butz, Ralf Herz, Matthias Ruhl, and Martin Schmidt.

**References**


The Application of Statistical Methods of Meta-Analysis for Heterogeneity Modelling in Medicine and Pharmacy, Psychology, Quality Control and Assurance

Dankmar Böhning¹, Uwe Malzahn¹, Peter Schlattmannn¹, Uwe-Peter Dammann², Wolfgang Mehnert³, Heinz Holling⁴, and Ralf Schulze⁴

¹ Department of Epidemiology, Free University Berlin, Haus 562, Fabeckstr. 60–62, 14195 Berlin, Germany, E-mail: boehning@zedat.fu-berlin.de
² Asta-Medica, Werk Künsebeck, Kantstr. 2, 33790 Halle-Künsebeck, Germany, E-mail: Uwe-Peter.Dammann@astamedica.de
³ Department of Pharmaceutical Technology, Free University Berlin, Kelchstr. 31, 12169 Berlin, Germany, E-mail: mehnert@zedat.fu-berlin.de
⁴ Occupational Psychology, University of Münster, Fliednerstr. 21, 48149 Münster, Germany, E-mail: holling@psy.uni-muenster.de

Abstract. In the past few years meta-analysis has become increasingly popular in many areas of science such as medicine and pharmacy, psychology and other social sciences. In these areas of application meta-analyses have been performed in order to obtain a pooled estimate of various single studies. Obtaining a single summary measure implicitly assumes homogeneity of these studies, i.e., the results of individual studies differ only by chance. In this case a combined estimate of the individual studies provides a powerful and important result. However this pooled estimate may be seriously misleading if study conditions are heterogeneous.

Thus, increasingly an approach has been advocated which considers meta-analysis as a study over studies. This approach seeks to investigate heterogeneity between studies. An important feature of this type of meta-analysis lies in the fact that it tries to identify factors which cause heterogeneity.

It has been the focus on this project (in corporation with the unit of quality assurance of ASTA Medica at location Künsebeck) to extend this approach appropriately to the area of quality control, where batches of the produced goods replace the role of studies in medicine or the social sciences. Clearly, in this setting an investigation of heterogeneity is equally attractive, since identification and modelling of heterogeneity helps to improve the production process.

1 Introduction

The paper reviews an approach which enables a global perspective on aspects of homogeneity and heterogeneity which occurs when applying methods of meta-analysis to clinical studies in medicine and pharmacy, psychology and other social sciences, but also in quality control and quality assurance in the pharmaceutical industry. In conventional meta-analysis investigations are done in such a way that a specific measure can be computed utilizing
numerous single studies. Frequently, statistical questions of efficiency are dominating in the literature (Hedges and Olkin, [1985]), which is achieved by pooling the various single studies and, thus by, achieving an increased sample size. This procedure, no doubt, is of great benefit, if the various studies to be combined in the meta-analysis, have emerged under comparable conditions and are different in a statistical sense only by chance. This is the situation of homogeneity. However, pooled analysis is often considered problematic if study conditions are heterogenous, especially if the interpretation of pooled estimators are kept in a traditional way.

1.1 Meta-Analysis of Clinical Studies in Medicine and Pharmacy

In clinical trials often a treatment group is compared with a control group, and the risk of some event (like survival after treatment) is compared between both groups. Let \( p_1 \) be the risk (probability) in the treatment group and \( p_0 \) the risk in the control group, then typical measures considered are the relative risk \( RR = p_1/p_0 \) or the risk difference \( RD = p_1 - p_0 \). These measures are estimated in several, say \( n \), studies and then pooled in a summary measure, for example in the case of the risk difference we have \( \sum_{i=1}^{n} w_i RD_i \) where the weights \( w_i \) are proportional to the inverse variance \( Var(RD_i), i = 1, ..., n \). There are numerous examples of this kind of meta-analysis and a recent reference to introductory reviews is Normand ([1999]) or earlier Jones ([1995]).

1.2 Meta-Analysis of Experimental Studies in Psychology

In the social sciences, primarily in psychology, an effect measure is computed for experimental or quasi-experimental studies which is often the standardized difference (difference of the means in treatment and control, then devided by the common standard deviation) or the correlation coefficient. A detailed discussion on the standardized difference is provided in Hedges and Olkin ([1985]), Cooper and Hedges ([1994]), and, in terms of the distributional aspects involved, in Malzahn, Böhning and Holling ([2000]). Details on the measure of the correlation coefficient are found in Hedges and Olkin ([1985]) and Cooper and Hedges ([1994]). Typically, the correlation coefficient \( \rho \) is used in it’s Fisher-transformed version having an approximate variance of \( Var\{log(z_i)\} = \frac{1}{m_i-3} \), where \( z_i = 0.5 \frac{1+\hat{\rho}_i}{1-\hat{\rho}_i} \) is the Fisher-transformation and \( m_i \) is the sample size of study \( i, i = 1, ..., n \). Consequently, the summary Fisher-transformed correlation coefficient takes on the simple form \( \frac{\sum_{i=1}^{n} m_i z_i}{\sum_{i=1}^{n} m_i} \) which is popular for it’s simplicity.

2 Meta-Analysis in Quality Control

The project and consequently the paper at hand investigates parallel aspects of meta-analysis and quality control. The cornerstone of this analogy are the
numerous batches which are drawn in quality control for monitoring purposes, which play the role of the single studies in meta-analysis. Measures of interest are here frequently count variables (counts of contamination particles) or other quality indexes. In this situation – even if homogeneity conditions are present – deviations from a given standard might occur, as well. It is quite important whether these deviations might have emerged from a homogenous process (as random variations) or are due to certain heterogeneities present in the production process. By means of the mixture distribution analysis we are able to model potentially present heterogeneity and further on, to classify each batch into one of the heterogeneity components. This might allow to diagnose certain common attributes and therefore be able to explore for the causes of heterogeneity.

2.1 Legal Background for Pharmaceutical Production

Pharmaceutical production of drug products and drug substances is worldwide regulated by the rules of Good Manufacturing Practices. For Europe and Germany producers have to follow the regulations of

1. Arzneimittelgesetz (AMG)
3. "Betriebsverordnung für pharmazeutische Unternehmer" (PharmBetrV 1994)

Production and quality control of drug products and drug substances have to recognize state of the art and current worldwide practices in accordance with the application. All procedures used in production and quality control must be validated and regularly revalidated. Drug products are mainly produced in batches, which should conform with the specification from batch to batch. Drug products brought into the market should be produced and controlled according to the application and the quality has to confirmed before a batch can be released for distribution.

The quality of a drug product or a drug substance is defined by identity, assay, chemical, physical and biological properties. A batch is the quantity of a drug produced under suitable uniform conditions to guarantee a homogeneous quality.

2.2 The Tasks and Objectives of Quality Assurance in Pharmaceutical Industry

The production of drugs is accompanied by

1. batch-and product related in-process controls (on line)
2. batch- and product related controls (off line)
3. not batch and not product related controls
Parenteral drugs are products which have to comply with additional, specific properties like sterility and essentially free of visible particles because of their parenteral application. Sterility is controlled by a sterility test which is destructive test on limited samples of a batch. In connection with in-process controls for the clean environment of rooms, air, surface and personnel hygiene during production especially parenteral drugs produced by aseptic processing sterility can be assured in all parts of a batch.

Each parenteral container is controlled by a 100%-inspection for particulate matter. The quality of this inspection is controlled by samples which are again inspected for subvisual particles. These are destructive tests on a limited number of samples. The quality is evaluated on the basis of quality index like the one which can be found in the Deutscher Arzneimittel Codex (DAC), Codex Probe no. 5. The particulate matter is evaluated for particles which can be seen easily, good or difficult. For instance:

1. No visible particle: no point
2. Particle difficult to be seen within 5 seconds: one point
3. Particle easily to be seen within 5 seconds: two points
4. Particles to be seen immediatly and in higher numbers: ten points

The formula for evaluation is: \[ Q_{TR} = \frac{A}{N}, \] where \( A \) stands for number of points recorded by three test persons and \( N \) stands for the number of controlled containers.

The results of all controls for one batch and from batch to batch is very important for the evaluation of the quality and the release for distribution. Trends for a homogeneous or heterogeneous process should be addressed and recognized as soon as it happens. Statistical evaluation of all available data is essential for the routine evaluation of the drug quality.

2.3 Meta-Analytic Modelling of Data Occurring in Quality Assurance

Very often quality assurance is based on the availability of a number of batches each having a certain number of items. For example, we might consider again \( Q_{TR} \) and define \( X \) as

\[ X = \text{Number of Times with } Q_{TR} \text{ positive in a series of } n \text{ investigations}. \]

This is best demonstrated by means of an example which is taken from the book of Derman and Ross ([1997]). The data are provided in Table 1.

As has been pointed out in the literature (Pettiti [1994], Cooper and Hedges [1994]), the area of meta-analysis has received various impulses during its historic development. In psychology the development of measures were achieved which could be suitably used for meta-analysis such as the standardized effect difference. Another impulse was the development of suitable statistical methods such as the appropriate form of a pooled mean.
Table 1. Number of defective items for 20 batches of 200 items each

<table>
<thead>
<tr>
<th>Batch</th>
<th>Number of defectives</th>
<th>Batch</th>
<th>Number of defectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>20</td>
<td>17</td>
</tr>
</tbody>
</table>

Meta-analysis experienced tremendous impulses by means of embedding important application areas such as evaluation research or health reporting. It is hoped that both areas discussed in this paper, namely quality control and assurance and meta-analysis, experience a similar impulse from each other.

It is quite obvious that in quality control the single batch can play the role of a single study in conventional meta-analysis. This can avoid various techniques including control charts and repeated testing which can be statistically flawed. For example, if for the data provided in Table 1 20 binomial tests are employed it can be expected that 1 of these will show a significant deviation from a desired standard though there is in fact no deviation from the desired standard (process is still in control). Similarly, if control charts are used it is well-known that the boundaries of these charts are reached for some batch, though the process is still in control. As consequence, investigators in quality assurance are forced to investigate for a non-existing source of deviation of the production process.

3 The Problem of Heterogeneity

3.1 The Occurrence of Heterogeneity

In fact, we are interested in separating random deviations which are occurring always in non-deterministic systems and systematic deviations. Only the latter are relevant and prone for further investigation and research.

How can we accomplish this separation? The first step is to model the situation when the process is in control which is called the situation of homogeneity. Typically, it is possible to derive some probability distribution for the measure of interest under homogeneity. We call the associated density

\[ f(x; \theta) \]

The question which system is deterministic and which is not is a mere philosophical question. Our point of view is that it is appropriate and useful to consider stochastic variation even measurements and processes are done with the highest accuracy.
of the measure of interest $X$: $f(x, \theta)$, where $\theta$ is some parameter involved in this density. In our example, the Number of Defective Items, $X$, follows a binomial distribution with density $f(x, \theta) = \binom{m}{x} \theta^x (1 - \theta)^{m-x}$, where $m$ is the size of the batch and the parameter $\theta$ corresponds to the allowed number of defectives. The question at hand is: what will happen if a deviation (loss in quality) occurs and how is this reflected in the statistical model. Clearly, if this happens homogeneity conditions no longer hold and the simple statistical model $f(x, \theta)$ will no longer be correct.

### 3.2 Diagnosis of Heterogeneity

There are some simple tests available which allow to diagnose this situation rather quickly. One of these tests is based upon the Chi-Square-statistic defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E(X_i))^2}{Var(X_i)}.$$  

Typically, $E(X_i)$ and $Var(X_i)$ will be functions of the unknown parameter $\theta$ and plug-in-estimates must be utilized. These plug-in estimators must be constructed with care to achieve $\chi^2$-distribution under homogeneity, at least approximately. To give a demonstration we note that in our binomial quality control example $E(X_i) = m\theta_i$ and $Var(X_i) = m\theta_i (1 - \theta_i)$ which might lead
to the plug-in estimates $E(X_i) = X_i$ and $Var(X_i) = X_i(1 - X_i/m)$. It can be shown that the associated distribution under homogeneity is quite different from a $\chi^2$-distribution with $n - 1$ df if sample sizes per batch, $m$, are small or moderate, even if the number of batches $m$ becomes large. The right thing to do here turns out to be a variance estimate utilizing information from all batches: $Var(X_i) = S_n(1 - S_n/m)$, where $S_n = \sum_{i=1}^{n} X_i/n$. The associated $\chi^2$-statistic (with $E(X_i) = S_n$) can be shown to be validly approximated by a $\chi^2$-distribution with $n - 1$ df even for small batch size $m$ (like $m = 5$). For further discussion see Böhning ([2000]) and Hartung ([1999]). To finish this aspect we find a value of $\chi^2 = 70.41$ with 19 df for the data of Table 1 which indicates strongly the presence of heterogeneity.

In the following section we will concentrate on the aspect: what can be done if heterogeneity is present?

4 Modelling Heterogeneity Using Mixture Distributions

If heterogeneity is present it is implied that the proportion of defectives in the batch is deviating in a systematic way from the required standard, in other words, it can be assumed that the hypothesis $\theta_1 = \theta_2 = ... = \theta_n = \theta$ is wrong and it is more reasonable to assume that for certain parts of the population of all possible batches a value (for the proportion of defectives) of $\theta_1$, for other parts a value of $\theta_2$ is valid and so forth. That is the population of possible batches consists of a proportion $p_j$ of batches with $\theta_j$, for $j = 1, ..., k$. It can be shown (Böhning [1999]) that then $X_i$ has a mixture distribution

$$f(x_i, P) = \sum_{j=1}^{k} f(x_i, \theta_j)p_j$$

which takes the form of a mixture of binomial distributions for our textbook example:

$$f(x_i, P) = \sum_{j=1}^{k} \binom{m}{x_i} \theta_j^{x_i}(1 - \theta_j)^{m-x_i}p_j$$

To mention a second example, let us consider the effect measure of the Fisher-transformed correlation coefficient: $\theta = 0.5\log\left[\frac{1+\rho}{1-\rho}\right]$. If we assume a normal distribution as conditional distribution for $\hat{\theta}_i$ with variance $1/(m_i - 3)$ in the $i$-study, then the mixture distribution takes on the form

$$f(\hat{\theta}_i, P) = \sqrt{m_i - 3} \sum_{j=1}^{k} \varphi(\sqrt{m_i - 3}(\hat{\theta}_i - \theta_j))p_j$$

where $\varphi$ is the standard normal density.
The distribution which gives probability mass $p_j$ to $\theta_j$ is called *mixing distribution* and is denoted by $P$. To estimate the parameters involved in (2), in other words the mixing distribution $P$ we use maximum likelihood estimation including the number of components in the mixture $k$. This can be accomplished with the computer package C.A.MAN, see Böhning et al. ([1992],[1998]). The associated maximum likelihood estimate (NPMLE) of $k$ and $\theta_j, p_j$ for $j = 1,...,k$ is called *nonparametric maximum likelihood estimate (NPMLE)* of the mixing distribution $P$. It is usually advisable to check whether the number of components $k$ can be reduced which can be accomplished by comparing log-likelihoods for reduced values of $k$ such as $k - 1, k - 2, ...$ until no significant drop in the log-likelihood is notable. For these fixed values of $k$ estimation is done via the EM-algorithm (Dempster et al. [1977]).

To provide a demonstration for this technique we again study the data of Table 1 and use the mixture model provided in (2). Tables 2 and 3 provide the results. There is empirical evidence for *heterogeneity* (see Table 2) and that this heterogeneity consists of 3 components (Table 3).

It can be seen that the population of batches can be separated into *three* components. One component consists of batches which are free of defective items (9.9%). The second component has 2.87 defective items per 100 (13.3%), whereas the last one has 8.6 defective items per 100 and this components represents the majority of all batches (76.8%).
Statistical Methods of Meta-Analysis 541

<table>
<thead>
<tr>
<th>Number of components $k$</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 (NPMLE)</td>
<td>−63.1454</td>
</tr>
<tr>
<td>3</td>
<td>−64.0984</td>
</tr>
<tr>
<td>2</td>
<td>−70.9835</td>
</tr>
</tbody>
</table>

Table 2. Heterogeneity structure for 20 batches of 200 items each

<table>
<thead>
<tr>
<th>Proportion $\theta_j$</th>
<th>Weight $p_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>0.0996</td>
</tr>
<tr>
<td>0.0287</td>
<td>0.1326</td>
</tr>
<tr>
<td>0.0865</td>
<td>0.7678</td>
</tr>
</tbody>
</table>

Table 3. Estimated mixing distribution for $k = 3$

Finally, it is even possible to allocate each observed (investigated) batch to one of the components in the mixture. This can be accomplished by utilizing Bayes theorem and calculate the posterior distribution of $\theta$ as

$$ f(\theta|x_i) = \frac{f(x_i, \hat{\theta}_j)p_j}{\sum_{l=1}^{k} f(x_i, \hat{\theta}_l)p_l} $$

and $f(\theta|x_i) = 0$ otherwise, where $\hat{\theta}_j$ and $\hat{p}_j$ corresponds to the maximum likelihood estimate identified in the previous estimation process. Each batch $i$ with Number of Defectives $X_i$ is allocated to that component $j$ for which $f(\theta_j|x_i)$ is largest of all $j = 1, ..., k$. This is done for the data in Table 1 and the results are provided in Table 4. Figure 2 also visualizes this reclassification. This technique might enable the practitioner to search for common sources for the occurred heterogeneity and finally identify sources for the loss in quality standards.

5 META – A Software Package for Meta-Analysis in Medicine, Social Sciences and the Pharmaceutical Industry

The software META has been developed to provide a tool which allows to perform meta-analyses within the areas of application described above. The focus of meta is on the analysis of heterogeneity, which may be considered here the unifying concept for several fields of application.

For different areas of application different measures of effect are important and necessary. Thus META enables the meta-analyst to choose out of a variety of measures of effect such as the relative risk in medicine, the standardized
<table>
<thead>
<tr>
<th>Batch $i$</th>
<th>$X_i$</th>
<th>Component $j$</th>
<th>Batch $i$</th>
<th>$X_i$</th>
<th>Component $j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>3</td>
<td>11</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>3</td>
<td>12</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>3</td>
<td>13</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>3</td>
<td>14</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>3</td>
<td>15</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>3</td>
<td>16</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>3</td>
<td>17</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>3</td>
<td>18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>3</td>
<td>19</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>3</td>
<td>20</td>
<td>17</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4. Classification of each batch into the components

Fig. 3. Spreadsheet with original data and empirical Bayes estimates

difference in psychology and proportions in quality control, just to mention a few.

META provides various statistical methods to perform meta-analyses such as simple pooled estimates, random effects models and graphical procedures such as confidence interval plots, funnel plots etc. We will illustrate the possible use of META using a data set from psychiatric epidemiology.

5.1 A Worked Example from Psychiatric Epidemiology

The following meta-analysis investigates the prevalence of agoraphobia based on seven studies (Eaton, [1995]) in several countries all over the world. Ago-
raphobia may be defined as space anxiety, as a fear of being in public places. This psychiatric disorder may even lead to total avoidance of public places and thus may cause severe disability.

An initial step in any meta-analysis might be to plot the effect measure together with a 95-percent confidence interval. This may be done using META and its graphics facilities. Figure 2 shows a screen dump of META and its data window. The data window shows the prevalent cases of agoraphobia together with the population at risk of the respective study.

The simplest model possible assumes parametric density \( f(x, \theta, \sigma^2) \) for some random quantity \( X \) where \( \theta \) is a parameter of interest and \( \sigma^2 \) is a nuisance parameter which might or might not be present in the model. In the example at hand, \( f(x, \theta) = \binom{m}{x} \theta^x (1 - \theta)^{m-x} \). In this case all studies are assumed to measure the same overall effect \( \theta \) and that they only differ in variability. Thus the summary measure needs to assign weights according to the inverse of the variance of the individual study in order to obtain the summary measure.

Looking at the confidence interval plot (Figure 5) there seems to be a large degree of variability to be present. However frequently one is interested in obtaining a summary measure for all studies. Using META we obtain the following results:

POOLED ESTIMATOR FOR PROPORTIONS

RESULTS

\[
\text{Pooled estimate: } 0.048892 \\
\text{Common variance: } 0.00000145 \\
\text{95 percent confidence interval (0.04654, 0.05125)}
\]

\[\chi^2\text{-test for homogeneity of proportions:} \]
\[115.23539 \text{ df = 6 p-value: 0.00000}\]

Clearly looking at the value of the \( \chi^2 \)-test of homogeneity we reject the null-hypothesis and conclude that there is substantial heterogeneity in terms of the prevalence of agoraphobia in the countries studied. As a result the computation of an overall rate is not very meaningful, since we ignore the underlying heterogeneity.

In order to deal with heterogeneity a mixture model has been implemented in META, as described in the previous sections. As it is assumed that \( \theta \) is not constant, but is varying itself according to some further distribution \( P \), we are able to consider the moments \( E_P(\theta) = \mu \) and \( Var_P(\theta) = \tau^2 \) of the heterogeneity distribution \( P \). Frequently \( \tau^2 \) is called the heterogeneity variance. META offers modelling according to two different concepts in order to deal with heterogeneity: One is a moment approach and is based on equating the expected value of the \( \chi^2 \)-statistic to the observed one and the solve for \( \tau^2 \). Actually this is the approach by DerSimonian and Laird ([1986]). The
DerSimonian-Laird-Estimator is provided by

\[
\tau^2 = \frac{\chi^2 - (n - 1)}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}}
\]

where \( w_i = \text{Var}(\hat{\theta}_i)^{-1} \), the inverse of the variance of the measure of interest in the \( i \)-th study and \( \chi^2 = \sum_i w_i (\hat{\theta}_i - \hat{\mu})^2 \) with \( \hat{\mu} = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i} \). \( \tau^2 \) will only be computed if \( \chi^2 \) is larger than \( n - 1 \); otherwise it is set to zero. Having estimated \( \tau^2 \) a pooled estimator is computed using the weights \( w_i^* = \{\text{Var}(\hat{\theta}_i) + \tau^2\}^{-1} \).

The other approach does find the nonparametric maximum likelihood estimator of the mixture model as outlined in section 4.

We proceed in our analysis with the estimation of the DerSimonian-Laird estimator:

RESULTS
Pooled estimate: 0.0455
(adjusted for heterogeneity using DerSimonian-Laird)

Heterogeneity variance: 0.0003
Variance of pooled estimator: 0.0000465

0.04545 95 percent CI: (0.0321, 0.0588)

Please note that we find a substantial value for the heterogeneity variance \( \tau^2 \) in this data set. As expected incorporating heterogeneity leads to a larger variance for the DerSimonian-Laird estimator. As a result we obtain a much wider confidence interval compared to the pooled estimator where we assume a constant value for \( \theta \) (see also Figs. 3 and 6).

Frequently there is a debate, whether one should use a summary measure in the presence of heterogeneity. One might argue that this may be done, but one has to be careful how to interpret the results. Under the presence of heterogeneity a summary measure will reflect the overall mean in the population well knowing that this effect might be different in subparts of the population.

If the presence of heterogeneity has been identified one might wish to model the structure of this heterogeneity and for example find the levels of effect in subparts of the population. This can be accomplished using the finite mixture model approach outlined above.

A convenient computational strategy uses a fixed grid of potential support points (subpopulation means \( \theta_j \)) which may or may not receive weights \( p_j \).

Figure 4 shows the dialog box which allows the user to define a grid of potential support points. Depending on the current measure of effect an appropriate mixing kernel may be chosen by the user. In this case since we are dealing with rates and the binomial distribution is the natural choice.
Statistical Methods of Meta-Analysis

Choose mixing kernel
- Normal (Fixed variance)
- Poisson
- Binomial
- Normal

Please define a grid of support points
- Minimum: 0.0211
- Maximum: 0.0690
- Number of support points: 20
- Max. iterations: 1500

Fig. 4. Dialog box for flexible number of components mixture model

Initial number of components: 5
Parameter: 0.0211, Weight: 0.1441
Parameter: 0.0317, Weight: 0.2840
Parameter: 0.0530, Weight: 0.3073
Parameter: 0.0584, Weight: 0.1533
Parameter: 0.0690, Weight: 0.1113

Log-likelihood at iterate: -34.8009

Based on this grid META identifies five potential subpopulations. Now these grid points with positive support may be used to find a refined solution using the EM algorithm (Dempster, Laird and Rubin, [1977]). Here we keep the number of components fixed and update mixing weights and subpopulation means. Frequently some population means coincide and thus the number of components decreases. For our data at hand we find after applying the EM-algorithm four remaining components. (Results not shown here). Now a backward elimination approach may be used in order to reduce the number
of mixing components. This would imply that we test $k = 4$ vs $k = 3$ using a Likelihood Ratio test approach.

**NPMLE for Fixed support size**

**Number of components after combining equal parameter estimates:** 3

- **Parameter:** 0.0212, Weight: 0.1440
- **Parameter:** 0.0316, Weight: 0.2844
- **Parameter:** 0.0559, Weight: 0.5716

**Log-likelihood at iterate:** -34.3889

Clearly the log-likelihood is only slightly smaller for this three component mixture model and we would conclude that a three component solution is appropriate.

Once a mixture model has been chosen, one might be interested in classifying the individual study. Due to their discrete structure mixture models provide a natural way of classifying the individual study. This is achieved by applying Bayes theorem (see (4)). According to this rule the $i$-th study is
then assigned to that subpopulation $j$ for which it has the highest posterior probability of belonging. \textsc{META} offers the option to classify the studies and to store the results of this classification in the data spreadsheet.

\textsc{META} also computes the posterior expectation for the measure of effect for the individual study based on the assumed distribution. Likewise also the posterior expectations may be stored within the data frame.

5.2 Availability

\textsc{META} is designed to be platform independent and uses the wxWindows 2.0 class library (Julian Smart [2000]). \textsc{META} may be obtained for Microsoft Windows 9x/NT and for Unix/Linux operating systems. \textsc{META} is available from the authors on request.

6 META's Special Module: Quality Control

6.1 Quality Measures for Parenteral Drugs

As has been outlined in Sect. 2.2, parenteral drugs are products which have to comply with additional, specific properties like sterility and essentially free of visible particles because of their parenteral application. Each parenteral container is controlled by a 100%-inspection for particulate matter. The particulate matter is evaluated for particles which can be seen easily, good or difficult, leading to the index $Q_{TR} = \frac{A}{N}$, where $A$ stands for number
of points recorded by three test persons and \( N \) stands for the number of controlled containers. This measure and similar measures are forming the first part of this module which provides then a heterogeneity analysis in the sense of DerSimonian-Laird.

### 6.2 Estimating Mixture Distributions for Grouped Count Data

A more complete modelling is provided for the following situation. During the production process of a pharmaceutical product there are conducted a number of controlling measurements concerning bacteriological contamination in the air, on surfaces and on working clothes. This is denoted by microbiological in-process controlling (MIPC). From this process result counting data \( w_i, i = 1, \ldots, n \). Here \( w_i \) denotes the number of colony forming units per surface – respectively volume unit (CBU = colony building units), and \( n \) is the total number of measurements. The particular issue here consists in the fact that for the statistical analysis the data are not available as in its original form as exact count \( W_i \), but as grouped data: there is a disjoint decomposition of the positive half axis given in the form

\[
0 = r_0 < r_1 < \cdots < r_{L-1} < r_L := \infty , \quad r_l \in \mathbb{N}, \quad 1 \leq l \leq L - 1.
\]

Furthermore we have the following notation: \( R_l = \{r_{l-1}, \ldots, r_l - 1\}, \quad l = 1, \ldots, L - 1, \quad R_L = \{r_{L-1}, r_{L-1} + 1, \ldots\} \). The observations available for the analysis are

\[
n_l := \# \{i \in \{1, \ldots, n\} : w_i \in R_l\}
\]

that means, \( n_l \) is the number of measurements for which the number of detected CBU falls within the group \( R_l \). We have a loss of information: instead of the vector \( \mathbf{w} = (w_1, \ldots, w_n) \) of original data the vector of ‘observed’ data is now \( \mathbf{y} = (n_1, \ldots, n_L) \).

Let us consider an example from microbiological environmental monitoring. The raw data (Table 5) consist of the number of detected CBU per surface unit on the overall of personal staff in an inprocess control department. Here, the total number of measurements is \( n = 164 \) with \( n_0 = 84 \) steril measurements (e.g. those samples with \( w_i = 0 \)), and we have \( L = 16 \) groups.

The following heterogeneity model is implemented in our analysis. The variables \( W_i \) are independent distributed according to a mixture of Poisson distributions with mixing distribution \( P \). This mixture distribution \( P \) is mixing \( k \) Poisson components \( Po(\theta_j) \):

\[
P(W_i = w|\theta_j) = \frac{\theta_j^w}{w!} e^{-\theta_j} , \quad w = 0,1,\cdots \text{ with } \theta_j > 0 , \quad j = 1,\cdots k.
\]

If we interpret the objects resp. individuals on which the measurements are done as representatives of a heterogenous population and for which the
Table 5. CBU counts for personal overall in inprocess control department

<table>
<thead>
<tr>
<th>Group number $l$</th>
<th>Left bound for the group $r_{l-1}$</th>
<th>Number of CBU $n_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>51</td>
<td>1</td>
</tr>
</tbody>
</table>

subpopulation membership in unknown, then the mixture model arises as the marginal model where the margin is taken over the latent variable denoting the subpopulation membership. The population weight of the component number $j$ is denoted by $p_j$ with

$$0 < p_j \leq 1, \quad j = 1, \ldots, k, \quad \sum_{j=1}^{k} p_j = 1.$$ 

The following quantities have to be determined resp. to be estimated:

1. $k$, the number of components,
2. the mixing fractions $p_j$,
3. the corresponding parameter $\theta_j$, the means within the single components.

The mixture distribution model is of the following form:

$$f(w; k, \Psi) := P_{k, \Psi} (\{W = w\}) = \sum_{j=1}^{k} p_j f(w; \theta_j) = \sum_{j=1}^{k} p_j \frac{\theta_j^w}{w!} e^{-\theta_j}, \quad (7)$$

with $\Psi = (\theta_1, \ldots, \theta_k; p_1, \ldots, p_{k-1})^T$. Combining (6) and (7) the estimation problem presented here exhibits two aspects:

1. There are to be estimated the parameter of a mixture distribution, for which
2. the data are at hand are of grouped form.
We use the maximum-likelihood principle for estimating the parameter $\Psi$. For this we have to begin with forming the likelihood function of the data $y$, which are available. It appears, that the problem can be divided into two subproblems:

(i) The determination of the number $k$ of mixture components and supplying initial estimations resp. starting values for part (ii).

(ii) The estimation of the parameter vector $\Psi$.

With regard to an algorithm for the numerical solution of the estimation problem it turns out, that the likelihood based on the ‘observed’ data $y$, $l(y)$, is not much helpful. For example, within the score equation $(\partial / \partial p_j)l(y)$ it is not possible to separate the parameter $p_j$. Therefore we preferred a solution by means of the EM-algorithm via constructing a ‘complete-data-likelihood’ (see for instance McLachlan and Krishnan [1997]). META is using the derived iterative equations to sequentially approximate the maximum-likelihood estimation according to the EM-algorithm. These iterative equations are of the structure

$$\theta_j^{(t+1)} = \theta_j^{(t)} F_\theta(j, \Psi^{(t)}) , \quad p_j^{(t+1)} = p_j^{(t)} F_p(j, \Psi^{(t)}).$$

At this $F_\theta$ and $F_p$ are expressions depending on the parameter values of the preceding step.

To solve the subproblem (i) we developed an initial algorithm. Within this algorithm initially we set $k = L$ and then generate a sequence of estimations $k(s), \Psi(s) = (\theta_1, \ldots, \theta_k(s); p_1, \ldots, p_{k(s)}-1)^T$ which is stopped if a stoppping rule is met, say at step number $s = S_0$. Then, we set $k := k(s_0)$ and $\Psi(s_0)$ is the vector of starting values for the algorithm of subproblem (ii).

For our example of counted CBU on personal overall the output of the algorithm is a mixing distribution with $k = 7$ components (see Table 6). Note that the weight is concentrated on components with small mean value, nevertheless there are three components with very small weight but rather large mean value:

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
<th>Mean value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.26176</td>
<td>0.00000</td>
</tr>
<tr>
<td>2</td>
<td>0.30838</td>
<td>0.40775</td>
</tr>
<tr>
<td>3</td>
<td>0.25031</td>
<td>1.71035</td>
</tr>
<tr>
<td>4</td>
<td>0.14184</td>
<td>5.92506</td>
</tr>
<tr>
<td>5</td>
<td>0.02016</td>
<td>22.4260</td>
</tr>
<tr>
<td>6</td>
<td>0.01269</td>
<td>43.4509</td>
</tr>
<tr>
<td>7</td>
<td>0.00486</td>
<td>59.5728</td>
</tr>
</tbody>
</table>

Table 6. Estimated mixing distribution for the example data in Table 5
In the light of the application we interpret the result as follows. It appears that the practitioner will focus attention on the components 5, 6, 7 and, consequently, on the samples which are associated with these components, in order to identify potential sources of deficiencies.

7 Discussion

We touched upon an approach which explicitly allows the modelling of heterogeneity. To do this it is important to emphasize that an appropriate measure of interest (describing the quality standards) has to be chosen. Given the chosen measure of interest it is furthermore equally important to find the corresponding statistical model under homogeneity conditions and further the associated mixture model which models potential heterogeneity. Important measures of interest have been considered from the areas of medicine and pharmacy (relative risk and risk difference), psychology (standardized difference and correlation coefficient) and – as a new area of application – quality control and assurance (quality index and count index). For these three application areas different modules have been developed and assembled to form a package META which allows the user in a simple way to analyze data in his/her application. As a special feature of the package META heterogeneity analysis is provided for each application area on the basis of mixture modelling.

Important aspects of future research will be:

1. valid computation of standard errors of the parameters involved in the mixture model and their associated confidence intervals
2. extension of the mixture models to allow covariate modelling
3. inclusion of these aspects in the package META

At the current stage not much is known about the correct computation of standard errors for the parameters of a mixture model. Basford, Greenway, McLachlan and Peel ([1997]) compare two methods of getting standard errors for the parameters of a mixture model of normal distributions. One method used in their computation is based on the conventional method of inversion of the information matrix. The other is based on the technique of the Bootstrap (Efron and Tibshirani [1993]). Their competitive analysis of both methods provides evidence that – though the information theoretic (and less computational expensive) approach is frequently close to the Bootstrap method – the Bootstrap seems to provide a useful alternative, especially for extreme cases of mixture models and small sample sizes. In addition, the paper by Basford, Greenway, McLachlan and Peel ([1997]) provides an example with larger sample size, in which the information based standard error is in considerable disagreement with the Bootstrap standard error. Therefore, it will be investigated whether it is possible to implement the Bootstrap approach for the mixture models used in our application areas to provide valid standard errors and, thus by, confidence intervals.
Another important area of further research will be the question in which way covariates can be included into the mixture modelling. Often additional information is provided to be used in the explanation of heterogeneity found in the data. For example, quality control sample data might come from different departments, from different shifts of quality control workers, or might have been collected at varying points in time. All these pieces of information might be collected and might form covariates which can be investigated for their potential in explaining heterogeneity. If the heterogeneity can be fully explained by the observed covariates, then the sources of variation in the quality control measure or index has been captured. It will be of importance to allow for heterogeneity given the covariates: we call this residual heterogeneity. If there is no residual heterogeneity, search for further sources of variation will be superfluous, at least in the sample data. Therefore, it will be important to model residual heterogeneity, which can be only validly accomplished if the univariate mixture model is extended for covariates.

It is targetted to extend META for these additional complexities in the near future.

Acknowledgements. This research is done under support of the Bundesministerium für Bildung und Wissenschaft, Forschung und Technologie (BMBF).

References


An Application for the Analysis of Human Tremor Time-Series

Josef Honerkamp¹, Michael Lauk¹, Jens Timmer¹, Carl-Herrmann Lücking², and Günther Deuschl³

¹ Center for Data-Analysis and Modeling and Department of Physics, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany
² Center for Neuroscience, University of Freiburg, Breisacherstr. 64, 79106 Freiburg, Germany
³ Department of Neurology, University of Kiel, Niemannsweg 147, Kiel, Germany

Abstract. For many diseases various methods for the diagnosis and treatment monitoring are available. Presently, such methods are not established for an investigation of human tremor diseases, although the different forms of tremor are common neurological symptoms and occur frequently in various neurological diseases and also other conditions. Human tremor is defined as an involuntary oscillation of one or more parts of the body. The goal of this project was to develop new methods and to improve and automate already proposed methods for the diagnosis and treatment monitoring of human tremors. We built an easy-to-use application for tremor-analysis and recording, running on conventional personal computers, that allows to investigate different forms of human tremor by advanced mathematical methods of time series analysis. The software is also applicable for users who are not familiar with these kind of advanced data analysis methods. It provides tools for the diagnosis and treatment monitoring under laboratory conditions, mainly based on the cross-spectral and spectral analysis of tremor acceleration time-series and electromyographic time-series. The software is now available on the market.

1 Introduction

Since many decades the electrophysiological and mathematical analysis of human tremor was subject to numerous studies. Tremor is defined as the involuntary, oscillatory movement of parts of the body, mainly the upper limbs. There are many different kinds of tremor, differentiated by clinical criterions. The better understanding of the generating mechanisms of the different human tremors could lead to an improvement of diagnosis and therapy of tremor diseases. Although the most common types of tremor, physiological tremor, enhanced physiological tremor, essential tremor and Parkinsonian Tremor, were investigated in many clinical and physiological studies, their mechanisms and origins are still unknown, for reviews see, e.g., [1–3]. Many other (sometimes very rare) tremors exist [1,3].

Shortly, the physiological tremor (PT) denotes a fine, low-amplitude oscillatory movement of the outstretched hand and is present more or less intense in all humans. Its origin is still under discussion. It was supposed to
originate from reflex loops [4-6], by random synchronization [7] or from a central oscillator [8-10]. The frequency of PT usually ranges from 7 Hz up to 11 Hz and depends on the weight of the hand, i.e. its frequency decreases if the outstretched hand is loaded with weights [3]. The so called enhanced physiological tremor (EPT) denotes a tremor whose frequency also depends on the load of the hand but with a pathological tremor amplitude. Prominent examples are the tremor caused by drug abuse, by excitement or by fear.

The essential tremor (ET) is a hereditary form of tremor and the most common tremor. Its frequency usually ranges from 4.5 Hz up to 10 Hz [3]. The Parkinsonian tremor (PD) is the second most common form of all pathological tremors. Its frequency usually ranges from 4 Hz up to 8 Hz [3]. It is one of the prominent symptoms of Parkinson’s Disease. The so called orthostatic tremor (OT) was supposed to be a special kind of ET [11-14] that only occurs during standing and shows an unusual high frequency around 15 Hz. However, recent observations suggest that OT is a separate form of tremor [15,16]. Although frequencies and amplitudes can differ substantially, they are not sufficient criterions for a reliable diagnosis of different pathological forms of tremor [1,3].

Despite the intensive work on tremor, until today, the electro-physiological analysis of tremor using advanced mathematical methods has not entered the daily clinical routine. E.g., the quantification of the tremor severity is almost done by looking at a persons shivering hand. An automatic computer based diagnose and treatment monitoring device is not available so far. The needed sensing techniques, amplification and computer hardware are available since many years. The missing part of such a device is an analysis software based on adequate mathematical methods to deal with various types of, in general, non-linear stochastic processes. A tool for tremor analysis, also available for people who are not specialized on this field could help to improve diagnosis and treatment monitoring of tremor diseases in widespread areas of neurology, rehabilitation and pharmacology.

2 Methods

2.1 Recording Techniques

We analyze three different kinds of time series. 1) accelerometer signals representing the mechanical movements of the recorded trembling limb (usually the hands), 2) surface electromyographies (EMG) representing the muscle activity as a potential difference between two electrodes placed on the skin over the respective muscle and 3) electroencephalographies (EEG) representing the cortical activity as potential measured over the scalp of a subject. To give an impression Fig. 1 shows three arbitrary time-series of a patient suffering from PD.

During the recording, subjects were sitting in a comfortable, heavy chair with their arms supported. The forearms were fixed proximal of the wrist.
with a strap. To measure the postural tremor, subjects were asked to hold their hands outstretched in pronated position and to avoid any voluntary movement. For the resting tremor measurements, subjects were asked to avoid any voluntary muscle contraction or movement of their hands. The duration of each record was 30 s. Hand acceleration (ACC) was recorded by light weight piezo-resistive accelerometers attached to the belly of the right and left hand [17,18]. Surface EMGs were recorded from the wrist flexor and
extensor muscles of the left and right forearm. In addition, in patients with orthostatic tremor and also in some patients with other tremors in the lower extremities, paravertebral muscles, and leg muscles were recorded.

ACCs and EMGs are band-pass filtered to avoid aliasing effects and undesired slow drifts (ACC: 0.5 Hz–50 Hz, EMG: 80 Hz–500 Hz). All data are simultaneously sampled at 1000 Hz and stored on computer using a special software [19]. The mean was subtracted from each time series. Finally, the series were tapered with a Bartlett-Window and normalized to unit variance. In addition, EMGs were digitally full wave rectified for spectral and cross-spectral analysis [20,1].

EEG recordings were performed using a 64-channel EEG system (Neuroscan). The potential field measured over the scalp was transformed into the reference free current density distribution which reflects the underlying cortical activity [21,22].

2.2 Spectral and Cross-Spectral Analysis

Spectral methods are mainly used to correctly quantify the tremors. A correct, robust and objective quantification is the basis for all clinical applications, in particular for the judgment of a treatment outcome. This usually implies the use of robust algorithms that work without defining individually changing parameter settings.

Cross-spectral methods provide a powerful tool to investigate the relation between simultaneously recorded signals. These methods have been used in tremor research to study the relation between electromyographies (EMG) and magnetoencephalograpgies (MEG) or electroencephalograpgies (EEG) [23–26], between EMGs and mechanical measurements [27–31,6,32–34,10,35–37], between EMGs [38,39,16,37] and between single motor units and EMGs [29,40–42].

The power spectrum $S(\omega)$ of a stationary, zero-mean time series $x(t)$ is defined as the Fourier transform of the autocorrelation function $A(\tau)$:

$$A(\tau) = <x(t)x(t-\tau)>$$

$$S(\omega) = \frac{1}{2\pi} \sum_{\tau} A(\tau) \exp(-i\omega \tau), \quad \omega \in [-\pi, \pi]$$

where “$<$” denotes expectation.

Analogously to the univariate quantities the cross-spectrum $S_{xy}(\omega)$, of two stationary, zero-mean time series $x(t)$ and $y(t)$ is defined as the Fourier transform (FT) of the cross-correlation function $C(\tau) = \langle x(t)y(t-\tau) \rangle$ [43,18,34]

$$S_{xy}(\omega) = \frac{1}{2\pi} \sum_{\tau} C(\tau) \exp(-i\omega \tau) = <\tilde{x}(\omega)\tilde{y}^*(\omega)>$$

where $*$ denotes complex conjugation and $\tilde{x}(\omega)$ and $\tilde{y}(\omega)$ are the Fourier transforms of $x(t)$ and $y(t)$, respectively.
The coherency spectrum $\kappa(\omega)$ is defined as the modulus of the normalized cross-spectrum $S_{xy}(\omega)$

$$\kappa(\omega) = \frac{|S_{xy}(\omega)|}{\sqrt{S_x(\omega)S_y(\omega)}} ,$$

where $S_x(\omega)$ and $S_y(\omega)$ denote the auto-spectra of $x(t)$ and $y(t)$ [43,44,18,34]. The phase spectrum $\Phi(\omega)$ is given by the representation:

$$S_{xy}(\omega) = |S_{xy}(\omega)| \exp(i\Phi(\omega)) .$$

The estimation of the power and cross spectra is done by a direct spectral estimation [43–45,18,34], based on the discrete Fourier-transform (FT) of the data. The simplest form of such a procedure is a sliding average of the periodogram, i.e. the estimation by the so called Daniell-window. We chose a triangular window, or the Bartlett estimator [43,44,18].

**Adaptive estimation of spectra** A crucial point within the estimation of tremor spectra is the choose of the optimal width $2h + 1$ of the smoothing window since it depends on the curvature of the true spectrum. For large curvatures the width should be smaller than for a small one. Since the curvature of the spectrum is not constant with respect to frequency, a frequency dependent smoothing window can improve the spectral estimation. For the estimation of tremor spectra we use proposed a frequency dependent smoothing procedure [18]. This data driven algorithm was developed for an optimal estimation of the spectrum in the region of the main peak, i.e. the global maximum of the spectrum. A detailed description of the estimation procedure and applications to measured data are given in [18].

**Confidence intervals for peak frequencies** Usually, the peak frequency and the power of a biological oscillations are of major interest. It is desirable to get a confidence region of a peak frequency or to decide whether differences in the observed spectral peak frequencies of two estimated spectra are significant, i.e. to test the hypothesis of a zero peak frequency difference. Especially for treatment monitoring tasks the availability of confidence regions and tests for significant differences are needed to decide whether a change in frequency is due to a biological variation or only within the confidence bands of the estimator.

We proposed a procedure to test this hypothesis [46]. Furthermore, a similar approach was used to get confidence regions for spectral peak frequencies [47]. The procedures are inspired by the parametric bootstrap [48] and the theory of the spectral estimation [44]. General aspects of bootstrapping in the frequency domain are discussed in [49,50]. Shortly, a resampling procedure draws several new realizations of the $\chi^2$ distributed periodogram. For each of these periodograms the peak frequency is estimated. The confidence interval is estimated by the distribution of the estimated frequencies [47,46].
procedure is validated by a large number of simulation studies with both linear and non-linear stochastic processes [47,46].

To date this method has been used in two different clinical studies [10,51].

**Cross-talk** The cross-correlation function can be used to detect the so-called cross-talk between different EMGs. Cross-talk describes the effect of volume conduction of the measured EMG potentials [52] that complicates the estimation of phases between antagonistic muscle pairs. The magnitude of cross-talk depends on different parameters like muscle-volume, skin-resistance and placement of the electrodes that cannot be fully controlled. The amount of cross-talk can be estimated from the discontinuity at lag zero of the cross-correlation function estimated between the un-rectified EMGs because of its instantaneous effect (see Fig. 2 for an example). Because of the part of the cross-correlation not originating from the cross-talk there is no possibility to estimate reliably the amount of cross-talk directly as percentage of the total variance. We therefore use this method only for the detection of strong cross-talk.

![Cross-correlation function](image)

**Fig. 2.** Cross-correlation function between a un-rectified extensor and flexor EMG of a patient with ET. The sharp discontinuity at the lag zero indicates that a strong cross-talk (a volume conduction effect) between the two recorded EMGs is present.

**Coherency in case of high curvature** If the cross spectrum exhibits a high curvature, i.e. sharp peaks, the estimation of coherency is not straightforward. In particular, in this case, a test against the hypothesis of zero coherency using un-modified coherency estimates will lead to false positive results. In simulation studies we discussed the limits of pre-whitening techniques and propose a procedure to solve this problem [16]. In short, we use a similar frequency-dependent adaptive procedure as for the estimation of auto-spectra above. The proposed method is validated in simulation studies [16]. The size and the power of the test against zero coherency is acceptable for the investigated non-linear stochastic dynamics [16].
Phase-spectra and delay estimation

In general the estimated phase-spectrum is a sum of different effects. Firstly, there might be a time delay $\delta$. Secondly, a transfer function $b(t)$ could modify the phase spectrum between two correlated signals $x(t)$ and $y(t)$, giving the model equation

$$y(t) = \int_{\delta}^{\infty} b(t')x(t - t')dt'. \quad (6)$$

We investigated the interpretation of phase-spectra for cases where the transfer function $b(t)$ is different from the simple case where $b(t)$ is given by a $\delta$-distribution for both linear and non-linear stochastic processes. The goal was to develop a robust algorithm for time delay estimations if little is known on the underlying physiological system. The pointwise interpretation of a non-zero phase-spectrum as a time-delay generally requires strong assumptions on the underlying processes. This interpretation is rarely valid in biological systems.

In simulations we tested the performance of a Hilbert-transform method that is able to determine both the delay $\delta$ and the phase of the transfer function $[53,54]$. Furthermore the validity of the least-squares time-delay estimation $[55-57]$ in case of non-linear stochastic processes was investigated $[54]$.

The results show that, whenever little is known about the system a Hilbert-transform method leads to acceptable results. For the case of non-linear stochastic oscillations the results indicate that the phase on the oscillator frequency and on higher harmonics behave as they were part of a broad band phase-spectrum of a linear stochastic processes $[54]$. For two of the applications this method is clearly superior to the widespread used simple interpretation of phase-spectra and led to altered physiological interpretations $[54]$.

2.3 Non-Linear Analysis Methods

Pathological tremors exhibit a nonlinear oscillation that is not strictly periodic. For time series of two types of pathological tremor (ET and PD), we investigated if the underlying dynamics should be regarded as chaotic or nonlinear stochastic oscillatory processes. The key question for a distinction is the order of the process. Chaotic processes are of at least third order, while nonlinear stochastic oscillatory processes can be of second order. Five time series analysis methods from nonlinear dynamics to determine the order of the processes were applied to tremor data $[58]$:

- The analysis of the local slopes of the correlation integrals $[59]$
- Poincaré sections and return maps $[60,61]$
- The local divergence of nearby trajectories $[62]$
- 'Deterministic versus stochastic' - plots $[63]$
- The so-called $\delta - \epsilon$ method $[64]$
The results suggest that the considered pathological tremors represent a nonlinear stochastic process of second order with additive dynamical noise.

It is a one of the standard tasks in time-series analysis to investigate whether characteristics extracted from measured data are able to support the differential diagnosis between healthy and different types of pathological states usually performed by clinical observations, see [65] for a collection of examples from different biomedical domains.

Visual inspection of clinically classified time series motivated Timmer et al. [66] to investigate asymmetric behavior of the time series with respect to the direction of time and to a change of sign [66]. Algorithmically, the first was captured by a measure for time reversibility based on the difference of conditional expectations forward and backward in time. The asymmetry with respect to a change of sign was best captured by a measure for the asymmetric decay of the autocorrelation function. We calculated this by the difference $\gamma_d$ of the absolute values of the first and the second maximum of the autocorrelation function.

In [66,67], the classification results for a set of 15 essential and 25 Parkinsonian tremor time series have been reported. The classification rate was 75% for the measure of time reversibility and 90% for the asymmetric decay of the autocorrelation function. The classification rate based on the linear measures of amplitude and frequency is 30%, evaluated for a sample of 62 Parkinsonian and 52 essential tremor time series [3]. We repeated the nonlinear analysis for a larger set of 75 essential and 112 Parkinsonian clinically classified tremor time series [58]. Figure 3 compares the former results based on the small sample with the new results based on the larger sample. Assuming the larger sample to be representative, it can be seen that the high classification rates of

![Fig. 3. Comparison of the distribution of classification features between the results based on small and on large samples of essential (ET) and Parkinsonian disease (PD) tremor time series. (a) Time reversibility (b) Asymmetric decay of the autocorrelation function. +: essential tremor, small sample, x: essential tremor, large sample o: Parkinsonian tremor, small sample, Δ: Parkinsonian tremor, large sample](image-url)
the former analysis were caused by a non-representative sample of essential tremor time series. For both features the variance of their distributions is much smaller for the small sample. Thus, the variability of the processes is underestimated.

For the larger sample investigated here, the classification rate decreases to 30%, the same rate as found by the linear measures of amplitude and frequency.

3 Physiological and Clinical Studies

A large number of physiological and clinical studies accompanied this project. Clinical test trials and studies form an essential part of such a project to validate a proposed method and to build databases with normative values.

In order to identify the role of reflexes in the generation of physiological tremor, we investigated in model simulations phase-spectra and and auto-spectra under the two hypothesis that 1) reflexes contribute significantly and 2) that they do not play a significant role [34,10]. In the model study, we described the tremor dynamics by a second-order linear stochastic difference equation that is driven by a non-white external force [34,10]. The possibly present reflex-feedback was modeled by a non-linear positive feedback system [10]. This model presents a non-linear, stochastic, inhomogeneous delay differential equation [10]. In 35 time series recorded from 19 subjects with the respective type of physiological tremor, we clearly found that reflexes contribute to the tremor [10].

Given the fact that the tremors introduced in section 1 occur always (physiological tremor [PT], enhanced physiological tremor [EPT], essential tremor [ET] and orthostatic tremor [OT]) or often (Parkinsonian tremor [PD]) bilaterally, showing similar frequencies on both sides, it is of interest to study whether this can be explained by a common source or commonly transmitted signals or by two distinct oscillators. To clarify this question, we investigated coherency estimates of these types of tremor using the method mentioned in Sect. 2.2 [16]. The orthostatic tremor (OT) was side-to-side correlated whereas all other forms of tremor were not. This gives the possibility of a reliable diagnosis of OT.

Further studies using the proposed analysis methods were done to clarify a participation of central structures in the generation of PT [9], to investigate correlations between electroencephalographies and EMGs [26,37,68], and to investigate the involvement of cranial muscles in orthostatic tremor [15]. A normative multi-center study that was recently finished [69] provides an important database for the daily clinical application of the software. Furthermore, the biological variability of frequencies and phases between antagonistic muscle pairs has been studied [51].
4 The Application

In this section we describe the last step in the development of the analysis software. At a first glance one might get the impression that the programming of a software tool for a daily use of the proposed methods is mainly a problem of the applied sciences and information technology. This is not the case. To our experience, this last step to build a bridge should be done by the theoretician. Unfortunately the importance of this step is often underestimated by the researchers working on the more theoretical side of a problem. A lot of good approaches to solve “real-world” time-series analytical problems never reach the application because the developer of the methods do not provide an applicable, easy-to-use software tool. A detailed description of the software is given in [19].

One of the problems was that this software was intended to serve two different user groups. Firstly, people who want to analyze tremor records in a standardized daily clinical procedure, and secondly, users who work in the different fields of tremor research. The difference between the requirements of these two groups is the way they are working with an application: For the first group the software should provide the possibility to get the results of a tremor analysis by just “pushing one or two buttons” on the desktop, whereas researchers usually need the possibility to modify configurations and parameters and to export the different results in other applications.

The software was planned, written and tested together with clinical laboratories during the different stages of the development. Therefore, we could secure that our application is able to fulfill different criteria that would be necessary to build a software which will be accepted by clinical and laboratory staff. The program provides the following features:

- Data recording.
- Empirical tests for the quality of recorded time-series.
- Analysis: the spectral and cross-spectral methods described above, highly automated and easy-to-use.
- If possible results are given with confidence regions for a particular value to allow a comparison to well investigated populations of different groups (i.e. controls and patients with different tremor diseases).
- Supporting the user in interpreting qualitative results of spectral and cross spectral methods in terms of physics and physiology.

Finally, a comfortable graphical and interactive user-interface with support for mouse, menus, icons and buttons provides a user-friendly, common interface which gives potentially any person the possibility to work with it.

The application uses a multi-document-interface (MDI). Thus, the user can work with it in the same way he is used to work with common applications, e.g. word processing programs. A further advantage of this strategy is the possibility to compare different records and results directly on the screen. The only restriction is the amount of available system resources.
Our software shows that it is possible to integrate advanced mathematical methods in an easy-to-use application. In the meantime, our application is accepted among clinical staff and medical doctors as a convenient, powerful and fast but also secure tool. It has been used for daily clinical routine as well as research purposes in different European hospitals. More than 15000 tremor data-sets were recorded and analyzed during the last 5 years. Figure 4 displays a screen-shot of the application frame window to give an impression of the software.

5 Discussion

The examples in sections 3 demonstrate the usefulness of cross-spectral methods in physiology and medicine. One example for the use in a daily clinical routine is the diagnosis of orthostatic tremor by a side-to-side coherency. Together with the high frequency (>12Hz) this is a highly reliable diagnosis criterion, that can be used without any further clinical knowledge about this tremor disorder.
The first applications of the differential diagnosis between ET and PD gave good results [66,67]. However, we saw that an application of the proposed methods to larger samples of patients led to a poor differentiation between ET and PD. This is a good example that it is always possible to separate a small collective of a few typical patients from another small sample, if one looks “long enough” for a differentiating parameter. The real value of a new method as a diagnostic tool can be certainly only estimated by large prospective studies.

The first goal of lots of different proposed tremor analysis methods was always to differentiate ET from PD. Standard spectral and cross-spectral parameters as frequency, power and coherency between different EMGs do not sufficiently separate the two tremors [3,70]. Also the phase between antagonistic muscle pairs is not a diagnostic criterion to differentiate ET and PD [38]. In the recently published paper by our group [58] it has been shown that there is also no considerable dynamical difference between ET and PD. In fact, we see more similarities than differences between PD and ET. Therefore, it has been hypothesized that the tremor in ET and PD could be driven by the same generator [70].

We always tried to work as close as possible with the future users, i.e. medical doctors and their laboratory assistants. This was one of the most important issues in developing such an application. Based on the results of the clinical tests and suggestions of clinical staff or complaints of the users, we are permanently modifying the application to improve its performance and user acceptance. Furthermore, we implemented (and still implement) newly established analysis methods as fast as possible and adopt new results from different fields of tremor research.

The concept to develop specialized software for particular tasks might seem to be expensive and time consumable. On the other hand, it is the only way to provide applications which are accepted by those who are not familiar with the used mathematical methods and/or algorithms. Other software packages which basically provide tools for time series analysis are available. But usually they require fully-fledged staff, which is too expensive for a daily clinical routine. Secondly, numerous mathematical methods cannot be applied “naively” to arbitrary time-series. It is often essential to modify existing concepts to treat a special problem.

In summary, a tool for an objective, reliable and fast quantification of human tremors is the first step towards a measurement device for human tremors that also defines a new “gold standard” for tremor measurements. Also for an objective quantification of treatment outcomes such a device is essential. To date almost all neurologists quantify tremor disorders and treatment outcomes by eye. To propagate our methods we developed an easy-to-use software tool for tremor recording and analysis that is now available on the market. The value of the proposed methods for a computerized, stand-alone differential diagnosis between ET and PD in a daily clinical routine is poor. But in combination with further clinical knowledge the measurements
are valuable parameters for the diagnosis of ET or PD. Especially the reliable detection of sub-clinical symptoms, i.e. very mild symptoms at the beginning stage of a disease, is only possible with the use of technical measurements and robust data-analytical tools based on adequate mathematical methods.

References


IX. Optimization in Design and Production

Free Material Optimization
J. Zowe, M. Kovcvara

Automatic Layout and Labelling of State Diagrams
P. Mutzel, G. W. Klau

Optimization Problems in a Semi-Automatic Device for Cutting Leather
A. Pott, H. Glaab

Stochastic Programming for Power Production and Trading under Uncertainty
R. Schultz, M. P. Nowak, R. Nürnberg, W. Römisch, M. Westphalen

Scheduling Scarce Resources in Chemical Engineering
R. H. Möhring, M. Uetz
Free Material Optimization

Jochem Zowe and Michal Kočvara

Institut für Angewandte Mathematik, Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
E-mail: zowe@am.uni-erlangen.de, URL: http://www2.am.uni-erlangen.de

Abstract. Free material design deals with the question of finding the stiffest structure with respect to one or more given loads which can be made when both the distribution of material as the material itself can be freely varied. We present the single and multiple-load situation (understood in the worst-case sense). We further introduce a software tool moped for free material optimization of general two-dimensional bodies. The graphical post-processor of moped is tailored for a design of fibre-reinforced composite materials used by our industrial partner DaimlerChrysler Aerospace. Finally, we generalize the above approach to the case of uncertain loads in order to design an optimal robust material.

1 Introduction

One of the basic problems of structural engineering reads: For a given set of boundary conditions and a given set of loads, find the stiffest structure of a given volume that is able to carry the loads. In this paper, the desired optimal structure is considered to be a two- or three-dimensional continuum elastic body and the design variables are the material properties which may vary from point to point. Our aim is to optimize not only the distribution of material but also the material properties themselves. We are thus looking for the ultimately best structure among all possible elastic continua, in a framework of what is now usually referred to as “free material design”.

Optimization of structures is traditionally performed through the variation of sizing variables (e.g., thicknesses of bars in a truss) and shape variables (e.g., splines defining the boundary of a body). With the appearance of composites and other advanced man-made materials it has been natural to extend this variation to the material choice itself. Moreover, the study of generalized shape design and topology design leads one to consider design of material properties, through the requirement for relaxation of initially ill-posed problems, and the quite general problem considered in this paper can be seen as a natural generalization of such settings (the relaxation approach [1] or the homogenization method [5]). Thus the impetus to study material design in a structural optimization framework has come both from the increased industrial use of advanced materials as well as from developments within the structural optimization field itself.
2 Free Material Optimization

We study the optimization of the design of a continuum structure that is loaded by multiple independent forces. In order to deal with the problem in a very general form, we consider the distribution of the material in space as well as the material properties at each point as design variables. The idea to treat the material itself as a function of the space variable goes back to the works [6,12] and has also been studied in various other contexts; see [5].

For an easier understanding of the problem we start with the single-load case in two-dimensional space and only sketch the (technical) multiple-load case in the subsequent section. A generalization of both problems to three-dimensional space is straightforward.

2.1 Single-Load Case

Our goal is to find optimal distribution and properties of an elastic material within a given body \( \Omega \subset \mathbb{R}^2 \). By \( u(x) = (u_1(x), u_2(x)) \in [H^1(\Omega)]^2 \) we denote the displacement vector at point \( x \) of the body under load, by

\[
e_{ij}(u(x)) = \frac{1}{2} \left( \frac{\partial u_i(x)}{\partial x_j} + \frac{\partial u_j(x)}{\partial x_i} \right) \quad \text{for } i,j = 1,2
\]

the (small-)strain tensor and by \( \sigma_{ij}(x) (i,j = 1,2) \) the stress tensor. We assume that our system is governed by linear Hooke's law, i.e., the stress is a linear function of the strain

\[
\sigma_{ij}(x) = E_{ijkl}(x)e_{kl}(u(x)) \quad \text{(in tensor notation),}
\]

where \( E \) is the so-called (plain-stress) elasticity tensor of order 4. The strain and stress tensors are symmetric and also \( E \) is symmetric. We use a standard engineering notation and write the strain and stress tensors as vectors

\[
e = (e_{11}, e_{22}, \sqrt{2}e_{12})^T \in \mathbb{R}^3, \quad \sigma = (\sigma_{11}, \sigma_{22}, \sqrt{2}\sigma_{12})^T \in \mathbb{R}^3.
\]

Correspondingly, the 4-tensor \( E \) can be written as a symmetric \( 3 \times 3 \) matrix. In this notation, equation (1) reads as

\[
\sigma(x) = E(x)e(u(x)).
\]

For the elasticity tensor \( E \) and a given external load function \( f \) (applied at a part of the boundary \( \partial \Omega \) denoted by \( \Gamma_2 \)) the potential energy of an elastic body as a function of the displacement function \( u \) is given by

\[
\Pi(E, u) := \int_{\Omega} \frac{1}{2} \langle E(x)e(u(x)), e(u(x)) \rangle dx - \int_{\Gamma_2} f(x) \cdot u(x) dx.
\]
The system is in equilibrium (outer forces and inner reaction forces balance each other) for any \( u \) which minimizes the potential energy over a convex set \( K \subset [H^1(\Omega)]^2 \), i.e.,

\[
\min_{u \in K} \Pi(E, u). 
\] (2)

We now come to the crucial issue. Whereas (2) is the job of the nature, the designer will try to find a material (i.e., a matrix function \( E \)) for which the optimal potential energy from (2) is as close to zero as possible. That means, the response of the body to the given load (the displacements and strains) is small; the material is as stiff (with respect to \( f \)) as possible. Hence, as the optimal value in (2) is always a non-positive number, we try to find such \( E \) (under physical and economical constraints) that the “\( \min \)” in (2) is as big as possible. Physics tells us that the elasticity matrix \( E \) has to be symmetric and positive semi-definite on all of \( \Omega \), what we write as

\[ E(x) \succeq 0 \quad \text{for all } x \in \Omega. \]

As “cost” constraints on \( E \) we use the trace of \( E \) and require with given positive \( V \):

\[ \int_{\Omega} \text{tr}(E(x))\,dx \leq V. \]

To exclude singularities, we further require with some \( \tilde{t} > 0 \)

\[ 0 \leq \text{tr}(E(x)) \leq \tilde{t} \quad \text{for all } x \in \Omega. \]

When we collect all the constraints into a set

\[ \mathcal{E} := \left\{ E \in [L^\infty(\Omega)]^{3 \times 3} \mid E \succeq 0, \int_{\Omega} \text{tr}(E)\,dx \leq V, \ 0 \leq \psi(E) \leq \tilde{t} \right\}, \]

our design problem becomes

\[ \max_{E \in \mathcal{E}} \min_{u \in K} \Pi(E, u). \] (3)

**Theorem 1 ([13]).** Problem (3) has an optimal solution \( (E^*, u^*) \in L^\infty(\Omega) \times H^1(\Omega) \).

Given the existence of an optimal elasticity matrix \( E^* \), we ask how to “compute” this matrix function \( E^* \). It proves to be crucial for this step to separate the local part of the design variables (material properties at each \( x \)) and the global aspect (overall distribution of the material described by the cost function). This separation will allow an explicit partial analytic maximization with respect to \( E \) and leaves us with an optimization problem in \( u \) and \( \rho := \text{tr}(E) \) only. That means, the 6 components of \( E \) reduce to one real variable \( \rho \) after this partial analytic solution of the problem. Moreover, the positive semi-definiteness constraint disappears.

We skip here these technical steps and arrive directly at the reduced problem (a detailed analysis is given in [13]):
Theorem 2. The optimal material problem (3) reduces to
\[
\max_{\rho \in L^\infty(\Omega)} \min_{u \in K} \left\{ \int_\Omega \left[ \frac{1}{2} \rho(x)(e(u(x)), e(u(x))) \right] \, dx - \int_{\Gamma_2} f(x) \cdot u(x) \, dx \right\}.
\] (4)

This problem has an optimal solution \((\rho^*, u^*)\) from which an optimal \(E^*\) can be recovered as
\[
E^*(x) = \frac{\rho^*(x)}{\|e(u^*(x))\|_2^2} e(u^*(x)) e^T(u^*(x)).
\]

Remark. If we choose a coordinate system based on the orthonormal eigenvectors corresponding to the eigenvalues \(e_I, e_{II}\) of the strain tensor, the optimal elasticity matrix reduces to
\[
E^* = \frac{1}{e_I^2 + e_{II}^2} \begin{pmatrix}
e_I & e_{II} \\
e_{II} & e_I \\
0 & 0
\end{pmatrix}.
\] (5)

The optimal material (5) is orthotropic. Also, it has only one non-zero eigenvalue and can only carry strain fields which are re-scalings of the given strain field for which the optimization was undertaken.

In order to solve our (infinite-dimensional) problems numerically, we have to discretize them. For that we use the finite element method, working with piece-wise constant approximation of the “density” function \(\rho\) and piece-wise bilinear approximation of the displacement field \(u\). With \(A_i\) being element stiffness matrices and \(f\) the discretized right-hand side, the discrete version of (4) takes the form
\[
\max_{\rho \geq 0} \min_{u \in K} \left[ \frac{1}{2} u^T \left( \sum_{i=1}^m \rho_i A_i \right) u - f^T u \right],
\] (6)

where \(m\) is the number of finite elements.

In the final step, we eliminate the variable \(\rho\). Using a standard Minimax Theorem and LP theory, (6) can be reduced to minimization of a finite max-function:
\[
(6) = \min_{u \in K} F(u) \quad \text{with} \quad F(u) := \max_{1 \leq i \leq m} \left( \frac{mV}{2} u^T A_i u - f^T u \right).
\] (7)

\(F\) is a nonsmooth convex function whose minimization requires special software. To convert the problem to a smooth one, we use a standard reformulation of (7) by adding an auxiliary variable \(\alpha\)
\[
(6) = \min_{\alpha \in \mathbb{R}, u \in K} \left\{ \alpha - f^T u \mid \alpha \geq \frac{mV}{2} u^T A_i u \text{ for } i = 1, \ldots, m \right\}.
\] (8)
The quadratically constrained quadratic program (8) is open to the powerful modern modified-barrier methods and interior-point codes introduced in [8], [7].

### 2.2 Multiple-Load Case

A solution of the single-load problem (3) will be optimal precisely for the one considered load \( f \) and might be extremely unstable (even collapse) under other loads than \( f \). In a realistic scenario one should look for a structure which can withstand a whole set of loads \( f_\ell, \ell = 1, \ldots, L \), and which is in addition the best one, e.g., in the worst-case sense. This leads to the following multiple-load problem, in which we seek the design function \( E \) which yields the smallest possible worst-case compliance

\[
\inf_{E \in \mathcal{E}} \sup_{\ell = 1, \ldots, L} \sup_{u \in \mathcal{K}} \left\{ -\frac{1}{2} \int_\Omega (E e(u), e(u)) \, dx + \int_{\Gamma_2} f^\ell \cdot u \, dx \right\}. \tag{9}
\]

By introducing a weight vector \( \lambda \), we can eliminate the discrete character of the " \( \sup \) " and formulate (9) as a saddle-point problem. For the new formulation we can prove again existence of the optimal material \( E^* \) [2].

To further rearrange our problem (and to be able to solve it numerically), we discretize it in the same way as the single-load problem. The resulting problem is, however, much more complicated. After a series of steps, described in detail in [2], the discretized problem can be formulated as a *Semidefinite Programming problem* (SDP)

\[
\inf_{(u^1, \ldots, u^L; \lambda) \in \mathcal{V}} \sup_{\alpha \in \mathbb{R}} \frac{\alpha V - 2}{\sum_{\ell=1}^L (f^\ell)^T u^\ell}
\]

s.t.

\[
Z_i(u^1, \ldots, u^L; \lambda) \succeq 0 \quad \text{for } i = 1, \ldots, m,
\]

where \( Z_i \) are \((2 + 4L) \times (2 + 4L)\) symmetric matrices and \( m \), as before, the number of finite elements.

The SDP problem (10) can be efficiently solved by modern interior-point polynomial time methods. The question of recovering the optimal elasticity matrices \( E_1^*, \ldots, E_m^* \) from the solution of (10) is rather technical; again we refer the reader to [2].

### 3 Program MOPED

This section presents a collection of tools for conceptual structural design, called MOPED. The underlying model is the *free material optimization* problem introduced above. This problem gives the best physically attainable material and can be considered the "ultimate" generalization of the structural
optimization problem. The method is supported by powerful optimization and numerical techniques, which allow us to work with bodies of complex initial design and with very fine finite-element meshes, giving thus quite accurate solutions even in “difficult” parts and for complex geometries.

3.1 Program Structure

The package MOPED consists of three main programs: the mesh generator DU-MESH, the code MATOPT including the discretization and optimization routines, and the graphical post-processor VIRES.

Mesh generator DU-MESH generates quadrilateral finite element meshes on arbitrary two-dimensional domains consisting of regions and holes. The domain boundary is specified by straight segments and arches.

Program MATOPT consists of three main parts – optimizer, finite element code and linear solver. The kernel is the Penalty/Barrier Multiplier method for minimization of convex functions due to Ben-Tal and Zibulevsky [4].

The finite element code provides the optimization algorithm with the necessary data, like value and gradient of the objective function and the constraints. The code includes three routines that read the mesh data files generated by DU-MESH.

The Penalty/Barrier Multiplier algorithm makes a number of steps and in each step it solves a system of linear equations. This system has always the same structure, is sparse and very large. The sparsity structure is given by the numbering of the finite elements which is done in DU-MESH and is more or less random. Hence, before solving the system, one has to perform a re-ordering of the system matrix in order to solve the system with high efficiency. Moreover, as we are approaching the solution, the condition number of the matrix becomes higher and higher; therefore we have to use a robust method for the solution of the linear system. Based on the experience with interior-point codes, the method of choice was a modern robust version of the Cholesky algorithm combined with a minimum degree ordering algorithm.

3.2 Data Post-Processing

The computation gives us two main informations about the optimal design. The first (rough) one is the stiffness distribution \( \rho = \text{tr}(E^*) \) and the second (fine) one is the optimal material \( E^* \). Rather than curtail our results and interpret them in the usual zero-one sense, we want to utilize the full information obtained in these results, in order to design an attainable advanced material. This obviously depends on the type of the advanced material and on the manufacturing technology. The composite materials lend themselves to a realization of the computed results. We consider a particular class of composite materials manufactured by the so-called tape-layering technology. The tape-layering machines used by our industrial partner build the whole piece by successive adding adhesive tapes including glass or carbon fibres
in a matrix material (epoxy). In the post-processing phase, we plot curves which indicate the way how to layer these tapes; they also show the proposed thickness of the tapes. The graphical post-processing tool is called viRES (VIsualization of REsultS).

Mathematically, the visualization problem reads as follows: We have given points (element mid-points) $x_k \ (k = 1, \ldots, m)$ in the two-dimensional space with values of $\rho_k$ and strain tensors $\epsilon_k$ from which we can compute for each point the principal strain directions $e_{k, I}$ and $e_{k, II}$. The task is to show these two different vector fields in just one picture such that the human being can get an idea of how the stress directions and magnitudes look like. The visualization of fibres (i.e., continuous lines going through the body) is also done by the program. A typical output of viRES is presented in Sect. 3.3.

3.3 Examples

Example 1. This example from our industrial partner shows a cross-section of an aircraft which has to carry huge forces coming from the wings when flying in extreme situations. This problem was originally solved as a truss, by means of the tools of truss topology optimization. The initial layout and the optimal truss, as computed in [9], are shown in Fig. 1. We now solve the

Fig. 1. Example 1 – initial layout, truss solution, distribution of $\rho$ and principal stress lines
example by MOPED using a mesh of 20000 elements. The resulting "density" distribution, as well as the lines of principal stresses are also presented in Fig. 1. The reader sees very nice coincidence with the truss result – the stress lines from the MOPED result follow the bars of the optimal truss.

Example 2. In this example we try to model a road bicycle. This is, of course, purely academic modelling, since bicycles are typical examples of multiple-load problems, where the important loads are three-dimensional (torque). Moreover, this is also a typical example of multicriteria optimization – we not only want the bike to be stiff, it should also be comfortable, aerodynamic, etc. Nevertheless, let us try to design the stiffest bike with respect to a typical loading condition. The distribution of the "density" parameter \( p \) is shown in Fig. 2 which also presents the lines of principal stresses.

4 Robust Material Optimization

Quite often, particularly in the single-load case, the optimum design (material) is very unstable with respect to small incidental forces. This is, in fact, a sign of optimality: the structure is designed such that it can withstand the given loads in the best way. Such a design is, unfortunately, not of much use in practice, where one has to encounter small incidental forces. As an example consider the model problem shown in Fig. 3. The optimum design for the problem data on the left-hand figure (with an isolated force) is a vertical bar, which is, obviously, extremely unstable with respect to any arbitrarily small horizontal load.

Hence we ask: can we modify the formulation of the free material optimization problem such that the optimal solution of the modified problem is robust with respect to small incidental loads? This question was answered by Ben-Tal and Nemirovski [3] in the context of truss topology design and their approach can be directly implemented in our problem. The basic idea of robust design in the sense of [3] is to consider the problem in the multiple-
load formulation and augment the set of given loads by mutually orthogonal small incidental loads.

Unfortunately, in the context of free material optimization, the resulting dimension of the problem (given by the number of load-cases) is very large: basically, it is the number of the finite element nodes times two. It is impossible to solve the associated SDP problem (10) with the current software, even for very coarse discretizations. And here comes the idea of cascading [10]. Note that we have certain freedom in the choice of the incidental loads; there are many orthonormal bases. The idea is to choose that one which includes the “most dangerous” incidental loads and to solve the robust problem only with these dangerous loads, ignoring the others. The “dangerous” loads are the eigenvectors associated with largest eigenvalues of certain eigenproblems. This idea leads to an iterative algorithm where in each step we solve a multiple-load problem with a small number of loads and thus numerically tractable.

Example 3. Consider a model example discretized by $29 \times 9$ mesh with the data depicted in Fig. 4; here the small circles denote nodes of the finite element mesh, black dots stand for nodes with homogeneous Dirichlet boundary conditions and the lines represent the prescribed forces.

As expected, the single-load solution shown in Fig. 4 is extremely unstable with respect to small perturbations; see Table 1, Step 0. The column “robustness” shows the increase of the optimal compliance (computed for the given load) when we applied the worst incidental load. The optimal elasticity matrices in all elements are of the type

$$
\begin{pmatrix}
t & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
$$

and can be, e.g., represented by fibres in the $x_1-$ direction.
After only two steps of the cascading algorithm, we obtain a design shown in Fig. 5, together with the principal strains and stresses associated with the original load. Within a small tolerance of 5 per cent, this design is robust; see Table 1, Step 2.

**References**

Automatic Layout and Labelling of State Diagrams

Petra Mutzel\textsuperscript{1} and Gunnar W. Klau\textsuperscript{2}

\textsuperscript{1} Vienna University of Technology, Algorithms and Data Structures Group, Favoritenstr. 9-11, 1040 Wien, Austria. E-mail: mutzel@apm.tuwien.ac.at
\textsuperscript{2} Max Planck Institute for Computer Science, Algorithms and Complexity Group, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany; currently Vienna University of Technology. E-mail: guwek@mpi-sb.mpg.de

Abstract. We consider the problem of automatically generating readable layouts for state diagrams. Such diagrams appear in the field of automation engineering in the design process of control systems. Our industrial partner, the Siemens AG, realised that due to the complex nature of these diagrams, automatic layout tools lead to a better design and documentation of control systems.

The layout problem turns out to be difficult, since not only a graph drawing problem has to be solved but also an additional labelling problem. In this article we study the combined graph layout and labelling problem and present new results for the two-dimensional compaction problem in graph drawing, the label number maximisation problem and the combined graph labelling problem.

1 Introduction

State diagrams are used for designing and running control systems like, e.g., production controls or robot controls in the area of automation engineering. Figure 1 illustrates a typical hand-drawn state diagram of a control system produced at our industrial partner, the Siemens AG.

A state diagram describes all possible states of a control system and the state transitions. A state can be, e.g., the initial state, an error state or states like “motor is running” and is displayed by a circle which is also called a state node in the diagram. A transition from a state $A$ to state $B$ is displayed as a connection line from state node $A$ to state node $B$ via a transition node. Rectilinear labels which are attached to the state and transition nodes contain further information – in many cases several lines of program code.

The drawing of Fig. 1 is very hard to read and understand, as many lines cross each other, labels often overlap with transition lines and many labels are placed far away from their associated nodes. Because of the complex nature of state diagrams, our industrial partner asked us to provide a tool for automatically generating readable layouts of state diagrams. First, we developed a prototype that used state-of-the-art layout techniques for graph drawing.\textsuperscript{1}

\textsuperscript{1} Note that a state diagram can be seen as a graph in which the nodes correspond to the states and transitions, and the edges correspond to the transition lines.
Figure 2 shows the same state diagram as Fig. 1, now semi-automatically produced by our prototype. We first used a graph layout algorithm and then a method for map labelling. This approach forces some labels to either overlap parts of the drawing or places them far away from their associated nodes.

We realised that this approach – first draw, then label – will not lead to nice layouts for state diagrams. When drawing the graph, the appropriate space for the labels should already be reserved. This leads to combining the
graph drawing and the map labelling problem. So far, in the literature, these problems have been treated separately.

The research aim of this project was to investigate the combined drawing and labelling problem in order to produce readable automatically generated layouts of state diagrams. For drawing the graphs we use the so-called topology-shape-metrics paradigm (see, e.g., [5]). This approach produces nice drawings with only a small number of crossings in which the line segments are mainly orthogonal, i.e., horizontal or vertical. Here, a first step (crossing minimisation problem) determines the topology of the layout so that the number of crossings is small. A second step (bend minimisation problem) fixes the shape of the layout, i.e., essentially the bends and the angles, trying to introduce as few bends in the edges as possible. In a third step (compaction problem), the lengths of the line segments are calculated. Here, the objective is to minimise the total edge length or the area.

In this scenario, it seems natural to investigate the combination of the compaction problem with the labelling problem. We call this problem a graph labelling problem in contrast to a problem from map labelling where the coordinates of the points are fixed in the plane. In Sect. 2 we give the mathematical formulations of all problems we consider in this article.

We first investigated the compaction problem. Formally, the pure two-dimensional compaction problem for orthogonal graph drawing is to determine coordinates for the vertices and bends, so that the total edge length is minimised and the shape of the drawing is preserved. This problem is closely related to problems in VLSI-design and has been shown to be NP-hard. We have found a new graph-theoretical formulation for the compaction problem. In [13], we have characterised the set of feasible solutions for the two-dimensional compaction problem in terms of paths in the so-called constraint graphs in $x$- and $y$-direction which code horizontal and vertical positioning relations. Similar graphs (known as layout graphs) have already been used for one-dimensional compaction in VLSI-design, but this is the first time that a direct connection between these graphs has been established. Given the pair of constraint graphs, the two-dimensional compaction task can be viewed as extending these graphs by new arcs so that certain conditions are satisfied and the total edge length is minimised. We can recognise those instances having only one such extension; for these cases we solve the compaction problem in polynomial time.

Hence, we have transformed the compaction problem, which is of rather geometrical nature, into a graph-theoretical one which can naturally be formulated as an integer linear program. Our computational experiments show that the approach works well in practice. Section 3 deals with the pure compaction problem.

Our idea of using constraint graphs for the compaction problem could be transformed to the map labelling problem. Again, a pair of constraint graphs in $x$- and $y$-direction is linked by a set of additional constraints, thus characterising all feasible solutions of the label number maximisation problem.
Formally, the label number maximisation problem asks for the maximum number of labels from a given set that can be placed in the plane so that each such label is attached to its corresponding point and no two labels overlap. Using our new approach, we can not only express various discrete models (e.g., the four-position model in which one of the label corners is attached to the point) but also the slider labelling model. The slider model allows a continuous movement of a label around its point feature, leading to a significantly higher number of labels that can be placed. To our knowledge, we have presented the first algorithm that computes provably optimal solutions in the slider model. Our experimental results on instances created by a widely used benchmark generator indicated that the new approach is applicable in practice. See Sect. 4 for a detailed description.

Section 5 shows that it is possible to integrate both approaches into one in order to attack the combined compaction and labelling problem. Formally, the graph labelling problem can be stated as follows: Given an orthogonal representation – as produced by algorithms within the topology-shape-metrics paradigm – the task is to generate a labelled orthogonal embedding with minimum total edge length.

Unfortunately, so far we have not been able to transfer our theoretical results for the combined problem into practice. Our first experimental results using a simple prototype are promising, however. For the remaining project time, we will work on integrating our approach into a practical tool, which can be used by our industrial partner to generate nice and readable layouts for state diagrams. In Sect. 6 we summarise our results for the combined compaction and labelling problem.

2 Mathematical Problem Formulation

As mentioned in the introduction, we concentrate on the compaction phase of the topology-shape-metrics paradigm in order to assign lengths and place labels at the same time. In this section, we present precise mathematical formulations – first for the two subproblems, and then for the combined compaction and labelling problem. We assume familiarity with planarity and basic graph theory.

The output of the second phase within the paradigm, orthogonalisation, is an orthogonal representation $H$ which contains the information about the planar topology and the shape of the drawing. We call an orthogonal drawing simple if its number of bends is zero. The shape of a simple drawing is given by the angles inside the faces, i.e., the angles occurring at consecutive edges of a face cycle. Note that the notion of shape induces a partitioning of drawings in equivalence classes. Two orthogonal drawings belong to the same class if one can be obtained from the other by modifying the lengths

\footnote{We can always assume that an orthogonal representation is simple by treating bends as artificial vertices.}
of the horizontal and vertical edge segments without changing the angles formed by them. Formally, for a simple orthogonal drawing, the orthogonal representation \( H \) is a function from the set of faces \( F \) to clockwise ordered lists of tuples \((e_r, a_r)\) where \( e_r \) is an edge, and \( a_r \) is the angle formed with the following edge inside the appropriate face.

In order to compute a drawing for \( H \), we must assign coordinates to the vertices. In this article, we concentrate on \emph{pure orthogonal embeddings}. They are only admissible for 4-planar graphs, i.e., planar graphs whose maximum vertex degree does not exceed four. A pure orthogonal embedding maps vertices and bends to distinct points and edge segments to horizontal or vertical non-crossing line segments of some minimum length connecting the images of their endpoints. As with representations, we call orthogonal embeddings simple if they do not contain bends. A special case are \emph{pure orthogonal grid embeddings} as defined in \cite{17}: Here, vertices and bends must have integer coordinates and the minimum length equals one. Our ideas can however be adapted to other drawing standards such as \emph{Kandinsky-like embeddings} (introduced in \cite{7}) or \emph{orthogonal box embeddings}, used, e.g., by the Giotto algorithm (\cite{18}); subclasses of this standard are the \emph{big node model} from \cite{8} and the \emph{TSS} model from \cite{1}, a related class are the \emph{quasi-orthogonal embeddings} as defined in \cite{11}. We state the pure two-dimensional compaction problem as follows:

\begin{problem}[Compaction problem for orthogonal drawings, \textsc{comp}]
Given an orthogonal representation \( H \), find a pure orthogonal grid embedding \( \Gamma \) for \( H \) of minimum total edge length.
\end{problem}

Patrignani has shown \( \text{NP} \)-hardness of \textsc{comp} in \cite{16}. The second subproblem is the labelling task. Let \( R \) be the set of axis-parallel rectangles in the plane. A \emph{labelling} for a set \( P = \{p_1, \ldots, p_k\} \) of \( k \) points in the plane, a set \( L = \{\lambda_1, \ldots, \lambda_l\} \) of \( l \) labels, two functions \( w, h : L \to \mathbb{Q} \) and a function \( a : L \to P \) is an assignment \( r : L \to R \), so that the following conditions hold:

1. \((L1)\) Rectangle \( r(\lambda) \) has width \( w(\lambda) \) and height \( h(\lambda) \) for every \( \lambda \in L \).
2. \((L2)\) Point \( a(\lambda) \) lies on the boundary of \( r(\lambda) \) for all \( \lambda \in L \).
3. \((L3)\) The open intersection \( r(\lambda) \cap r(\mu) \) is empty for all distinct \( \lambda, \mu \in L \).

Properties \((L1)\) and \((L2)\) make sure that each label is attached correctly to its point feature and drawn with the given size. Property \((L3)\) prohibits overlaps between the labels. We allow, however, that two labels touch each other. An interesting version of the labelling subproblem is its maximisation variant:

\begin{problem}[Label Number Maximisation, \textsc{lnm}]
Given a set \( P = \{p_1, \ldots, p_k\} \) of \( k \) points in the plane, a set \( L = \{\lambda_1, \ldots, \lambda_l\} \) of \( l \) labels, two functions \( w, h : L \to \mathbb{Q} \) and a function \( a : L \to P \), find a labelling for a subset \( L_P \subseteq L \) of largest cardinality.
\end{problem}
Also LNM is NP-hard in most axis-parallel labelling models. We can now formally state the combined compaction and labelling problem. A labelled orthogonal embedding $\Gamma_L$ for an orthogonal representation $H$ of a planar graph $G = (V,E)$, a label set $L$, size functions $w, h : L \rightarrow \mathbb{Q}$ and a function $a : L \rightarrow V$ is a tuple $(\Gamma, r)$ where $\Gamma$ is an orthogonal embedding of $H$ and $r$ is a labelling for $L$ with point set \{r(v) | v \in V\}.

**Problem 3 (Compaction and Labelling, COLA).** Given an orthogonal representation $H$ for a planar graph $G = (V,E)$, a label set $L$ with functions $w, h : L \rightarrow \mathbb{Q}$, and $a : L \rightarrow V$, find a labelled orthogonal embedding for $H$ with minimum total edge length.

Note that in this article we have included proofs only if they are especially important or necessary for understanding our techniques, e.g., constructive proofs building the basis of algorithms.

### 3 Optimal Two-Dimensional Compaction

In this section we present our results for the pure two-dimensional compaction problem COMP. We also introduce the combinatorial framework which will be needed throughout the rest of this article. A detailed discussion of our approach to solving the compaction problem to optimality appeared in [13], a full version of the paper is [10].

So far, in graph drawing only heuristics have been used for compacting orthogonal grid drawings. Tamassia suggested in [17] refining the shape of the drawing into one with rectangular faces by introducing artificial vertices and edges. If all the interior faces are rectangles, COMP can be solved in polynomial time using minimum-cost network flow algorithms. In general, however, the solution is far from the optimal solution for the original graph without the artificial vertices and edges. Another method with better results but the

![Fig. 3. The result of four different compaction algorithms: traditional, longest paths, flow, optimal](image-url)

3 The models are defined by further restrictions of property (L2), see Sect. 4 for details. For the unrestricted version as above, a complexity proof appears in [19].
same drawback is given in [2]. Of course, once a first sketch of the graph has been computed, VLSI-based techniques can be used to further improve the quality of the layout. In one-dimensional compaction, the goal is to minimise the width or height of the layout while preserving the coordinates of the fixed dimension. The one-dimensional version of the compaction problem can be solved in polynomial time using the so-called layout graphs. Figure 3 illustrates the differences in the output of four different compaction methods for a given graph with fixed shape.

But the result of this further improvement is also often not satisfying. Figure 4a shows an orthogonal drawing with total edge length \(2k+5\) which cannot be improved by one-dimensional compaction methods. The reason for this lies in the fact that these methods are based on visibility properties. A two-dimensionally optimal solution is Fig. 4b in which the total edge length is only \(k+6\).

3.1 The Combinatorial Framework

Initial observation leads to the following notion: In an orthogonal drawing, incident edges of same direction can be combined; the emerging sets of consecutive edges form the objects to compact. Let \(\Gamma\) be a simple orthogonal drawing of a graph \(G=(V,E)\). It induces a partition of the set of edges \(E\) into the horizontal set \(E_h\) and the vertical set \(E_v\). A horizontal (respectively vertical) subsegment in \(\Gamma\) is a connected component in \((V,E_h)\) (respectively \((V,E_v)\)). If the component is maximally connected it is also referred to as a segment. Three observations are crucial (also see Fig. 5).

1. Each edge is a subsegment.
2. Each vertex \(v\) belongs to one unique horizontal and one unique vertical segment, denoted by hor\((v)\) and vert\((v)\).
3. The limits of a subsegment \(s\) are given as follows: Let \(v_l,v_r,v_b,\) and \(v_t\) be the leftmost, rightmost, bottommost, and topmost vertices on \(s\). Then \(l(s) = \text{vert}(v_l), r(s) = \text{vert}(v_r), b(s) = \text{hor}(v_b),\) and \(t(s) = \text{hor}(v_t)\).

Let \(S_h\) and \(S_v\) denote the set of horizontal and vertical segments, respectively. The following lemma implies that, in an orthogonal drawing for a graph \(G=(V,E)\), the total number of segments is \(2|V| - |E|\).
\[ S_h = \{s_1, s_2, s_3\} \quad S_v = \{s_4, s_5\} \]

<table>
<thead>
<tr>
<th>i</th>
<th>V(s_i)</th>
<th>E(s_i)</th>
<th>l(s_i)</th>
<th>r(s_i)</th>
<th>b(s_i)</th>
<th>t(s_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{v_1}</td>
<td>{}</td>
<td>s_4</td>
<td>s_4</td>
<td>s_1</td>
<td>s_1</td>
</tr>
<tr>
<td>2</td>
<td>{v_2, v_5}</td>
<td>{(v_2, v_5)}</td>
<td>s_4</td>
<td>s_5</td>
<td>s_2</td>
<td>s_2</td>
</tr>
<tr>
<td>3</td>
<td>{v_3, v_4}</td>
<td>{(v_3, v_4)}</td>
<td>s_4</td>
<td>s_5</td>
<td>s_3</td>
<td>s_3</td>
</tr>
<tr>
<td>4</td>
<td>{v_1, v_2, v_3}</td>
<td>{(v_1, v_2, v_2, v_3)}</td>
<td>s_4</td>
<td>s_4</td>
<td>s_1</td>
<td>s_3</td>
</tr>
<tr>
<td>5</td>
<td>{v_4, v_5}</td>
<td>{(v_4, v_5)}</td>
<td>s_5</td>
<td>s_5</td>
<td>s_2</td>
<td>s_3</td>
</tr>
</tbody>
</table>

Fig. 5. Segments of a simple orthogonal grid drawing and its limits

**Lemma 4.** Let \( \Gamma \) be a simple orthogonal grid drawing of a graph \( G = (V, E_h \cup E_v) \). Then \( |S_h| = |V| - |E_h| \) and \( |S_v| = |V| - |E_v| \).

We provide a necessary and sufficient condition for all feasible solutions of a given instance of the compaction problem. This condition is based on existing paths in the so-called constraint graphs. Nodes in these graphs represent the segments; their weighted arcs characterise relative positioning relations between the segments. We denote by \( c(s_i) \) the coordinate of segment \( s_i \). A weight \( \omega_{ij} \in \mathbb{Q} \) for an arc \((s_i, s_j)\) indicates that the coordinate difference \( c(s_j) - c(s_i) \) must be at least \( \omega_{ij} \). Unlike the layout graphs which are based on visibility properties, the pair of constraint graphs arises from the shape of the drawing.

**Definition 5.** A coordinate assignment for a pair of constraint graphs \((D_x, D_y)\) with \( D_x = (V_x, A_x) \), \( D_y = (V_y, A_y) \) and arc weights \( \omega \in \mathbb{Q}^{A_x \cup A_y} \) is a function \( c : V_x \cup V_y \to \mathbb{Q} \). We say \( c \) respects an arc set \( A \subseteq A_x \cup A_y \) if \( c(v_j) - c(v_i) \geq \omega_{ij} \) for all \((v_i, v_j) \in A\).

We will show later that we can use a pair of constraint graphs together with a coordinate assignment which respects this pair to construct a feasible orthogonal grid embedding — as long as the graphs satisfy certain conditions. The following theorem expresses an important connection between constraint graphs and coordinate assignments.

**Theorem 6.** Let \( D = (V, A) \) be a constraint graph with arc weights \( \omega \in \mathbb{Q}^{|A|} \). A coordinate assignment \( c \) that respects \( A \) exists if and only if \( A \) does not contain a directed cycle of positive weight.

**Proof.** For the forward direction assume otherwise. Without loss of generality, let \( C = ((v_1, v_2), (v_2, v_3), \ldots, (v_{|C|}, v_1)) \) be a directed cycle of positive weight in \( A \). Since \( c \) respects \( A \), it must respect in particular the arcs in \( C \). It follows that \( c(v_2) - c(v_1) \geq \omega_{(1,2)} \), \( c(v_3) - c(v_2) \geq \omega_{(2,3)} \), \ldots, \( c(v_{|C|}) - c(v_{|C|}) \geq \omega_{(|C|,1)} \). Summing up the left sides yields zero, summing up the right sides yields \( \sum_{a \in C} \omega_a \). We get \( 0 \geq \sum_{a \in C} \omega_a > 0 \), a contradiction.
For the backward direction of the proof, let $\mathcal{A}c \geq \omega$ be the set of inequalities describing the requirements for a coordinate assignment. By Farkas' Lemma, there is a feasible solution if and only if no vector $y \geq 0$ with $y^T A = 0$ and $y^T \omega > 0$ exists. Assume otherwise, i.e., there is such a vector $y$. Then $y$ corresponds to a flow in $A$ with positive weight. Since all supplies in the corresponding network are zero, this flow must be circular and thus corresponds to a directed cycle of positive weight.

Figure 6 shows an example of a pair of constraint graphs. The arcs specify exactly the relative relationships known from the shape of the graph: We call such special pairs of constraint graphs $((S_v, A_h), (S_h, A_v))$ shape graphs or a shape description. Note that each horizontal edge in the original graph defines a relative positioning constraint between two vertical segments. Similarly, vertical edges determine constraints between horizontal segments.

The characterisation of feasible solutions for the two-dimensional compaction problem is based on the following three observations:

1. The arcs of the shape graphs are contained in every pair of constraint graphs corresponding to a drawing reflecting the given shape.
2. Generally, the information in the shape graphs is not sufficient to produce an orthogonal embedding. If this is the case, however, we will see that they fulfil a special property which we will refer to as completeness.
3. There are in general many possibilities for extending the shape graphs to a complete pair of constraint graphs.

We denote by $s_i \xrightarrow{+} s_j$ a path of positive weight between $s_i$ and $s_j$. The following is a precise characterisation for complete pairs of constraint graphs in terms of paths that must be contained in the arc sets.

**Definition 7.** A pair of constraint graphs is complete if and only if both arc sets do not contain positive cycles and for every pair of segments $(s_i, s_j)$ $\in$
If one of the positive paths in Def. 7 exists we call the pair of segments \((s_i, s_j)\) separated.\(^4\) We can now express a one-to-one correspondence between these complete extensions and orthogonal embeddings. On the basis of the following theorem, the two-dimensional compaction task can be seen as the search for a complete extension of the given shape graphs leading to minimum total edge length.

**Theorem 8.** For each simple orthogonal embedding with shape description \(\sigma = ((S_v, A_h), (S_h, A_v))\) there exists a complete extension \(\tau = ((S_v, B_h), (S_h, B_v))\) of \(\sigma\) and vice versa.

**Proof (Sketch).** To prove the first part of the theorem, we consider a simple orthogonal grid drawing \(\Gamma\) with shape description \(\sigma = ((S_v, A_h), (S_h, A_v))\). Let \(c(s_i)\) denote the fixed coordinate for segment \(s_i \in S_h \cup S_v\). We construct a complete extension \(\tau = ((S_v, B_h), (S_h, B_v))\) for \(\sigma\) as follows: \(B_h = \{ (s_i, s_j) \in S_v \times S_v \mid c(s_i) < c(s_j) \}\), i.e., we insert an arc from every vertical segment to each vertical segment lying to the right of \(s_i\). Similarly, we construct the set \(B_v\). Clearly, we have \(A_h \subseteq B_h\) and \(A_v \subseteq B_v\). We show the completeness by contradiction: Assume first that there is some pair \((s_i, s_j)\) which is not separated. According to the construction this is only possible if the segments cross in \(\Gamma\), which is a contradiction. Now assume that there is a cycle in one of the arc sets. Again, the construction of \(B_h\) and \(B_v\) prohibits this case. Hence \(\tau\) is a complete extension of \(\sigma\).

We give a constructive proof for the second part of the theorem by specifying a simple orthogonal grid drawing for the complete extension \(\tau\). To accomplish this task we need to assign lengths to the segments. Any topological sorting algorithm, e.g., one based on longest paths computations in the acyclic graphs in \(\tau\), computes an integer coordinate assignment \(c\) as defined in Def. 5 for \(\tau\). Given \(c\), the following simple and straightforward method assigns coordinates to the vertices. Let \(x \in \mathbb{N}^V\) and \(y \in \mathbb{N}^V\) be the coordinate vectors. Then simply setting \(x_v = c(\text{vert}(v))\) and \(y_v = c(\text{hor}(v))\) for every vertex \(v \in V\) results in a correct grid drawing. \(\square\)

We have transformed the compaction problem into a graph-theoretical problem. Our new task is to find a complete extension of a given shape description \(\sigma\) that minimises the total edge length. If a shape description already satisfies the conditions of a complete extension (see Fig. 7a), the compaction

\(^4\) In the special case that all arc weights equal one – as in the problem COMP – the notion of completeness is simpler: Both arc sets must be acyclic and for every segment pair one of the four paths must be present.
Fig. 7. Three types of shape descriptions. Dotted lines show the orthogonal grid drawings, thin arrows arcs in shape descriptions and thick gray arcs possible completions

problem can be solved optimally in polynomial time: The resulting problem is the dual of a minimum cost flow problem. Sometimes the shape description is not complete but it is only possible to extend it in one way (see Fig. 7b). In these cases it is also easy to solve the compaction problem. But in most cases it is not clear how to extend the shape description since there are many different possibilities (see Fig. 7c).

The graph-based characterisation can be used to obtain an integer linear programming formulation for the problem COMP. Since the pure compaction problem is the special case of the combined compaction and labelling problem in which the label set is empty, we defer the presentation of our approach until Sect. 5 and mention at this point some of our experimental results for the pure compaction problem.

3.2 Computational Results

We compare the results of our method – which we will refer to as OPT – to the results achieved by two other compaction methods: ORIG is an implementation of the traditional method proposed in [17]. It divides all the faces of the drawing into sub-faces of rectangular shape and assigns consistent edge lengths in linear time. 1DIM is an optimal one-dimensional compaction algorithm. It first calls ORIG to get an initial drawing and then runs iteratively a visibility-based compaction, alternating the direction in each phase. It stops if no further one-dimensional improvement is possible.

The algorithms ORIG, 1DIM, and OPT have been tested on a large test set. This set, first mentioned in [6], contains more than 11,000 graphs representing data from real-world applications. For our experimental study, we planarise each of the graphs by computing a planar subgraph and reinserting the edges, representing each crossing by a virtual vertex. After fixing the planar embedding for every graph we compute its shape using an extension
of Tamassia's bend minimising algorithm, presented in [11]. The resulting orthogonal representation is the input for the three compaction algorithms. We compare the total edge length and the area of the smallest enclosing rectangle of the drawings produced by ORIG, 1DIM, and OPT. Furthermore, we record their running times.

All the examples can be solved to optimality on a SUN Enterprise 10000. For all instances, the running times of ORIG and 1DIM are below 0.05 and 0.43 seconds, respectively. OPT solves the vast majority of instances (95.5%) in less than one second, few graphs need more than five seconds (1.1%) and only 29 graphs need more than one minute. The longest running time, however, is 68 minutes.

The average improvement of the total edge length computed by OPT over the whole test set of graphs is 2.4% compared to 1DIM and 21.0% compared to ORIG. Just looking at hard instances where OPT needs more than two seconds of running time we yield average improvements of 7.5% and 36.1%, respectively. Figures 8 and 9 show this fact in more detail: The x-axis denotes the size of the graphs, the y-axis shows the improvement of total edge length in percent with respect to 1DIM and ORIG. We compute the minimal, maximal and average improvement for the graphs of the same size. The average improvement values are quite independent from the graph size, and the minimum and maximum values converge to them with increasing number of vertices. Note that in some cases OPT yields improvements of more than

![Improvement in %](image)

**Fig. 8.** Quality of OPT compared to 1DIM
30% in comparison to the previously best strategy; Fig. 3 at the beginning of this section shows the drawings for such a case (here, the improvement is 28%). For the area, the data look similar with the restriction that in a few cases the values produced by 1DIM are slightly better than those from OPT: short edge length does not necessarily lead to small area. The average area improvements compared to 1DIM and ORIG are 2.7% and 29.3%, however.

In general, we make the following observations: Instances of COMP divide into easy and hard problems, depending on the structure of the corresponding graphs. On the one hand, we are able to solve some randomly generated instances of biconnected planar graphs with 1,000 vertices in less than five seconds. In these cases, however, the improvement compared to the results computed by 1DIM is small. On the other hand, graphs containing tree-like structures are hard to compact since their number of fundamentally different drawings is in general very high, but in these cases the improvement is much greater.

4 Optimal Labelling in the Slider Model

In this section we investigate the label number maximisation problem LNM as defined in Sect. 2. Like the pure compaction problem, this problem is interesting in its own right and has direct applications, e.g., in cartogra-
Many papers have been published on the label number maximisation problem (for an overview, see the bibliography [21]). So far, most previous work on map labelling has concentrated on the discrete model, which allows only a finite number of positions per label. The most popular discrete model is the four-position model (see Fig. 10); the two- and one-position models have been introduced rather for theoretical purposes. In [4], Christensen, Marks and Shieber have presented a comprehensive treatment of LNM in the four-position model including complexity analysis, heuristic methods and a computational study. They have introduced a procedure for randomly creating labelling instances which has become a widely used benchmark generator in the map labelling literature. The only practically efficient algorithm for computing provably optimal solutions in the discrete model has been suggested by Verweij and Aardal in [20]. They treat the problem as an independent set problem and solve it using a branch-and-cut algorithm. The algorithm is able to optimally label up to 800 point features (using the benchmark generator from [4]) within moderate computation time (about 20 minutes).

More natural than the discrete model is the slider model, which allows a continuous movement of a label around its point feature. Although Hirsch already considered this model in 1982, it was not further investigated until very recently. In [19], van Kreveld, Strijk and Wolff introduce several variations of the slider model (see Fig. 10). They prove NP-hardness of LNM in the four-slider model and suggest a polynomial time algorithm which is able to find a solution that is at least half as good as an optimal solution. Moreover, their computational results show that the slider model is significantly better than the discrete model. The four-slider model allows the placement of up to 15% more labels in real-world instances and up to 92% more labels in pseudo-random instances.

In this section, we will present an algorithm for solving instances of LNM to optimality that works in any of the above mentioned labelling models. We

![Fig. 10. Axis-parallel rectangular labelling models. A label can be placed in any of the positions indicated by the rectangles and slid in the directions of the arcs](image-url)
allow several labels per point feature and labels of different sizes. Figure 11 shows a provably optimal labelling for 700 point features computed with a first implementation of our new approach; 699 labels could be placed. Like the algorithm for the compaction problem, it is based on the pair of constraint graphs defined in Sect. 3. In the following, we show how to treat labels inside our combinatorial framework.

4.1 Extending the Combinatorial Framework

Modelling point features. The positions of the $k$ point features are specified in the input set $P$. For each $p_i \in P$ with coordinates $x(p_i)$ and $y(p_i)$ we introduce a node $x_i$ in $V_x$ and a node $y_i$ in $V_y$; one for its $x$-, one for its $y$-coordinate. We fix the positions of the point features by inserting four directed paths $P_x = (x_1, \ldots, x_k)$, $P_{-x} = (x_k, \ldots, x_1)$, $P_y = (y_1, \ldots, y_k)$ and $P_{-y} =$
(y_k, \ldots, y_1) with weights \( \omega_{x_i, x_{i+1}} = x(p_{i+1}) - x(p_i), \omega_{x_{i+1}, x_i} = x(p_i) - x(p_{i+1}), \omega_{y_i, y_{i+1}} = y(p_{i+1}) - y(p_i) \) and \( \omega_{y_{i+1}, y_i} = y(p_i) - y(p_{i+1}) \) for \( i \in \{1, \ldots, k-1\} \).

We call the directed edges on these paths fixed distance arcs and refer to them as \( A_F \). Figure 12 shows a set of point features and its representation in the constraint graphs.

**Lemma 9.** A coordinate assignment \( c \) that respects \( A_F \) results in a correct placement of point features (up to translation).

**Modelling labels.** Each label \( \lambda \in L \) has to be represented by a rectangle \( r(\lambda) \) of width \( w(\lambda) \) and height \( h(\lambda) \). Additionally, we have to ensure that \( \lambda \) will be placed correctly with respect to its point feature \( a(\lambda) \), i.e., \( a(\lambda) \) must lie on the boundary of \( r(\lambda) \).

Straightforwardly, we model a label \( \lambda \) by two nodes in \( V_x \) and two nodes in \( V_y \), representing the coordinates of \( r(\lambda) \). Following the terminology in Sect. 3, we call these nodes the left, right, bottom and top limit of \( \lambda \) and refer to them as \( l_\lambda, r_\lambda, b_\lambda \) and \( t_\lambda \), respectively. We introduce four *label size arcs* \( A_S(\lambda) = \{(l_\lambda, r_\lambda), (r_\lambda, l_\lambda), (b_\lambda, t_\lambda), (t_\lambda, b_\lambda)\} \) in order to model the size of \( r(\lambda) \). The weights of these arcs are \( \omega_{l_\lambda, r_\lambda} = w(\lambda), \omega_{r_\lambda, l_\lambda} = -w(\lambda), \omega_{b_\lambda, t_\lambda} = h(\lambda) \) and \( \omega_{t_\lambda, b_\lambda} = -h(\lambda) \), see Fig. 13a.

A label \( \lambda \) must be placed close to its point feature \( a(\lambda) \). Let \( x \) and \( y \) be the nodes representing point \( a(\lambda) \) in the constraint graphs. We add four *proximity arcs* \( A_P(\lambda) = \{(x, r_\lambda), (l_\lambda, x), (y, t_\lambda), (b_\lambda, y)\} \), as illustrated in Fig. 13b. These arcs have zero weight and exclude the point feature \( a(\lambda) \) from lying outside the rectangle \( r(\lambda) \).

The point feature may still lie inside \( r(\lambda) \). We disallow this by adding at least one of the four *boundary arcs* \( \{(r_\lambda, x), (x, l_\lambda), (t_\lambda, y), (y, b_\lambda)\} \), each of weight zero. Note that these arcs are inverse to the proximity arcs for label \( \lambda \). If, e.g., \( (r_\lambda, x) \) is present in \( D_x \), it forces – together with its inverse proximity arc \( (x, r_\lambda) \) – the coordinate of the right side of \( r(\lambda) \) to be equal to the coordinate of \( x \); the label has to be placed at its leftmost position. See also Fig. 13c.

![Fig. 13. Modelling labels](image)
At this point we can influence the labelling model. We define a set $A_B(\lambda)$ containing the boundary arcs for the different models:

<table>
<thead>
<tr>
<th>Labelling model</th>
<th>Boundary arcs $A_B(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slider model</td>
<td>Four-slider: $(r_\lambda, x), (x, l_\lambda), (t_\lambda, y), (y, b_\lambda)$</td>
</tr>
<tr>
<td></td>
<td>Two-slider: $(t_\lambda, y), (y, b_\lambda)$</td>
</tr>
<tr>
<td></td>
<td>One-slider: $(y, b_\lambda)$</td>
</tr>
<tr>
<td>Discrete model</td>
<td>Four-position: $(r_\lambda, x), (x, l_\lambda), (t_\lambda, y), (y, b_\lambda)$</td>
</tr>
<tr>
<td></td>
<td>Two-position: $(y, b_\lambda), (t_\lambda, x), (x, l_\lambda)$</td>
</tr>
<tr>
<td></td>
<td>One-position: $(y, b_\lambda), (x, l_\lambda)$</td>
</tr>
</tbody>
</table>

For a slider model, at least one arc $a \in A_B(\lambda)$ must be contained in the constraint graph, for a discrete model at least two. E.g., for the four-slider model, the set $A_B(\lambda)$ consists of all four boundary arcs, one of which must be present in the appropriate constraint graph. Note that we can express all six axis-parallel rectangular labelling models we have introduced at the beginning of this section as additional requirements on the constraint graphs.

**Lemma 10.** Let $\lambda$ be a label, $c$ be a coordinate assignment respecting $A_S(\lambda)$, $A_P(\lambda)$ and at least $d$ boundary arcs from $A_B(\lambda)$. Then $c$ results in a placement in which $\lambda$ is represented by a rectangle $r(\lambda)$ of width $w(\lambda)$ and height $h(\lambda)$. The label is placed so that point feature $a(\lambda)$ lies on the boundary of $r(\lambda)$ if $d = 1$ and on a corner of $r(\lambda)$ if $d = 2$.

**Avoiding label overlaps.** To this point we have ensured that each label is placed correctly with respect to its point feature. We must still guarantee that the intersection of rectangles is empty. A crucial observation is that it suffices to consider only the pairs of labels that can possibly interact. If there is any overlap, such a pair must be involved.

Consider two different labels $\lambda$ and $\mu$ and their corresponding rectangles $r(\lambda)$ and $r(\mu)$. We call the pair vertically separated if $r(\lambda)$ is placed either above or below $r(\mu)$. Similarly, $\lambda$ and $\mu$ are horizontally separated if one rectangle is placed left to the other. Two labels overlap if they are neither vertically nor horizontally separated, we can exclude this by introducing one of the following four label separation arcs $A_S(\lambda, \mu) = \{(t_\mu, b_\lambda), (t_\lambda, b_\mu), (r_\mu, l_\lambda), (r_\lambda, l_\mu)\}$. Label separation arcs have zero weight.

Let $R_\lambda$ be the boundary of the region in which label $\lambda$ can be placed. Note that $R_\lambda$ is defined by lower left corner $(x(a(\lambda)) - w(\lambda), y(a(\lambda)) - h(\lambda))$ and upper right corner $(x(a(\lambda)) + w(\lambda), y(a(\lambda)) + h(\lambda))$. Likewise, we determine $R_\mu$ for label $\mu$. If the intersection of $R_\lambda$ and $R_\mu$ is empty, $\lambda$ and $\mu$ can never overlap, and we do not have to add any label separation arcs for this pair. In this case we set $A_S(\lambda, \mu) = \{\}$.

Consider now the case in which the intersection of $R_\lambda$ and $R_\mu$ is not empty, as depicted in Fig. 14. Depending on the position of the corresponding point features $a(\lambda)$ and $a(\mu)$, $A_S(\lambda, \mu)$ contains the following label separation arcs:
Fig. 14. Label separation arcs between two labels $\lambda$ and $\mu$

1. If $x(a(\mu)) \geq x(a(\lambda))$ we have $(r_{\lambda}, l_{\mu}) \in A_S(\lambda, \mu)$.
2. If $x(a(\lambda)) \geq x(a(\mu))$ we have $(r_{\mu}, l_{\lambda}) \in A_S(\lambda, \mu)$.
3. If $y(a(\mu)) \geq y(a(\lambda))$ we have $(t_{\lambda}, b_{\mu}) \in A_S(\lambda, \mu)$.
4. If $y(a(\lambda)) \geq y(a(\mu))$ we have $(t_{\mu}, b_{\lambda}) \in A_S(\lambda, \mu)$.

Note that the only case in which $A_S(\lambda, \mu)$ contains all four label separation arcs occurs if $\lambda$ and $\mu$ label the same point feature, i.e., $a(\lambda) = a(\mu)$.

**Lemma 11.** Let $\lambda$ and $\mu$ be two labels that can possibly overlap and let $c$ be a coordinate assignment respecting $A_S(\lambda)$, $A_S(\mu)$ and $A \subseteq A_S(\lambda, \mu)$ with $|A| \geq 1$. Then $c$ results in a placement in which the two rectangles $r(\lambda)$ and $r(\mu)$ do not overlap.

We refer to the boundary and label separation arcs as potential arcs

$$A_{pot} = \bigcup_{\lambda \in L} A_B(\lambda) \cup \bigcup_{\lambda, \mu \in L, \lambda \neq \mu} A_S(\lambda, \mu)$$

and state the label number maximisation problem in a combinatorial way. The task is to choose additional arcs from $A_{pot}$ for a maximum number of labels without creating positive directed cycles.

**Problem 12 (Constraint Graph Fulfilment problem, CGF).** Given an instance of LNM, let $(D_x, D_y)$ be the pair of constraint graphs including only fixed distance arcs, label size arcs and proximity arcs. Let $d = 1$ if in the slider model and $d = 2$ if in the discrete model and let $A_x$ and $A_y$ be the arc sets of $D_x$ and $D_y$, respectively. Find a set $L_P \subseteq L$ of greatest cardinality and an arc set $A \subseteq A_{pot}$ with the following properties:

(F1) $|A \cap A_B(\lambda)| \geq d$ for all $\lambda \in L_P$.
(F2) $|A \cap A_S(\lambda, \mu)| \geq 1$ for all $\lambda, \mu \in L_P$, $\lambda \neq \mu$, $R_{\lambda} \cap R_{\mu} \neq \emptyset$.
(F3) $A \cap A_x$ and $A \cap A_y$ do not contain a positive cycle.

**Theorem 13.** Problems LNM and CGF are polynomially equivalent.
Proof. Let \( A \) be a solution of \( \text{CGF} \). We extend \((D_x, D_y)\) by adding \( A \) to the arc sets. Because of (F3) and Thm. 6, there is a coordinate assignment that respects both the horizontal and vertical arc set. Lemmas 9, 10 and 11 ensure that properties (L1), (L2) and (L3) are fulfilled, thus we have a solution for \( \text{LNM} \).

For the other direction, we start with the given coordinate assignment \( c \) resulting from the placement of labels. We create the set \( A \) of additional arcs as follows: For each label \( \lambda \) we add one or two boundary arcs, depending on how \( r(\lambda) \) is placed with respect to point feature \( a(\lambda) \). Similarly, we add appropriate arcs from \( A(\lambda, \mu) \) for pairs of labels \( \lambda \) and \( \mu \), depending on the relative position of \( \lambda \) and \( \mu \) in the labelling. Note that we have chosen the additional arcs so that they are respected by \( c \). Properties (F1) and (F2) follow by construction, property (F3) follows by Thm. 6. \( \Box \)

We propose a zero-one integer linear programming formulation for solving \( \text{CGF} \). The goal is to find the set of additional boundary and label separation arcs \( A \) and to determine which labels are to be placed.

We introduce two types of binary variables for this task: For each label \( \lambda \) there is a variable \( y_\lambda \in \{0,1\} \), indicating whether \( \lambda \) will be placed \((y_\lambda = 1)\) or not \((y_\lambda = 0)\). Additionally, there is a variable \( x_a \in \{0,1\} \) for each potential additional arc \( a \in A_{\text{pot}} \) which is one if \( a \) is active and zero otherwise. Recall that \( d \) determines the labelling model – we have \( d = 1 \) in the slider and \( d = 2 \) in the discrete model.

We present the zero-one integer linear program (1) and show that it corresponds to \( \text{CGF} \). We define \( C_p = A_{\text{pot}} \cap C \).

\[
\text{max} \quad \sum_{\lambda \in L} y_\lambda \\
\text{subject to} \quad \sum_{a \in A_B(\lambda)} x_a - dy_\lambda \geq 0 \quad \forall \lambda \in L \quad (1.1)
\]

\[
\sum_{a \in A_S(\lambda, \mu)} x_a - y_\lambda - y_\mu \geq -1 \quad \forall \lambda, \mu \in L, \lambda \neq \mu \quad (1.2)
\]

\[
\sum_{a \in C_p} x_a \leq |C_p| - 1 \quad \forall \text{ positive cycles } C \quad (1.3)
\]

\[
y_\lambda \in \{0,1\} \quad \forall \lambda \in L \quad (1.4)
\]

\[
x_a \in \{0,1\} \quad \forall a \in A_{\text{pot}} \quad (1.5)
\]

We refer to (1.1) as boundary inequalities, to (1.2) as label separation inequalities and to (1.3) as positive cycle inequalities.

**Theorem 14.** Each feasible solution \((y,x)\) to (1) corresponds to a feasible solution of \( \text{CGF} \) and vice versa. The value of the objective function equals the cardinality of \( L_P \).
Corollary 15. An optimal solution of (1) corresponds to an optimal labelling.

Corollary 15 and Thm. 6 (the existence of a coordinate assignment) suggest an algorithm for attacking practical instances of the label number maximisation problem: In a first step, we solve (1) using integer programming techniques. The solution tells us which boundary and label separation arcs should be added to the arc sets of the constraint graphs \((D_x, D_y)\). We use this information in a second step for computing the corresponding coordinate assignment via minimum cost flow (see proof of Thm. 6).

Our implementation is based on LEDA ([15]) and the ILP-solver in CPLEX ([9]). Due to the fact that there may be an exponential number of positive cycles in the constraint graphs, we use an iterative cutting plane approach. Our separation procedure uses the Bellman–Ford algorithm for detecting negative cycles applied to the constraint graphs after multiplying the arc weights \(w \in \mathbb{Q}^{A_x \cup A_y}\) by \(-1\). Our implementation is based on the one given in [3].

We have tested our algorithm on a widely used benchmark generator for randomly creating instances of LNM, following the rules described in [4]: First, we construct a set of \(n\) points with random coordinates in the range \([0, \ldots, 792]\) for the \(x\)- and \([0, \ldots, 612]\) for the \(y\)-coordinates. To each point feature belongs a label of width 30 and height 7. We have run the algorithm with both the slider and the discrete labelling models for rectilinear map labelling. Figure 11 in Sect. 1 shows an optimal solution for an instance with \(n = 700\). In the optimal solution for the four-slider model, 699 labels can be placed, whereas at most 691 can be placed in the four-position model.

Evidently, more freedom in the labelling model results in a higher number of labels that can be placed. Two main factors influence the running time of our implementation for instances of the same size: On the one hand, this is the number of labels that cannot be placed in an optimal solution, i.e., the difference \(|L| - |L_P|\): To our surprise, we had the longest running times in the one-position model. On the other hand, the tightness of the inequalities seems to have an impact on the running time; the more restrictions on the variables, the faster the algorithms. Both factors, however, interrelate: In the more restricted models we can also place fewer labels.

5 Solving the Combined Problem

In this section, we combine the results from Sects. 3 and 4 and express the combined compaction and labelling problem COLA as a combinatorial optimisation problem in the pair of constraint graphs. More details can be found in [12]. Furthermore, we present an integer linear program and show that each feasible solution corresponds to a solution for the COLA problem.

We have to extend the notion of shape description and completeness in order to integrate the labels: A labelled shape description \(\sigma_L\) of an orthogonal
representation $H$ with label information $L, w, h, a$ is a tuple $((S_v \cup V_{Lv}, A_h \cup A_{Lh}), (S_h \cup V_{Lh}, A_v \cup A_{Lh}))$ where $((S_v, A_h), (S_h, A_v))$ is a shape description for $H$ and

$$
V_{Lv} = \bigcup_{\lambda \in L} \{l_\lambda, r_\lambda\}, \quad A_{Lh} = \bigcup_{\lambda \in L} (A_{S_h}(\lambda) \cup A_{P_h}(\lambda)),
$$

$$
V_{Lh} = \bigcup_{\lambda \in L} \{b_\lambda, t_\lambda\}, \quad A_{Lv} = \bigcup_{\lambda \in L} (A_{S_v}(\lambda) \cup A_{P_v}(\lambda)).
$$

Here, $A_{S_h}(\lambda)$ and $A_{S_v}(\lambda)$ are the label size arcs and $A_{P_h}(\lambda)$ and $A_{P_v}(\lambda)$ correspond to the proximity arcs, both defined in Sect. 4, i.e.,

$$
A_{S_h}(\lambda) = \{(l_\lambda, r_\lambda), (r_\lambda, l_\lambda)\} \quad A_{S_v} = \{(b_\lambda, t_\lambda), (t_\lambda, b_\lambda)\}
$$

$$
A_{P_h}(\lambda) = \{(l_\lambda, r_a(\lambda)), (r_a(\lambda), l_\lambda)\} \quad A_{P_v} = \{(b_\lambda, t_a(\lambda)), (b_a(\lambda), t_\lambda)\}.
$$

As in the pure labelling problem, the label sizes determine the weights of arcs in $A_{S_h}$ and $A_{S_v}$, the proximity arcs have zero weight. Observe that each instance of the COLA problem uniquely determines a labelled shape description. We can now give a more general formulation of completeness by considering not only the segments but also the labels. This generalisation leads to the subsequent main theorem whose proof is similar to the proof of Thm. 8.

**Definition 16.** A pair of labelled constraint graphs is complete if and only if the arc sets do not contain a directed cycle of positive weight and each distinct pair $o, p \in S_h \cup S_v \cup L, o \neq p$ is separated.

**Theorem 17.** For each simple labelled orthogonal embedding with shape description $\sigma = ((S_v, A_h), (S_h, A_v))$ and label information $L, w, h, a$ a complete labelled extension $\tau = ((S_v \cup V_{Lv}, A_h \cup A_{Lh}), (S_h \cup V_{Lh}, A_v \cup A_{Lh}))$ of $\sigma$ exists and vice versa.

### 5.1 Integer Linear Programming Formulation

We use this extended combinatorial framework to obtain an integer linear program formulation for the combined compaction and labelling problem COLA. Let an instance of the problem be characterised by the 4-planar graph $G = (V, E)$, a simple orthogonal representation $H$ which divides $E$ into $E_h$ and $E_v$, and the label information $L, w, h$ and $a$. This determines the labelled shape description $\sigma_L = ((S_v \cup V_{Lv}, A_h \cup A_{Lh}), (S_h \cup V_{Lh}, A_v \cup A_{Lh}))$. Recall that the node set of this pair of constraint graphs is given by the set of segments $S = S_h \cup S_v$ and additional nodes $V_{Lv} \cup V_{Lh}$ which model the limits of the labels. The arc sets consist of the arcs in the shape description for $H$ and the size and proximity arcs for the labels.

We consider the set of potential additional arcs $A_{pot}$ that might be in some complete extension $\tau_L$ of $\sigma_L$. For a distinct pair $o, p \in S \cup L, o \neq p$, let
$A_{\text{sep}}(o, p)$ be the arcs between the limits of $o$ and $p$ which ensure separation of the two objects (cf. Def. 7), so we have

$$A_{\text{pot}} = \bigcup_{o \neq p \in S \cup L} A_{\text{sep}}(o, p).$$

As shown for the pure labelling problem and sketched for the pure compaction problem, a set $A_{\text{sep}}(o, p)$ rarely must have cardinality four, in most cases it can be empty. We introduce a variable $x_a \in \{0, 1\}$ for each arc $a \in A_{\text{pot}}$ which is one if $a$ is contained in the extension and zero otherwise. Additionally, we have a variable vector $c \in \mathbb{Q}^{|S \cup V_h \cup V_L|}$ to model the coordinate assignment for the pair of constraint graphs. Clearly, if we know $c_v$ for all nodes $v$ in the constraint graphs we can produce a labelled drawing of the graph with the given label information. Our integer linear programming formulation for the COLA problem is as follows ($M$ is a very large number):

$$\begin{align*}
\text{min} & \quad \sum_{e \in E_h} (c_{r(e)} - c_{l(e)}) + \sum_{e \in E_v} (c_{t(e)} - c_{b(e)}) \\
\text{subject to} & \quad \sum_{a \in A_{\text{sep}}(o, p)} x_a \geq 1 \quad \forall o \neq p \in S \cup L \quad (2.1) \\
& \quad c_j - c_i \geq \omega_a \quad \forall a = (i, j) \in A \quad (2.2) \\
& \quad c_j - c_i - (M + \omega_a)x_a \geq -M \quad \forall a = (i, j) \in A_{\text{pot}} \quad (2.3) \\
& \quad x_a \in \{0, 1\} \quad \forall a = (i, j) \in A_{\text{pot}} \quad (2.4)
\end{align*}$$

Inequalities (2.1) model the characterisation of separation, i.e., the existence of necessary paths between distinct objects in an extension. Inequalities (2.2) force the coordinates to obey the distance rules coded by the weighted arcs in the underlying labelled shape description; the value $\omega_{ij}$ denotes the minimum distance between $s_i$ and $s_j$. The same must hold true for the potential additional arcs: Whenever a variable $x_a$ is one, i.e., the potential arc $a$ is part of the extension, we want an inequality of type (2.2), otherwise there should be no restriction on the coordinate variables. This situation is modelled by inequalities (2.3) with the help of a large constant $M$. We have shown in [13] that inequalities (2.2) and (2.3) ensure the absence of positive cycles in the extension.

Note that we do not introduce additional decision variables to integrate the labelling task; the choice of where to place a label is performed by the separation properties. A label can be placed wherever it does not interfere with other objects as long as it stays in the neighbourhood of the vertex to which it belongs. Moreover – as in the pure labelling problem – we do not restrict the label placement to a finite number of prescribed places. Of course, the number of possibly non-separated objects is much higher than in an instance of a pure compaction problem. Observe also that the boundary arcs, that serve in the pure labelling problem to force the boundary of a label...
to touch the appropriate point feature are a special case of the separation arcs.

Like the one-to-one correspondence between complete extensions and labelled orthogonal embeddings there is a one-to-one correspondence between feasible solutions of the integer linear program and complete extensions of the given labelled shape graphs:

**Theorem 18.** For each feasible solution \((x, c)\) of the integer linear program for a given labelled shape description \(\sigma_L\), there is a labelled orthogonal embedding with appropriate shape and label information and vice versa.

**Proof (Sketch).** A solution of the integer linear program corresponds to a complete extension of the labelled shape description \(\sigma\) with appropriate length assignment. The forward direction follows with Thm. 8. Conversely, we can use the information of a labelled embedding to construct a feasible solution of the integer linear program. We can show that none of the inequalities is violated and that the value of the objective function equals the total edge length.

### 6 Conclusions

In order to devise an algorithm that automatically draws and labels state diagrams as they occur, e.g., in automation engineering, we have studied the NP-hard combined compaction and labelling problem. This problem lies in the intersection of two research areas: graph drawing and map labelling. The task is to simultaneously assign consistent edge lengths for a given orthogonal representation which determines the shape of the drawing and to label subsets of vertices. The resulting drawing should have small total edge length and all labels should be placed at the corresponding vertices so that they do not overlap other objects. Already the two subproblems – the pure compaction task and the pure labelling task – are NP-hard.

We have introduced combinatorial formulations for the two-dimensional compaction problem and the maximisation variant of the labelling problem. Both subproblems have themselves interesting direct applications. We have given integer linear programming formulations for the combinatorial versions of the problems. Our experimental results show that this approach works well in practice – we can optimally compact graphs that have as many as 1,000 vertices and optimally label up to 700 point features in moderate computation time.

We have shown how to combine the combinatorial formulations of the two subproblems in order to model the combined compaction and labelling problem. We have presented an integer linear program to solve it, and preliminary computational experiments with instances from our industrial partner indicate that our approach can be successful in practice.
References


Optimization Problems in a Semi-Automatic Device for Cutting Leather

Alexander Pott\textsuperscript{1} and Holger Glaab\textsuperscript{2}

\textsuperscript{1} Institut für Algebra und Geometrie, Otto-von-Guericke-Universität Magdeburg, 39016 Magdeburg, Germany
\textsuperscript{2} Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany

Abstract. Cutting large leather skins is a two step process, a “decision step” and a “cutting step”. In the decision step, it has to be decided which piece is cut from which part of the skin. This is called nesting. After the nesting process the actual cutting takes place. The cooperation company HUMANTEC GmbH in Wemding (Germany, Bavaria) has developed a machine where the nesting is done by a person and not by a computer. The machine just assists the worker and helps her to find a good nesting. Only the cutting process is basically fully automatic. In both parts, the nesting and the cutting process, mathematical problems related to the travelling salesman problem occur.

1 Introduction

In this introductory section we describe a semi-automatic machine called Lasernest. This machine is designed to cut large leather skins. It supports the two processes nesting and cutting that occur when leather skins are cut. Nesting: There is a list of pieces which have to be cut off the leather skin, for instance the different pieces that are necessary for the production of a leather sofa. Leather is a natural product: Not every part of the skin has the same quality. There are sometimes differences in the colour or some parts of the skin are damaged due to injuries (for instance bites) of the animal. Regions of less quality may have been created also during the preparation of the leather. As a consequence, it is important to decide which piece is actually cut from which part of the skin.

There are also some quality requirements for the pieces that are cut from the skin. For instance, some parts of a leather sofa are unvisible or less exposed than others. These less important pieces or parts of pieces may be cut off from regions which have minor quality.

The final goal in this nesting process is to minimize waste of leather. However, one cannot simply use algorithms to minimize this quantity since we have to take into account the uniqueness of each skin, hence the different regions of quality as described above.

In practice, two different approaches have been suggested to overcome this difficulty. One can partition the skin into regions of different quality and one does the same for the pieces. The quality requirements can be translated
into mathematical constraints that a feasible nesting (or packing) has to satisfy. With these new constraints, classical algorithms to optimize packings can be adopted. However, the judgement about the quality of each skin and partitioning it into different quality regions is still done by humans.

The approach of our cooperation partner HUMANTEC GmbH and their machine Lasernest is different. A skilled worker may find a good packing of the pieces in time compared to a computer if she gets help from a machine. Note that even an automatic nesting device needs human expertise to judge the quality. This intermediate step is not necessary if the nesting is done directly by human workers.

How can a machine assist the workers during the nesting process? The answer is that the workers always see the contours of the pieces on the leather skin. The easiest way to accomplish this is to use templates. Unfortunately, this is rather time consuming. Lasernest uses the following concept. Basically, the nesting is done on a computer screen. You can click on the pieces and move them around as you know it from any geometry software. However, it is also necessary to project the image on the leather skin. Then the workers can easily decide whether the quality requirements are met.

In Lasernest, this projection is realized by several (typically between 4 and 8) laser rays. The rays are steered by two mirrors (also called scanner pairs) corresponding to two orthogonal directions. It is important to note that moving the mirrors takes time. Now the main point is the following: In order to obtain a steady and flicker free image, it has to be drawn about 25 times a second. Therefore, one should try to find paths for the laser rays which are small. At first view there is nothing which can be optimized since every piece in a cutting image has to be drawn. The answer is that it also takes time if the laser moves from one piece to another (so called idle paths). The mirrors have to be moved although the laser is dimmed. Moreover, drawing acute angles is more time consuming than drawing obtuse ones. This is due to the fact that the laser rays are directed by the scanner pairs; moving the pairs to draw an acute angle takes more time.

These two facts (minimize idle paths and avoid acute angles) are the main reasons why it is possible to speak about “good” paths for the laser rays. We call the problem to find paths which are minimal with respect to scanner movements the Lasernest multiple scanner problem (m-LP) where \( m \) denotes the number of scanner pairs (i.e. Laser rays).

**Cutting:** The second process in manufacturing leather skins is the cutting of the skin. Similar to \( m \)-LP, one should try to find a “good” path for the cutting knife. In this situation it is not really necessary to find a short path. It is much more important to get sharp borders rather than fuzzy ones. The quality of the boundary of the pieces, after cutting, depends partly on the path of the knife. For instance it is sometimes necessary to cut from outside the leather into the skin. This should preferably be done orthogonal to the boundary of the skin. If the angle between the path of the knife and the boundary of the skin is too small it is impossible to obtain sharp edges.
Moreover, it is helpful if small pieces are cut first. We abbreviate the problem to find a good path for the cutter by CP.

The two problems $m$-LP and CP are related but there are also several differences:

- Both problems are variants of a travelling salesman problem (TSP). In case of $m$-LP we have $m$ salesmen.
- Problem $m$-LP has to be solved “online”. As soon as the worker moves a piece, the laser path has to be re-optimized. In practice, one of the lasers draws the piece that is inserted since this piece will be moved steadily. It is also possible to move pieces that have been inserted already however that does not happen very often. In contrast, when the nesting process is finished, the optimal cutter path has to be computed only once.
- It takes about 10 minutes for a skilled worker to construct a good nesting. The cutting takes only about one minute. Lasernest has two cutting tables A and B. When the pieces are nested on table A, the skin on table B can be cut. This arrangement is the reason that there is sufficient time (about 1 min) to solve the cutter problem.

It was our goal in the project to provide the cooperation partner with fast algorithms to solve the laser problem $m$-LP and the cutter problem CP. We modelled $m$-LP as an integer linear programming problem (ILP). It turns out that this problem is very hard. Even if we could solve it exactly or close to optimality, it would take by far too much time to be of any use for our cooperation partner. Therefore, we used the ILP formulation basically only to find good lower bounds by relaxing the integer restrictions. Good solutions have been constructed by heuristics. We will explain the ILP formulation in Sect. 2. Section 3 contains a description of our heuristics. Finally, we describe in Sect. 4 our model of the cutter problem and our approach to solve it. Since in this case the time constraints are less severe, integer linear programming techniques might be useful to find good solutions.

We hope that the description of the two problems was detailed enough so that the reader can understand the following models. For $m$-LP, we also refer to [5].

2 The Lasernest Problem with $m$ Scanners

We consider a cutting image consisting of $S$ pieces. We assume the existence of $m$ scanners, i.e. $m$ independent laser rays. For each laser, we have to find a path such that each piece is drawn or, in other words visited by at least one laser. For simplicity we assume that each piece is visited by precisely one laser.

We have to define the objective function carefully. We do not simply want to minimize the length of each laser path. The goal is that the length of each
of the paths is small in order to get a flickerfree image. Therefore, the length of the smallest path should be small.

At first view, the pieces are bounded by closed curves and therefore there is no candidate for a start- and endpoint on the curves. As we pointed out in the introduction, the sum of the lengths of the paths between the pieces should be small. In order to describe the distance (which is not the Euclidean one) we have to know a start- and endpoint where the laser starts drawing the piece and where it finishes. We also mentioned that it is useful to avoid the drawing of acute angles. These two observations motivate to split the border of each piece into several small paths. These are usually straight lines or other simple polygons. We assume that the boundaries of the S pieces split into N curves which are not closed, each of it having a well defined start- and endpoint. We can think of these as the vertices V of a complete directed graph G. There are two possible ways to traverse a vertex i, namely from start- to endpoint or vice versa. Let us call these two directions $i$ and $\bar{i}$. So somehow we double the set of vertices. For the quality of the drawing of the cutting image it is unimportant whether $i$ or $\bar{i}$ is drawn. Therefore, only one of the two vertices $i$ and $\bar{i}$ has to be “visited”.

There are four arcs associated with $i,j$, see Fig. 1. If the laser draws the piece $i$ and directly after that the piece $j$, the ray has to be directed from the endpoint of $i$ to the startpoint of $j$. This arc is denoted by $(i,j)$. We call the length of this arc $c_{ij}$. Analogously, we define $c_{ij}, c_{i\bar{j}}$ and $c_{i\bar{j}}$. We have $c_{ij} = c_{ji}$ and $c_{ij} = c_{ji}$. We emphasize that the distance is not a metric and it is not the Euclidean distance. For instance, $c_{ij}$ measures the time that is takes to travel from the endpoint of $i$ to the endpoint of $j$. We have to include the time that it takes to direct the laser ray on the idle path. This time depends basically on the angle between the piece and the idle path. The reader may imagine that this time is different for the arcs $(i,j)$ and $(\bar{i},\bar{j})$ in Fig. 1. The cost function $c_{ij}$ is calculated by a module of our cooperation partner.

Drawing the piece $i$ also takes time, say $d_i$. Again, $d_i$ will be calculated by a programme of HUMANTEC. In TSP language, $d_i$ means that travelling through a city is also time consuming (a realistic assumption). Since we have

![Fig. 1. Illustration of the four arcs corresponding to $(i,j)$](image-url)
to draw each piece exactly once, this amounts to a total time $\sum d_i$. If we had only one (sub)tour (as in the usual TSP) we could forget this constant term. We could also forget about it if our goal would be to minimize the sum of the lengths of the $m$ laser tours. However, we want to minimize the maximum length of a tour and therefore we have to take the $d_i$ into account. We do this by defining a modified distance function $\tilde{c}_{ij} = c_{ij} + \frac{d_i + d_j}{2}$, similarly $\tilde{c}_{ij}, \tilde{c}_{ij}$ and $\tilde{c}_{ij}$.

In order to define $m$-LP as an integer linear programming problem, we need one more bit of notation. Let $S_1, \ldots, S_{m+1}$ be a partition of the vertices $V$ of our graph $G$. We define

$$A(S_1, \ldots, S_m) := \{(i, j) \in A : i \in S_k, j \in S_l, 1 \leq k < l \leq m\}$$

and call it a directed $m$-cut. We describe $m$ laser tours by variables $x_{ij}^k$, $k = 1, \ldots, m$, where

$$x_{ij}^k = \begin{cases} 1 & (i, j) \text{ is an arc in the } k\text{'th laser tour,} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we define $x_{ij}, x_{ij}$ and $x_{ij}$. In order to represent all components belonging to an $m$-cut, we write

$$x^k(A(S_1, \ldots, S_m)) = \sum_{ij \in A(S_1, \ldots, S_m)} (x_{ij}^k + x_{ij}^k + x_{ij}^k + x_{ij}^k)$$

We obtain the following formulation of $m$-LP:

$$\begin{align*}
\text{min} & \quad w \\
\text{s.t.} & \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{j \neq i} (\tilde{c}_{ij} x_{ij}^k + \tilde{c}_{ij} x_{ij}^k + \tilde{c}_{ij} x_{ij}^k + \tilde{c}_{ij} x_{ij}^k) - w \leq 0, \quad k = 1, \ldots, m \quad (1) \\
& \quad \sum_{k=1}^{m} \sum_{j=1}^{N} (x_{ij}^k + x_{ij}^k + x_{ij}^k + x_{ij}^k) = 1, \quad i = 1, \ldots, N \quad (2) \\
& \quad \sum_{i=1}^{N} (x_{ij}^k + x_{ij}^k) - \sum_{l=1}^{N} (x_{jl}^k + x_{jl}^k) = 0, \quad j = 1, \ldots, \bar{N}, N, \bar{N} \quad (3) \\
& \quad \sum_{k=1}^{m} x^k(A(S_1, \ldots, S_{m+1})) \geq 1, \quad (S_1, \ldots, S_{m+1}) \in P_{m+1}(V) \quad (4) \\
& \quad \sum_{i,j \in V} (x_{ij}^k + x_{ij}^k + x_{ij}^k + x_{ij}^k) \geq 2, \quad k = 1, \ldots, m \quad (5) \\
& \quad x_{ij}, x_{ij}, x_{ij}, x_{ij} \in \{0, 1\}, \quad k = 1, \ldots, m, ij \in A. \quad (6)
\end{align*}$$
In order to obtain a linear objective function we introduced an artificial variable \( w \) and a class of artificial constraints (1). Equations (2) and (3) ensure that each vertex has exactly one successor and one predecessor. Moreover, (3) guarantees, that each piece is assigned to exactly one of the laser tours and that each piece has a unique passing direction. Together with (2) the last condition also implies that any two laser tours are vertex disjoint. The so-called shorter subtour elimination constraints (SSEC) (4) exclude the existence of more than \( p \) different laser tours. Finally, the inequalities (5) ensure that each circuit consists of at least two arcs. Additionally, (5) in connection with (6) ensure that each feasible solution consists of exactly \( m \) laser tours.

It seems rather difficult to investigate the corresponding polytope. However, such an investigation is the basis of any branch&bound or branch&cut algorithm. In other words, it seems hopeless to solve this problem by standard ILP techniques, in particular in view of the condition that \( m \)-LP should be solved online. Therefore, we used the ILP formulation only to find good lower bounds for the optimum solution.

In connection with \( m \)-LP and its ILP formulation, we studied the (easier) Hamiltonian \( m \)-median problem and its corresponding polytope. This polytope is the convex hull of all the incidence vectors of \( m \) directed cycles \( \{C_1, \ldots, C_m\} \) whose vertex sets form a partition of the vertices of the graph \( G \). We obtained several results about this polytope, see [3]. It is already quite involved to determine the dimension of it.

Another lower bound can be easily constructed from a graph theoretic relaxation. We constructed a new matroid (\( m \)-forests) such that each solution of \( m \)-LP is in the matroid. Therefore, the optimum solution of the optimization problem on the matroid gives a lower bound on the optimum value of the original \( m \)-LP. Because of the special structure of our objective function, the value \( c(F^*)/m \) is a lower bound for our optimum value where \( F^* \) is the optimum \( m \)-forest. It should be noted that \( m \)-forests are an interesting relaxation for a vehicle routing problem without depots. We refer the reader to [4] for the definition and investigation of the matroid.

Finally, we used an assignment relaxation to find good lower bounds.

3 Heuristics for Solving \( m \)-LP

We describe our main heuristics that produce good solutions. In our experiments with real world data, the quotient between lower bound and best solution is about .95. This empirical data indicates that the heuristics described below are sufficiently good for practical purposes. Therefore one can say that one of our problems, namely to provide the cooperation partner with a fast and good algorithm for solving \( m \)-LP, has been solved.

We have to distinguish opening heuristics and improving steps. We implemented the following three heuristics to find a good solution to start with:
1. greedy;
2. restricted greedy;
3. COMBCRIT.

**Greedy:** We start with the empty graph $E$. We simply add edges to $E$ (starting with an arc of lowest cost) until we have obtained at most $m$ cycles. Then it may be necessary to split cycles into smaller ones in order to obtain $m$ cycles. Obviously, this procedure does not take into account that we are interested in "balanced" solutions where the length of the longest cycle is as small as possible.

**Restricted Greedy:** Using elementary facts about the statistical distribution of the arc lengths we give an estimate about the number $d_{\text{min}}$ of vertices that we can expect in each cycle. Then we do not allow an arc to close a path to a cycle if the number of vertices in the cycle is smaller than $d_{\text{min}}$.

**COMBCRIT:** We choose $m$ vertex disjoint subtours each consisting of two arcs. These $m$ subtours are the starting configurations to construct $m$ vertex disjoint cycles. The $m$ arc pairs are in different cycles and they remain in different ones during the run of COMBCRIT. Therefore, they should be chosen from regions which are far away from each other. During COMBCRIT, vertices are assigned to the $m$ laser subtours. The decision where the node is inserted is based on two considerations. One point is that it should be inserted in a path which is still short. Another criteria is that the length should not increase too much upon insertion. We made several test runs to get some intuition and advice how to mix these two criteria: It turned out to be better if we put more emphasize on the second one.

We are now going to describe two improving heuristics:

**3-opt:** This is not a simple adaptation of the standard 3-opt improving strategy in the classical TSP, see [7]. In our setting we choose two arcs from the cycle $T_{\text{max}}$ of maximum length and one arc from the cycle of smallest length $T_{\text{min}}$. Then we choose a best recombination of the pieces after we removed the arcs. The two possible types of recombinations are illustrated in Fig. 2.

![Fig. 2. Possible recombinations in 3-opt by exchanging the arc set \{(u_1,v_1), (u_2,v_2), (u_3,v_3)\}](image)
It is possible that in one step of 3-opt, the length of $T_{\text{max}}$ does not decrease, however the total length of the $m$ laser tours does. In this case (which occurs rather often) one should change the tour although the value of our objective function does not improve. Usually, in the steps after such a degenerate one, more improvements occur.

2-opt-splitting: This approach consists of three different steps. In a first step we join the two laser tours having maximum and minimum cost to a larger laser tour by removing and adding two arcs. Then we apply several

<table>
<thead>
<tr>
<th>$N$</th>
<th>RC3O</th>
<th>RC2O</th>
<th>GR3O</th>
<th>GR2O</th>
<th>RG3O</th>
<th>RG2O</th>
<th>BLB</th>
<th>$g^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>58</td>
<td>53*</td>
<td>53*</td>
<td>53*</td>
<td>53*</td>
<td>53*</td>
<td>47</td>
<td>12.77</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
<td>46*</td>
<td>46*</td>
<td>46*</td>
<td>46*</td>
<td>46*</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>109</td>
<td>99*</td>
<td>111</td>
<td>101*</td>
<td>111</td>
<td>101*</td>
<td>97</td>
<td>2.06</td>
</tr>
<tr>
<td>25</td>
<td>78</td>
<td>76</td>
<td>76</td>
<td>73</td>
<td>76</td>
<td>73*</td>
<td>69</td>
<td>5.80</td>
</tr>
<tr>
<td>25</td>
<td>94</td>
<td>91*</td>
<td>97</td>
<td>92</td>
<td>101</td>
<td>91*</td>
<td>86</td>
<td>5.81</td>
</tr>
<tr>
<td>30</td>
<td>149</td>
<td>139*</td>
<td>150</td>
<td>145</td>
<td>143</td>
<td>149</td>
<td>136</td>
<td>5.15</td>
</tr>
<tr>
<td>34</td>
<td>127</td>
<td>121*</td>
<td>134</td>
<td>122</td>
<td>133</td>
<td>125</td>
<td>115</td>
<td>6.09</td>
</tr>
<tr>
<td>33</td>
<td>179</td>
<td>163*</td>
<td>167</td>
<td>161*</td>
<td>174</td>
<td>164</td>
<td>154</td>
<td>5.55</td>
</tr>
<tr>
<td>35</td>
<td>182</td>
<td>180</td>
<td>191</td>
<td>178*</td>
<td>186</td>
<td>183</td>
<td>171</td>
<td>4.09</td>
</tr>
<tr>
<td>38</td>
<td>264</td>
<td>253*</td>
<td>262</td>
<td>253*</td>
<td>265</td>
<td>253*</td>
<td>242</td>
<td>4.55</td>
</tr>
<tr>
<td>42</td>
<td>104</td>
<td>97*</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>98</td>
<td>89</td>
<td>6.74</td>
</tr>
<tr>
<td>41</td>
<td>197</td>
<td>195</td>
<td>201</td>
<td>194</td>
<td>199</td>
<td>191*</td>
<td>188</td>
<td>1.60</td>
</tr>
<tr>
<td>49</td>
<td>140</td>
<td>136</td>
<td>144</td>
<td>134*</td>
<td>140</td>
<td>135</td>
<td>128</td>
<td>4.69</td>
</tr>
<tr>
<td>46</td>
<td>191</td>
<td>183</td>
<td>202</td>
<td>186*</td>
<td>194</td>
<td>186*</td>
<td>175</td>
<td>6.29</td>
</tr>
<tr>
<td>54</td>
<td>224</td>
<td>218*</td>
<td>226</td>
<td>220</td>
<td>226</td>
<td>220</td>
<td>213</td>
<td>2.82</td>
</tr>
<tr>
<td>58</td>
<td>300</td>
<td>281*</td>
<td>296</td>
<td>284</td>
<td>301</td>
<td>284</td>
<td>276</td>
<td>2.90</td>
</tr>
<tr>
<td>62</td>
<td>237</td>
<td>229*</td>
<td>238</td>
<td>230</td>
<td>240</td>
<td>229*</td>
<td>218</td>
<td>5.05</td>
</tr>
<tr>
<td>68</td>
<td>279</td>
<td>265*</td>
<td>275</td>
<td>267</td>
<td>277</td>
<td>267</td>
<td>258</td>
<td>2.69</td>
</tr>
<tr>
<td>72</td>
<td>209</td>
<td>293*</td>
<td>314</td>
<td>291*</td>
<td>309</td>
<td>292</td>
<td>282</td>
<td>3.19</td>
</tr>
<tr>
<td>79</td>
<td>250</td>
<td>243*</td>
<td>258</td>
<td>245</td>
<td>261</td>
<td>248</td>
<td>235</td>
<td>4.26</td>
</tr>
<tr>
<td>79</td>
<td>276</td>
<td>266*</td>
<td>288</td>
<td>272</td>
<td>285</td>
<td>275</td>
<td>262</td>
<td>3.82</td>
</tr>
<tr>
<td>71</td>
<td>293</td>
<td>284</td>
<td>302</td>
<td>283*</td>
<td>296</td>
<td>287</td>
<td>277</td>
<td>2.17</td>
</tr>
<tr>
<td>85</td>
<td>299</td>
<td>290*</td>
<td>301</td>
<td>294</td>
<td>303</td>
<td>293</td>
<td>279</td>
<td>5.02</td>
</tr>
<tr>
<td>86</td>
<td>281</td>
<td>272</td>
<td>293</td>
<td>269*</td>
<td>286</td>
<td>275</td>
<td>261</td>
<td>3.07</td>
</tr>
<tr>
<td>91</td>
<td>295</td>
<td>283</td>
<td>296</td>
<td>282*</td>
<td>300</td>
<td>282*</td>
<td>273</td>
<td>3.30</td>
</tr>
<tr>
<td>85</td>
<td>324</td>
<td>314*</td>
<td>324</td>
<td>316</td>
<td>321</td>
<td>312*</td>
<td>298</td>
<td>4.70</td>
</tr>
<tr>
<td>100</td>
<td>381</td>
<td>367*</td>
<td>397</td>
<td>374</td>
<td>388</td>
<td>372</td>
<td>362</td>
<td>2.76</td>
</tr>
<tr>
<td>99</td>
<td>297</td>
<td>290*</td>
<td>304</td>
<td>290</td>
<td>298</td>
<td>290*</td>
<td>278</td>
<td>4.32</td>
</tr>
<tr>
<td>111</td>
<td>378</td>
<td>370*</td>
<td>389</td>
<td>374</td>
<td>392</td>
<td>377</td>
<td>362</td>
<td>2.49</td>
</tr>
<tr>
<td>117</td>
<td>336</td>
<td>323*</td>
<td>345</td>
<td>323*</td>
<td>333</td>
<td>325</td>
<td>312</td>
<td>3.53</td>
</tr>
<tr>
<td>135</td>
<td>430</td>
<td>405*</td>
<td>422</td>
<td>406</td>
<td>416</td>
<td>404</td>
<td>393</td>
<td>2.80</td>
</tr>
<tr>
<td>150</td>
<td>475</td>
<td>457</td>
<td>476</td>
<td>455</td>
<td>476</td>
<td>455*</td>
<td>442</td>
<td>2.94</td>
</tr>
<tr>
<td>118</td>
<td>560</td>
<td>543</td>
<td>561</td>
<td>544</td>
<td>554</td>
<td>542*</td>
<td>529</td>
<td>2.64</td>
</tr>
<tr>
<td>151</td>
<td>433</td>
<td>413*</td>
<td>445</td>
<td>417</td>
<td>441</td>
<td>421</td>
<td>407</td>
<td>2.46</td>
</tr>
<tr>
<td>153</td>
<td>487</td>
<td>464*</td>
<td>490</td>
<td>464*</td>
<td>491</td>
<td>465</td>
<td>446</td>
<td>3.80</td>
</tr>
<tr>
<td>157</td>
<td>863</td>
<td>842*</td>
<td>864</td>
<td>846</td>
<td>855</td>
<td>843</td>
<td>827</td>
<td>1.45</td>
</tr>
<tr>
<td>163</td>
<td>901</td>
<td>876</td>
<td>897</td>
<td>876</td>
<td>897</td>
<td>874*</td>
<td>866</td>
<td>0.92</td>
</tr>
<tr>
<td>160</td>
<td>805</td>
<td>785</td>
<td>807</td>
<td>784*</td>
<td>801</td>
<td>785</td>
<td>770</td>
<td>1.82</td>
</tr>
<tr>
<td>174</td>
<td>1015</td>
<td>986*</td>
<td>1011</td>
<td>988*</td>
<td>1002</td>
<td>984*</td>
<td>971</td>
<td>1.34</td>
</tr>
<tr>
<td>176</td>
<td>880</td>
<td>848*</td>
<td>884</td>
<td>851</td>
<td>869</td>
<td>856</td>
<td>836</td>
<td>1.79</td>
</tr>
<tr>
<td>212</td>
<td>904</td>
<td>869*</td>
<td>889</td>
<td>870</td>
<td>895</td>
<td>872</td>
<td>857</td>
<td>1.52</td>
</tr>
<tr>
<td>204</td>
<td>1137</td>
<td>1111</td>
<td>1139</td>
<td>1109*</td>
<td>1137</td>
<td>1109*</td>
<td>1090</td>
<td>1.65</td>
</tr>
</tbody>
</table>

Table 1. Empirical comparison of heuristic solutions and the best lower bound (BLB) for 43 cutting images from practice: RC=COMBCRIT with random start routes, GR=greedy, RG=restricted greedy, 2O=2-opt, 3O=3-opt, BUB=best upper bound (signed by *), $g^* = (BUB - BLB)/BUB \times 100$, $N$ number of curves to be drawn
classical 2-opt moves to this new laser tour in order to reduce its cost. In the final step we split it into two smaller ones such that the cost maximum of the two resulting laser tours is as small as possible.

The advantage of this combined method is obvious: By using the structure of two laser tours we have more knowledge and more exchange combinations compared to 2-opt-moves in a single laser tour. A further advantage is that each iteration of 2-opt-splitting has a complexity of $O(N^2)$ whereas each step of 3-opt requires $O(N^3)$ time.

In Table 1 we compare the empirical performance of several combined heuristics. Combined heuristics means that we have constructed a feasible start solution in a first step (opening). This solution has been improved by local search methods in a second step.

Our data indicates that all combined heuristics using 2-opt-splitting as a local search method yield almost the same good results. Since the opening heuristics COMBCRIT is by far the fastest of the three candidates we recommend to use COMBCRIT for the online construction of the laser tours and to reoptimize the cutting image from time to time by the 2-opt-splitting algorithm.

4 The Cutting Problem

We consider a completely nested cutting image. Additionally to cutting out the pieces, certain places of the leather skin have to be marked and other places of the skin have to be punched out (punched holes). In order to perform the different tasks the cutter has specific tools. Depending on the cutter type which is used, Lasernest offers the following features:

1. The tools can be changed during the process such that cutting, marking and punching out can be done in an arbitrary order.
2. The different jobs have to be performed consecutively, i.e. one job has to be completed before the tool is changed and the next job starts.

We will model the more complex first case. Our modelling can easily be adopted to the second case.

Similarly to the $m$-LP we split the boundaries of of our pieces in curves and denote the set of all curves by $V_C$. Note that each curve has a start- and endpoint and therefore two different traversing directions. Since as a rule the places to be marked are also straight lines we also model them by two traversing directions $i$ and $i$ and denote the associated vertex set by $V_M$. In contrast to that, the punched holes are modelled as zero-dimensional objects without a direction. We denote this vertex set by $V_A$. Finally we have a depot node 0 where the cutter starts and finishes its cutter tour. If we permit all possible connections between curves, marks and punched holes we obtain a complete digraph $D = (V, A)$ with node set $V := V_A \cup V_C \cup V_M \cup \{0\}$ and
arc set

\[ A := \{(i,j) : i,j \in V \} \cup \{(i,j), (j,i) : i \in V, j \in V_C \cup V_M \} \]
\[ \cup \{(i,j) : i,j \in V_C \cup V_M \}. \]

Note that we have four arcs connecting vertices in \( V_C \cup V_M \), similar to the modelling of \( m-\)LP.

Moreover, we define a function \( t : A \rightarrow \mathbb{R}_0^+ \) which measures the time to move the cutter from the endpoint of vertex \( i \) to the startpoint of vertex \( j \) and to perform the task associated to vertex \( j \). If \( i \) and \( j \) belong to different node types, i.e. the tool has to be changed, additional reset costs have to be added. More precisely, if \( i \in V_C \) and \( j \in V_M \) then \( t_{ij} \) measures the time to travel from the endpoint of curve \( i \) to the endpoint of mark \( j \) and to change the tool and to mark the corresponding object. In the other cases we define \( t_{ij}, t_{ij}, t_{ij}, t_{ij} \) similarly.

Our optimization goal is to minimize the total processing time of the cutter. Hereby we have to consider certain restrictions:

**Precedence Clustering Constraints:** Due to the special mechanics of the cutter portal the cutting board is divided into several regions which are processed subsequently. Since these regions are given by the cutter system it might be necessary to split some curves and to enlarge the node set \( V \). For reasons of simplicity we will assume that no curves have to be split. We define a clustering \( (C_1, \ldots, C_g), g \in \mathbb{N} \) of \( V \{0\} \) with pairwise disjoint sets \( C_1, \ldots, C_g \) and \( \bigcup_{i=1}^g C_i = V \) where each vertex, except the depot, is uniquely assigned to a cluster \( C_i \). The clusters \( C_i \) are processed in an increasing order, i.e. the cutter starts in \( C_1 \).

As already mentioned in the introduction, it is helpful if the smaller pieces are cut first: Small jets below the skin produce low pressure to fix and smooth the skin. In order to obtain a good cutting it is necessary to keep this vacuum effect as long as possible. As soon as some pieces are cut out air infiltrates the skin and the vacuum effect decreases. Therefore, it seems reasonable to group the curves in a second clustering step according to their sizes. This second clustering is only applied to the set of curves \( V_C \) and not to the punch nodes and not to the marks. The impact of the latter ones on the vacuum effect can be neglected. We define on each cluster \( C_i \) a second clustering \( C_{i,1}, \ldots, C_{i,n_i} \) with pairwise disjoint subclusters \( C_{i,j} \) and \( \bigcup_{j=1}^{n_i} C_{i,j} = C_i \cap V_C \). Each curve is assigned to exactly one subcluster. If \( j < h \) holds we want that all the curves contained in \( C_{i,j} \) are cut out before the first curve in \( C_{i,h} \) is processed.

**Tangent structure precedence constraints** The quality of the cutting process mainly depends on sharp edges. This means that we have to avoid paths of the cutter knife which from “tangent-like” structures. Such a constellation can occur if an oval boundary of piece, represented by two curves \( v \) and \( w \) is (almost) touched by a straight line, represented by a curve \( u \). Moreover the angle between \( u \) and \( v \) respectively \( u \) and \( w \) has to be small enough. We see this situation in Fig. 3.
In the constellation of 3, we will probably obtain fuzzy edges in the area around $T$ and $T'$ if the curves $v$ and $w$ are cut out in the same passing direction. There are two solutions for this problem, we only consider solution (a) where $u$ is a straight line that can be traversed in both directions. We want that $u$ is cut before $v$ and $w$, i.e. we impose two precedence relations $TS u \prec v$ and $TS u \prec w$.

Additionally, we fix the traversing directions of curve $v$ and curve $w$ by $v$ and $w$. This means that all arcs incident to node $w$ and node $v$ become infeasible. Let $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ denote the Euclidean distance in $\mathbb{R}^2$, let $\alpha_{TS}$ denote the critical angle and let $\epsilon_{TS}$ denote the critical distance. We represent a tangent structure by a node triple $(u, v, w)$ and define by

$$TS := \{ (u, v, w) \in V \times V \times V : \min_{p \in u} d(p, v \cap w) \leq \epsilon_{TS}, \angle(u, v) \leq \alpha_{TS} \}$$

the set of tangent structures. The tangent structures can be identified algorithmically in a preprocessing step. This can be implemented with complexity $O(|V_C| \log |V_C|)$.

Since the precedence constraints resulting from clustering and tangent structures have to be fulfilled by each feasible solution of our cutter problem, we can exclude several infeasible arcs of $A$. We delete these infeasible arcs from $A$ again in a (second) preprocessing step. We denote by $D_0 := (V, A_0)$ the resulting feasible digraph for CP. We can refer to each solution of CP as a cutter tour if the cutter has to return to the depot node 0 after it has fulfilled all tasks. Let $x \in \{0, 1\}^{|A_0|}$ be the incidence vector of a cutter tour defined by

$$x_a = \begin{cases} 1 & \text{if } a \text{ is an arc in the cutter tour,} \\ 0 & \text{otherwise.} \end{cases}$$
Moreover, we define by
\[
W := \{v, w \mid \text{there exists an } u \in V \text{ with } (u, v, w) \in TS\}
\]
the set of curves where the passing direction is determined by a tangent structure constraint. Moreover,
\[
B_0 = \{ij : \{ij, \overline{ij}, \overline{ji}, \overline{ij}\} \cap A_0 \neq \emptyset\}
\]
is the aggregated arc set of the multidigraph $D_0$ which we obtain when we ignore the passing directions. Consequently, the (up to four) arcs $(i, j), (\overline{i}, j), (i, \overline{j})$ and $(\overline{i}, \overline{j})$ are replaced by only one arc $(i, j)$. Similarly, we define by
\[
y \in \{0, 1\}^{|B_0|}, y_{ij} := \sum_{a \in A_0, a^+ \in \{i, \overline{i}\}, a^- \in \{j, \overline{j}\}} x_a \text{ for } ij \in B_0
\]
the aggregated incidence vector.

In order to formulate the CP as an integer linear program we need a little bit more notation. For $j \in V$ we define
\[
x(\delta^- (j)) := \sum_{ij \in A_0} x_{ij} + \sum_{ij \in A_0} x_{ji} \quad \text{and} \quad x(\delta^+ (j)) := \sum_{ji \in A_0} x_{ji} + \sum_{ji \in A_0} x_{j\overline{i}}.
\]
Let $S_1, S_2$ be subsets of $V$. Then we define $y(S_1, S_2) := \sum_{ij \in B_0(S_1, S_2)} y_{ij}$.

We obtain the following ILP-formulation for CP:
\[\min \quad t_C := \sum_{ij \in A_0} t_{ij} x_{ij}\]
s. th. \[\sum_{ij \in B_0} y_{ij} = 1, \quad j \in V\] (7)
\[\sum_{ij \in B_0} y_{ij} = 1i \in V\] (8)
\[x(\delta^- (j)) - x(\delta^+ (j)) = 0, j \in (V_C \setminus W) \cup V_M\] (9)
\[y(B_0(S)) \leq |S| - 1\] (10)
\[y(\{v, w\}, Q) + y(B_0(Q)) + y(Q, \{u\}) \leq |Q|\] (11)
\[y(\{v\}, S) + y(B_0(S)) + y(S, \{u\}) \leq |S|\] (12)
\[x_a \in \{0, 1\}, a \in A_0.\] (13)

The equations (7) and (8) ensure that each node has exactly one predecessor and one successor in the cutter tour. The equations (9) impose the uniqueness of the traversing direction for each node $v \in (V_C \setminus W) \cup V_M$. Since the return arcs from clusters of less priority to clusters of higher priority are already eliminated, it suffices to impose the subtour elimination constraints (10) only on the node subsets $S \subset C_k, |S| \geq 2$, of a cluster.
The tangent structure precedence constraints (11) have to be satisfied for all tangent structures \((u,v,w)\) where \(Q \subset V \setminus \{u,v,w\}\) and \(|Q| \geq 1\). They guarantee that the curve \(u\) is processed before the curves \(v\) and \(w\) are cut off. If not, then there would exist a directed path from node \(v\) respectively \(w\) to node \(u\). If we define \(Q\) as the set of intermediate nodes on this directed path we have \(y(\{v,w\},Q) + y(B_0(Q)) + y(Q,\{u\}) = |Q| + 1\), a contradiction to (11).

The subcluster precedence constraints (12) have to be satisfied for all \(u \in C_{k,l-1}, v \in C_{k,l}\) with \(1 \leq k \leq g, 2 \leq l \leq n\) and \(S \subset C_{k} \setminus \{u,v\}, |S| \geq 1\). These restrictions ensure that all the nodes of a subcluster of higher priority are processed before the first node of the next subcluster is processed.

Finally, since \(A_0(C_l,C_k) = \emptyset\) holds for \(g \geq l > k \geq 1\) the cluster subtour elimination constraints (10) imply that \(y(C_k,C_l) = 1\) also holds for \(1 \leq k < l \leq g\).

Hence, the CP is a variant of a precedence-constrained asymmetrical travelling salesman problem with additional constraints (9). We can show that under certain conditions the CP can be decomposed into \(g\) asymmetric Hamiltonian path problems with specified start node and unspecified end node. In this case we have to find a shortest (with respect to \(t\)) Hamiltonian path \(P_k\) in each cluster \(C_k\) where the start node of \(P_k\) is the end node of \(P_{k-1}\) in the previously visited cluster \(C_{k-1}\).

In order to use integer linear programming techniques we investigated the corresponding polytope. It turned out that already the dimension of the polytope can only be determined under certain assumptions (\(|C_k| \geq 5\) for all \(k, C_k \cap (V_M \cup V_A) \neq \emptyset\)). Nevertheless, we could apply the polyhedral knowledge of the precedence constrained ATSP using ABACUS, see [2]. Based upon our polyhedral investigations we developed a branch&cut algorithm [1] which has the following main features:

**Preprocessing:** We eliminate all infeasible arcs and identify all tangent structures which impose precedence constraints. Moreover, we define a subclustering according to sizes and assign each curve to a subcluster.

**Initialization:** We calculate initial upper bounds by insertion and greedy-like heuristics. The best achieved objective is taken as initial upper bound.

**Lower bounds:** In the initial LP we only add the equations (7)–(9). The associated solution from the initial lower bound. In further iterations cutting planes are generated and added to the current LP in order to improve the best lower bound. Moreover, we exploit the knowledge of each LP-solution by calling LP-based heuristics in order to improve the current best upper bound.

**Cutting Planes:** We have developed separation algorithms for several classes of inequalities. Most of them are known inequalities from the precedence constrained ATSP-polytope [2]. We also use the tangent structure precedence constraints as cutting planes which can be separated in \(O(|V|^{4})\) time by determining a minimum capacity \((v',0)\)-cut in the support digraph \((v'\text{ is a pseudonode})\) of the current LP-solution.
5 Summary

In our project we tried to solve two problems (m-LP and CP). The first of these problems had to be solved very fast (basically online), therefore only heuristic solution approaches could be used. We were able to design very fast and good algorithms and tested their quality empirically. We also found good heuristics for CP. In this case, there is more time to find good solutions, therefore we decided to use more sophisticated arguments to find them. The basic skeleton of the algorithm is already developed but not yet implemented and tested.

References

Stochastic Programming for Power Production and Trading Under Uncertainty

Rüdiger Schultz\(^2\), Matthias P. Nowak\(^2\), Robert Nürnberg\(^1\), Werner Römisch\(^1\), and Markus Westphalen\(^2\)

\(^1\) Humboldt-University Berlin, Inst. of Mathematics, 10099 Berlin, Germany, E-mail: romisch@mathematik.hu-berlin.de
\(^2\) Gerhard-Mercator-University Duisburg, Dept. of Mathematics, 47048 Duisburg, Germany, E-mail: schultz@math.uni-duisburg.de

Abstract. Optimization models under uncertainty for mid-term cost-optimal operation of a hydro-thermal power system and for simultaneous power production and day-ahead trading at a power exchange are presented. Algorithms for solving these models are sketched and initial numerical experience is reported.

1 Introduction

Uncertainty is of increased importance in today’s power management. Traditional sources of uncertainty, such as weather conditions infecting the load curve, are accompanied by novel sources that are due to the liberalization of power markets. Trading now becomes an issue, and new market instruments arise that, to more or less extent, involve some level of uncertainty.

In the present paper we report on two mathematical optimization models that are tailored to coping with randomness in power production and power trading. The paper grew out of a cooperation with the VEAG Vereinigte Energiewerke AG Berlin, a utility running a hydro-thermal power system in the Eastern part of Germany. Our mathematical instruments stem from (mainly two-stage) stochastic integer programming.

In Sect. 2 we present basic issues of mathematical modelling of generation units and their interaction in the VEAG system. Section 3 then addresses a planning model for mid-term cost optimal power production under uncertainty of electrical load, fuel prices and electricity prices in power contracts. In Sect. 4 we develop some ideas on how to couple day-ahead trading at a power exchange with the basic production model from Sect. 2. In both Sects. 3 and 4 we place accent on modelling issues, sketch the optimization algorithms for problem solution, and display initial numerical experience.

2 Power Production – The VEAG System

The generation system of VEAG Vereinigte Energiewerke AG Berlin comprises pumped-storage hydro as well as coal and gas fired thermal power
plants. Throughout, we work with a (mostly hourly) discretization of the planning horizon into $t = 1, \ldots, T$ subintervals, and we assume that there are $I$ thermal as well as $J$ pumped-storage hydro units. The variable $u_{it} \in \{0, 1\}, i = 1, \ldots, I; t = 1, \ldots, T$ indicates whether the thermal unit $i$ is in operation at time $t$. Variables $p_{it}, s_{jt}, w_{jt}, i = 1, \ldots, I, j = 1, \ldots, J; t = 1, \ldots, T$ are the output levels for the thermal units, the hydro units in generation and in pumping modes, respectively. The variables $l_{jt}$ denote the fill (in energy) of the upper dam of the hydro unit $j$ at the end of interval $t$.

The power output of units and the fill of the upper dams have to fit the following bounds

$$\begin{align*}
p_{it}^{\min} u_{it} &\leq p_{it} \leq p_{it}^{\max} u_{it}, i = 1, \ldots, I, \ t = 1, \ldots, T \\
0 &\leq s_{jt} \leq s_{jt}^{\max}, \quad j = 1, \ldots, J, \ t = 1, \ldots, T \\
0 &\leq w_{jt} \leq w_{jt}^{\max}, \quad j = 1, \ldots, J, \ t = 1, \ldots, T \\
0 &\leq l_{jt} \leq l_{jt}^{\max}, \quad j = 1, \ldots, J, \ t = 1, \ldots, T
\end{align*}$$

(1)

Here, $p_{it}^{\min}, p_{it}^{\max}, s_{jt}^{\max}, w_{jt}^{\max}$ denote minimal and maximal outputs, respectively, and $l_{jt}^{\max}$ is the maximal fill of the upper dam. Load coverage is modeled by the constraints

$$\sum_{i=1}^{I} p_{it} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) \geq d_{t}, \quad t = 1, \ldots, T,$$

(2)

where $d_{t}$ denotes the electrical load at time $t$.

In addition to load coverage, reserve management is a key issue in power production. Quite different reserve schemes are employed in practice. At least the following requirement involving a so-called spinning reserve $r_{t}$ has to be met for the thermal units:

$$\sum_{i=1}^{I} (u_{it} p_{it}^{\max} - p_{it}) \geq r_{t}, \quad t = 1, \ldots, T.$$

(3)

For the pumped-storage plants we have the following balances that interconnect different time intervals:

$$\begin{align*}
l_{jt} &= l_{jt-1} - (s_{jt} - \eta_{j} w_{jt}), \\
l_{j0} &= l_{j0}^{\text{in}}, \quad l_{jT} = l_{jT}^{\text{end}}, \quad \left\{ j = 1, \ldots, J, \ t = 1, \ldots, T \right\}
\end{align*}$$

(4)

Here, $l_{j0}^{\text{in}}, l_{jT}^{\text{end}}$ are the initial and final fills (in energy) of the upper dams, $\eta_{j}$ denote the pumping efficiencies. Constraints avoiding simultaneous generation and pumping in the hydro plants are dispensable since it can be shown that such a deficiency can not occur in optimal points.

To avoid excessive thermal stresses in the coal fired blocks, they have to adhere to minimum up and down times $\sigma_{i}$ and $\tau_{i}$. These are modeled via

$$\begin{align*}
u_{it} - u_{i,t-1} &\leq u_{i,\sigma}, \quad \sigma = t + 1, \ldots, \min\{t + \sigma_{1} - 1, T\}, \\
u_{i,t-1} - u_{it} &\leq 1 - u_{i,\tau}, \quad \tau = t + 1, \ldots, \min\{t + \tau_{1} - 1, T\}, \quad i = 1, \ldots, I; t = 2, \ldots, T - 1.
\end{align*}$$

(5)
The constraints (1)–(5) provide a mathematical model for basic features and basic interaction of the generating units of the VEAG system. Typical objective functions to be minimized on the constraint set (1)–(5) concern fuel consumption for starting up and operating thermal units together with costs (or revenues) according to power contracts. Formalizing fuel cost minimization, for instance, leads to the objective function

\[
\sum_{t=1}^{T} \sum_{i=1}^{I} C_{it}(p_{it}, u_{it}) + \sum_{t=1}^{T} \sum_{i=1}^{I} S_{it}(u_{i}),
\]

where \(C_{it}\) denote the piecewise linear fuel costs and \(S_{it}(u_{i}) = c_{it} \max\{u_{it} - u_{i,t-1}, 0\}\) the start-up costs for the thermal unit \(i\), where \(u_{i} := (u_{it})_{t=0}^{T}\) and \(u_{i0}\) is a given initial value.

In Sects. 3 and 4, the above mathematical apparatus will be utilized as starting point for including the sources of uncertainty outlined in the introduction.

### 3 Mid-Term Power Planning

For the system introduced in Sect. 2 we describe a model for its mid-term cost-optimal power production planning under uncertainty on the electrical load and on the electricity and fuel prices. For mid-term planning models we are faced with stochastic data when considering time periods lying far in the future. In order to derive solutions that hedge against uncertainty it is necessary to incorporate the randomness of the data into the model. So far this is mainly done for operational models (cf. [4,6] and the references therein).

Since we regard future planning periods (e.g. next week or year), we assume that the quality of available information on the load uncertainty does not depend on time, i.e., it does not increase with the length of the planning horizon. Furthermore, the load is stochastic right from the beginning of the considered time period. The stochastic behaviour of the load \(d_{t}\), the spinning reserve \(r_{t}\) and the price for fuel and electricity – characterized by its coefficients \(a_{t}, b_{t}\) and \(c_{t} – \) is represented by a discrete-time stochastic process \(\{\xi_{t} := (d_{t}, r_{t}, a_{t}, b_{t}, c_{t})\}_{t=1}^{T}\) on some probability space \((\Omega, \mathcal{A}, P)\). Now, the decision process consists of two stages where the first-stage decisions correspond to the here-and-now schedules for all power generation units. The second-stage decisions, on the other hand, correspond to future compensation or recourse actions of each unit in each time period. The latter naturally depend on the environment created by the first-stage decisions and the load and price scenarios in that specific time period. Hence, the aim of such a two-stage planning model can be formulated as follows. Find an optimal schedule for the whole system and planning horizon such that the uncertain data can be compensated by the system, all system constraints are satisfied and the sum of the total generation costs and the expected recourse costs is minimal.
In order to give a mathematical formulation of the two-stage model let \((u, p, s, w)\) denote the first-stage decisions and \((u, p, s, w)\) be the stochastic second-stage decisions having the components \(u_{it}, p_{it}, s_{jt}, w_{jt}\), which correspond to the recourse actions of each unit at time period \(t\). In addition to the (non-stochastic) constraints for \((u, p, s, w)\), i.e. the capacity limits (1), the storage dynamics (4) and the minimum up- and down-times (5), we have to require that the recourse actions also satisfy the system constraints. These are the operating ranges of all units, the minimum up/down-time requirements for the thermal units and the reservoir capacity bounds:

\[
p_i^{\min} u_{it} \leq p_{it} \leq p_i^{\max} u_{it}, \, u_{it} \in \{0, 1\},
\]

\[
u_{it} - u_{i,t-1} \leq u_{i,\sigma}, \quad \sigma = t + 1, \ldots, \min\{t + \sigma_i - 1, T\},
\]

\[
u_{i,t-1} - u_{it} \leq 1 - u_{i,\tau}, \quad \tau = t + 1, \ldots, \min\{t + \tau_i - 1, T\},
\]

\(t = 2, \ldots, T - 1, \, i = 1, \ldots, I, \quad \mathbb{IP} - a. \, s.\)

\(0 \leq s_{jt} \leq s_{jt}^{\max}, \quad 0 \leq w_{jt} \leq w_{jt}^{\max}, \quad 0 \leq \ell_{jt} \leq \ell_{jt}^{\max},
\]

\(t = 1, \ldots, T, \quad j = 1, \ldots, J, \quad \mathbb{IP} - a. \, s.\)

\[
\ell_{jt} = \ell_{j,t-1} - s_{jt} + \eta_{jw_{jt}}, \quad t = 1, \ldots, T,
\]

\[
\ell_{j0} = \ell_{j0}^{\min}, \quad \ell_{jT} = \ell_{jT}^{\end}, \quad j = 1, \ldots, J, \quad \mathbb{IP} - a. \, s.
\]

Here some remarks concerning the interplay of the two stages are due. The first-stage solutions act as a basis for the recourse actions, which have to satisfy the second-stage constraints in a cost-optimal way. To this end we have to guarantee that the transition from the first to the second stage is feasible. While the static constraints (7a) and (7c) need no further consideration, we neglect the possible impact of the constraints (7d). This is justified since the generation system is thermal dominated. The minimum up- and down-times constraints (7b) for the thermal units, however, need some refinement. In order to enforce compatibility between the first- and second-stage decisions we introduce constraints that relate the scheduling behaviours of the two stages to each other. This means that we prevent a thermal unit from being switched on or off in the second stage, if the scheduling history in the first stage prohibits that. The same canonical dependency is required in the other direction as well, i.e. we restrict switching in the first stage subject to the constraints set by the second-stage scheduling. Thus we have the constraints:

\[
u_{it} - u_{i,t-1} \leq 1 - (u_{i,\sigma - 1} - u_{i,\sigma}), \quad \sigma = t, \ldots, \min\{t + \sigma_i - 1, T\},
\]

\(i = 1, \ldots, I, \quad t = 2, \ldots, T - 1.
\]

Observe the consequences of the compatibility constraints (8). The inequality (8b), say, represents a constraint for the second stage if and only if unit \(i\) is
switched off in the first stage at time $t$. In this case it enforces that the thermal unit will not be switched on in the second stage as long as the unit is cooling for its minimum down-time in the first stage. The remaining inequalities have similar effects.

Furthermore we introduce a subdivision of the set $\mathcal{I} := \{1, \ldots, I\}$ of all thermal units into two subsets $\mathcal{I}_1$ and $\mathcal{I}_2$ such that $\mathcal{I}_1 \cup \mathcal{I}_2 = \mathcal{I}$ and the conditions $u_{it} = u_{it}$, $i \in \mathcal{I}_2$, $t = 1, \ldots, T$, $IP$ - a. s., are satisfied. This means that only some of the available thermal units may change their on/off state when compensating uncertain data. From a modelling point of view this approach leads to a reduction of the number of binary variables corresponding to a unit $i \in \mathcal{I}_2$. Moreover the case $\mathcal{I}_2 = \mathcal{I}$ conforms with the view taken in [3]. There all on/off decisions of the thermal units have been regarded as long-term decisions and thus belonging to the first stage only. Observe that (8) is clearly satisfied for all $i \in \mathcal{I}_2$.

The loading constraints (2) have to be adapted to the new situation. Here we distinguish between the two stages. As mentioned before we are looking for a solution to the here-and-now decisions that satisfies the uncertain demand with a certain probability and furthermore allows an optimal scheduling in each of the scenarios. That is why the first-stage power outputs of all generation units have to satisfy at least the expected load, while the second-stage power outputs are required to satisfy the load $d_t$ with probability one. Hence the (modified) loading constraints are given by the following inequalities:

$$\sum_{i=1}^{I} p_{it} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) \geq \mathbb{E}(d_t), \ t = 1, \ldots, T, \ \text{(9a)}$$

$$\sum_{i=1}^{I} p_{it} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) \geq d_t, \ t = 1, \ldots, T, \ IP - \text{a. s.} \ \text{(9b)}$$

The reserve constraints (3) are modified in the same way. Note that from now on we use an equivalent characterization of the reserve constraints:

$$\sum_{i=1}^{I} u_{it} p_{it}^{\max} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) \geq \mathbb{E}(d_t + r_t), \ t = 1, \ldots, T, \ \text{(10a)}$$

$$\sum_{i=1}^{I} u_{it} p_{it}^{\max} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) \geq d_t + r_t, \ t = 1, \ldots, T, \ IP - \text{a. s.} \ \text{(10b)}$$

Again the second-stage decisions cover the specified amount with probability one, while the first-stage spinning reserve meets at least the expected demand. Finally we incorporate the stochastic fuel and electricity prices into the model. To this end we define the random variables $C_{it}$, that describe the costs for operating thermal unit $i$ in the second-stage during time period $t$, in the following way: $C_{it}(p, u) := \max_{l=1, \ldots, I} \{a_{ilt} p + b_{ilt} u\}$, where $a_{ilt}, b_{ilt}$ are components of the random variable $\xi_t$ that represent stochastic cost.
coefficients such that $C_{it}(\cdot, 1)$ is $\mathcal{P}$-almost surely convex and increasing on $\mathbb{R}_+$. The cost functions $C_{it}(p, u)$ for the first stage we define accordingly, taking the expected values $\mathbb{E}(a_{ilt}), \mathbb{E}(b_{ilt}), l = 1, \ldots, I$, as price coefficients, however. The effect of stochastic prices on the startup costs is modelled in a similar way. More precisely we have $S_{it}(u_i) := c_{it} [u_i - u_{i,t-1}]^+, S_{it}(u_i) := \mathbb{E}(c_{it})[u_{it} - u_{i,t-1}]^+$, where $c_{it}$ are stochastic startup cost coefficients and $u_i := (u_{it})_{t=0}^T, u_{i0} = u_{i0} \mathcal{P}$-a.s., $i = 1, \ldots, I$.

In consistency with common two-stage stochastic programming the objective function corresponds to the total costs for operating the thermal units in the first stage plus the expected costs in the second stage, i.e.,

$$
\sum_{i=1}^I \sum_{t=1}^T \left[ C_{it}(p_{it}, u_{it}) + S_{it}(u_i) \right] + \mathbb{E} \left( \sum_{i=1}^I \sum_{t=1}^T \left[ C_{it}(p_{it}, u_{it}) + S_{it}(u_i) \right] \right). \quad (11)
$$

The stochastic power production planning model consists then in minimizing the objective function (11) over all deterministic decisions $(u, p, w, s)$ and all stochastic decisions $(u, p, s, w) \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{P}; \mathbb{R}^{2T(I+J)})$ satisfying the constraints (1), (4), (5), (7)–(10). The model represents a two-stage stochastic mixed-integer program involving $2(I + J)T$ deterministic and $2(I + J)T$ stochastic decision variables.

The stochastic program elaborated above is almost separable with respect to the units, since only the constraints (9)–(10) couple different units. This structure allows us to apply a stochastic version of the classical Lagrangian relaxation idea. We relax the above mentioned coupling constraints by introducing Lagrange multipliers $\lambda := (\lambda^1, \lambda^2, \lambda^3, \lambda^4)$, where $\lambda^1, \lambda^2 \in \mathbb{R}_+^T$ and $\lambda^3, \lambda^4 \in L^1(\Omega, \mathcal{A}, \mathcal{P}; \mathbb{R}_+^I)$. Setting $x := (u, p, s, w)$ and $\lambda(x) := (u, p, s, w)$ the Lagrangian $L(x, x; \lambda)$ is formed in the following way. Each of the constraints (9)–(10) is first associated with a Lagrange multiplier and then enters the objective function (11). Here the deterministic multipliers $\lambda^1, \lambda^2$ are linked with the first-stage constraints (9a) and (10a), whereas the stochastic variables $\lambda^3, \lambda^4$ are paired with the second-stage constraints (9b), (10b). Note that $L(\cdot)$ takes its values in $\mathbb{R}$, since the latter constraints enter the objective function via their expected value (cf. [10]). With the dual function $D(\lambda) := \inf_{(x, x)} L(x, x; \lambda)$, where the infimum is taken subject to the constraints (1), (4), (5), (7) and (8), the dual problem reads

$$
\max \{ D(\lambda) : \lambda \in \mathbb{R}^{2T(I+J)} \times L^1(\Omega, \mathcal{A}, \mathcal{P}; \mathbb{R}_+^{2T}) \}. \quad (12)
$$

The optimal value of the dual problem (12) provides a lower bound for the optimal costs of the (nonconvex) primal model. The minimization of $L$ decomposes into stochastic single unit subproblems and the dual function

$$
D(\lambda) = \sum_{i=1}^I D_i(\lambda) + \sum_{j=1}^J \hat{D}_j(\lambda) + \sum_{t=1}^T \left[ \lambda^1_t \mathbb{E}(d_t) + \lambda^2_t \mathbb{E}(d_t + r_t) + \mathbb{E} (\lambda^3_t d_t + \lambda^4_t (d_t + r_t)) \right]. \quad (13)
$$
may be evaluated by solving thermal subproblems and hydro subproblems corresponding to $D_i(\lambda)$ and $D_j(\lambda)$, respectively. The two kinds of subproblems represent two-stage stochastic programming models for the operation of a single unit. While the thermal subproblems are mixed-integer two-stage stochastic program, that reduce to combinatorial two-stage stochastic problems, the hydro subproblems are linear two-stage models. For the latter we make use of a descent algorithm described in [9], whereas for the thermal subproblems we employ stochastic dynamic programming when it is necessary. This means in particular that for the solution of the thermal subproblems we first relax the compatibility constraints (8) in order to apply a scenario decomposition approach. Then the stochastic dynamic programming need only be used if the preliminarily solution violates the compatibility constraints (cf. [10]). Extending Lagrangian relaxation approaches for deterministic power management models, our method for solving the stochastic two-stage model consists of the components described in Fig. 1. The non-smooth optimization problem (12) is solved with the proximal bundle method described in [7]. Bundle methods make use of function value and subgradient information in order to determine descent directions. The subproblem solvers provide that necessary information. The bundle method delivers an optimal value $D(\lambda^*)$ of (12), which is a lower bound for the optimal cost of the primal model. In general, however, the dual optimal scheduling decisions violate the load and reserve constraints. With a Lagrangian heuristics, that takes its basic ideas from [5,12], we determine a primal feasible solution. For the case $I_2 = I$ we use a strategy that treats all scenarios simultaneously, whereas in all other instances each scenario is dealt with independently. Finally we fix the binary decisions and use a deterministic version of the economic dispatch algorithm in [9] for every scenario to find a nearly optimal primal solution. The interaction of the described components is illustrated in Fig. 1.

The stochastic Lagrangian relaxation algorithm was implemented in C++ except for the proximal bundle method, for which the FORTRAN-package NOA 3.0 [8] was used as a callable library. For testing the implementation a number of load scenarios was simulated from a time series model for the load process described in [6]. Furthermore the stochastic prices have been

![Fig. 1. Structure of the stochastic Lagrangian relaxation method](image-url)
simulated by a discretized geometric Brownian motion. Test runs have been carried out for the weekly production planning (i.e. \( T = 168 \)) of the hydrothermal power generation system of VEAG comprising 25 thermal units and 7 pumped storage plants and for a number of scenarios ranging from 5 to 100. The corresponding primal optimization problems have up to 400,000 binary and 650,000 continuous variables, and more than 1,300,000 constraints. Table 1 shows computing times and gaps for different choices of the optimality tolerances.

Table 1. Computing times and gaps (NOA 3.0: NGRAD= 20)

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>( \mathcal{I}_2 )</th>
<th>( \text{opt. tol: } 10^{-3} )</th>
<th>( \text{opt. tol: } 10^{-4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>time [min]</td>
<td>gap [%]</td>
<td>time [min]</td>
</tr>
<tr>
<td>5</td>
<td>( \emptyset )</td>
<td>0:16</td>
<td>1:00</td>
</tr>
<tr>
<td>5</td>
<td>( \emptyset )</td>
<td>0:18</td>
<td>0:35</td>
</tr>
<tr>
<td>10</td>
<td>( \emptyset )</td>
<td>0:57</td>
<td>2:50</td>
</tr>
<tr>
<td>10</td>
<td>( \mathcal{I} )</td>
<td>0:44</td>
<td>2:04</td>
</tr>
<tr>
<td>5</td>
<td>( \mathcal{I} )</td>
<td>0:10</td>
<td>0:28</td>
</tr>
<tr>
<td>10</td>
<td>( \mathcal{I} )</td>
<td>0:27</td>
<td>0:59</td>
</tr>
<tr>
<td>50</td>
<td>( \mathcal{I} )</td>
<td>9:00</td>
<td>12:04</td>
</tr>
<tr>
<td>100</td>
<td>( \mathcal{I} )</td>
<td>30:17</td>
<td>35:31</td>
</tr>
</tbody>
</table>

Fig. 2. Solution for 10 scenarios and \( \mathcal{I}_2 = \emptyset \)
tolerance for the proximal bundle method. The test runs have been performed on an HP 9000 (780/J280) Compute-Server with 180 MHz frequency and 768 MByte main memory under HP-UX 10.20. The results show that a smaller optimality tolerance leads to smaller gaps at the expense of the computing times. Here the gap refers to the relative difference of the costs of the determined scheduling decisions \((x, x)\) and the optimal value \(D(\lambda^*)\) of the dual problem. In the case of \(I_2 = \emptyset\) the performance of the algorithm is closely related to the efficiency of the thermal subproblem’s solver. In particular it depends on how often the stochastic dynamic programming algorithm is used during the dual maximization. In fact the complexity of the involved memory structures increases very fast, so that problem instances with more than 10 scenarios cannot be handled so far. Figure 2 provides a sample output of the algorithm. It is worth mentioning that typical solutions exhibit both switching on and off decisions at the transition from the first to the second stage. Thus a solution for this case in general yields a better objective function value than the solution to the corresponding problem, where the index set \(I_2\) consists of all thermal units (cf. [10]).

4 Day-Ahead Trading and Power Production

Power exchanges with future and spot markets are novel instruments arising at liberalized electricity markets [1]. Trading at a power exchange is becoming a means of ever growing importance in the utility’s total economic activity. Coordination of trading and power production is a key issue in this respect. In what follows we address simultaneous power production and day-ahead trading at a power exchange. The production component is modeled according to Sect. 2.

Day-ahead trading involves sealed selling and purchase bids for every individual hour of the day ahead. Each offer comprises volume and price. There is only one round of bidding and the market price is cleared by an independent operator such that the total exchange is maximized. Figure 3 depicts the price formation mechanism: selling and purchase offers are placed in (price-) ascending and descending orders, respectively, yielding two step curves where step length corresponds to volume and step height to price. The intersection of the curves determines both the market price and the total volume traded. Selling offers strictly below and purchase offers strictly above market price are executed completely. Vice versa, sellings strictly above as well as purchases strictly below do not become effective. Offers at market price in general are executed only partially, with specific splitting rules in case of multiple offers with identical price.

A simultaneous management of electricity production and day-ahead bidding thus has to take into account two main factors: the utility’s possibilities to generate or consume a certain amount of power at a certain price at a certain time, and the competitors’ bids for that very time period. The latter
price per volume

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Determining of the market price by mean of offers}
\end{figure}

being sealed and thus not available for the utility at the time of bidding, this induces considerable uncertainty for the utility at the moment of decision making.

The mathematical modelling of the phenomena sketched above is accomplished in two steps. First the price formation mechanism at the power exchange is formulated in mixed-integer linear terms and concatenated with the production model from Sect. 2. Afterwards, this deterministic optimization model is extended into a stochastic one by introducing probabilistic scenarios for the bidding behaviour of the competitors and heading for an optimization of the nonanticipative decisions to be taken by the utility. In what follows, we will display the first step in detail and only outline the extension towards stochastic programming.

To model price formation we consider the discretization \( t = 1, \ldots, T \) of the optimization horizon from Sect. 2. The price is discretized into \( m = 1, \ldots, M \) ascending levels with values \( b_{mt} \geq 0 \). For all \( t = 1, \ldots, T; m = 1, \ldots, M \) we introduce triplets \((v_{mt}^s, v_{mt}^p, v_{mt}^c) \in \{0,1\}^3\) of indicators. The subsequent constraints system ensures that \( v_{mt}^s = 1 \) (\( v_{mt}^p = 1 \)) iff the selling (purchase) offer with price level \( m \) at time \( t \) is strictly below (above) market price and hence executed completely. Moreover, we have \( v_{mt}^c = 1 \) iff price level \( m \) coincides with the valid market price at time \( t \):

\[
\begin{align*}
\sum_{m=1}^{M} v_{mt}^c &= 1, \ t = 1, \ldots, T; \\
v_{mt}^s &\geq v_{m+1,t}^s, \\
v_{mt}^p &\leq v_{m+1,t}^p, \\
v_{mt}^s + v_{mt}^c + v_{mt}^p &= 1,
\end{align*}
\]

(14)
We distinguish between variables for energy volumes of own offers of price level \( m \) in time period \( t \) (\( q^s_{mt}, p^p_{mt} \geq 0 \)) and variables for volumes of own executed offers of price level \( m \) in time period \( t \) (\( p^s_{mt}, p^p_{mt} \geq 0 \)). Executed offers cannot exceed offers:

\[
0 \leq p^s_{mt} \leq q^s_{mt}, \ 0 \leq p^p_{mt} \leq q^p_{mt}; \quad t = 1, \ldots, T, \ m = 1, \ldots, m. \tag{15}
\]

Own selling offers below and own purchase offers above market price are to be executed completely:

\[
p^s_{mt} \geq q^s_{mt} - C_1(1 - v^s_{mt}), \quad p^p_{mt} \geq q^p_{mt} - C_1(1 - v^p_{mt}), \quad t = 1, \ldots, T, \ m = 1, \ldots, m \tag{16}
\]

with a sufficiently big constant \( C_1 \). Own selling (purchase) offers must not be executed, if they are above (below) the valid price:

\[
p^s_{mt} \leq C_2(v^s_{mt} + v^c_{mt}), \quad p^p_{mt} \leq C_2(v^p_{mt} + v^c_{mt}), \quad t = 1, \ldots, T, \ m = 1, \ldots, m \tag{17}
\]

with a sufficiently big constant \( C_2 \).

The competitors' (or foreign) selling and purchase offers of price level \( m \) for time \( t \) are denoted by \( f^s_{mt}, f^p_{mt} \geq 0, m = 1, \ldots, M, t = 1, \ldots, T \), respectively. It is feasible that not the complete volume of a competitor's offer at market price is executed (as in our example, Fig. 3). \( \beta^s_t, \beta^p_t \geq 0 \) then denote the actually executed volumes of the competitors at valid market price:

\[
0 \leq \beta^s_t \leq \sum_{m=1}^{M} v^c_{mt} f^s_{mt}, \quad 0 \leq \beta^p_t \leq \sum_{m=1}^{M} v^c_{mt} f^p_{mt}, \quad t = 1, \ldots, T. \tag{18}
\]

Maximum exchange is reached at the equilibrium of supply and demand:

\[
\sum_{m=1}^{M} (v^s_{mt} f^s_{mt} + p^s_{mt}) + \beta^s_t = \sum_{m=1}^{M} (v^p_{mt} f^p_{mt} + p^p_{mt}) + \beta^p_t, \quad t = 1, \ldots, T. \tag{19}
\]

At market price, either a complete selling offer or a complete purchase offer or both a complete selling and a complete purchase offer have to be executed (cf. Fig. 4). Introducing another indicator \( v^0_t \in \{0, 1\} \) attaining the value 1 iff at least all selling offers at market price are executed, the price formation model is completed by the following constraints:

\[
\beta^s_t \geq \sum_{m=1}^{M} v^c_{mt} f^s_{mt} - C_3(1 - v^0_t), \quad \beta^p_t \geq \sum_{m=1}^{M} v^c_{mt} f^p_{mt} - C_3v^0_t, \quad t = 1, \ldots, T; \quad p^s_{mt} \geq q^s_{mt} - C_3(2 - v^c_{mt} - v^0_t), \quad p^p_{mt} \geq q^p_{mt} - C_3(1 - v^c_{mt} + v^0_t), \quad t = 1, \ldots, T, \ m = 1, \ldots, M \tag{20}
\]

with a sufficiently big constant \( C_3 \).
Let \( g_{mt} \in \mathbb{R} \) denote the net exchange at the power exchange of price level \( m \) at time \( t \). It has to fulfill

\[
\sum_{m=1}^{M} g_{mt} = \sum_{m=1}^{M} (p_{mt}^p - p_{mt}^g), \quad t = 1, \ldots, T, \\
-C_4 v_{mt}^c \leq g_{mt} \leq C_4 v_{mt}^c, \quad t = 1, \ldots, T, \quad m = 1, \ldots, M
\]

with a sufficiently big constant \( C_4 \). Finally, concatenation with the production model from Sect. 2 is achieved by modifying the load coverage constraints into

\[
\sum_{i=1}^{I} p_{it} + \sum_{j=1}^{J} (s_{jt} - w_{jt}) + \sum_{m=1}^{M} g_{mt} = d_t, \quad t = 1, \ldots, T
\]

and adding the term

\[
\sum_{t=1}^{T} \sum_{m=1}^{M} b_{mt} g_{mt}.
\]

The above model was validated with VEAG data for the power system and market prices of the Amsterdam Power Exchange (APX) [1] for the power exchange. The VEAG power system comprises 17 coal and 8 gas fired thermal units as well as 7 pumped storage plants. Using APX market prices from the time period July to December 1999 we constructed hourly foreign offers of different prices and volumes. To study impacts of power trading to power production we formed test problems both with actual APX price levels (problem B) and with levels deviating form the APX prices by up to 30% (problem C). In addition the model was run as a pure productions model, without trading (problem A). Model sizes varied from 18 000 to 63 000 constraints and 13 000 (2 500 integer) to 35 000 (6 000 integer) variables. Numerical tests were performed on a SUN E450 ultra SPARC, 300 MHz, using
Table 2. Calculations

<table>
<thead>
<tr>
<th>Prob.</th>
<th>Time</th>
<th>Best Solution</th>
<th>Lower Bound</th>
<th>Gap</th>
<th>Min. Saving</th>
<th>Selling</th>
<th>Purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0:20:14</td>
<td>46287933</td>
<td>46287933</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2:27:59</td>
<td>46185297</td>
<td>44892456</td>
<td>2.80%</td>
<td>0.22%</td>
<td>11200</td>
<td>2890</td>
</tr>
<tr>
<td>C</td>
<td>1:15:49</td>
<td>45767961</td>
<td>45242528</td>
<td>1.15%</td>
<td>1.12%</td>
<td>40900</td>
<td>1570</td>
</tr>
</tbody>
</table>

Table 2 reports solution times and optimality certificates (gaps) together with the impact of trading. For problems B and C gains of 0.22% and 1.12%, respectively, were achieved, compared with the pure-production instance A. Solution times behind these savings are quite substantial. However, savings of 0.14% and 0.48% were achieved for problems B and C after only 43 and 20 minutes of solution time, respectively. Trading activities (selling and purchase volumes) over time for problem C are displayed in Fig. 5. These initial numerical results indicate the potential of simultaneously optimizing power production and trading.

To remove the unrealistic model assumption of anticipating the competitors’ offers the latter are considered to be stochastic, i.e., given scenario-wise with certain probabilities. Then a two-stage stochastic program becomes appropriate. The principal modelling follows the lines sketched in a different context in Sect. 3: We consider an optimization horizon of 7 days and assemble all the decisions to be taken at the first day into the first stage. These comprise production decisions (variables \( p_{it}, s_{jt}, w_{jt}, l_{jt} \) for suitable \( t \) and all \( i, j \) ) and the utility’s offers for the subsequent day (variables \( q_{mt}^{s}, q_{mt}^{p}, g_{mt} \) for suitable \( t \) and all \( m \) ). The remaining variables enter the second stage. The objective function is a sum of direct costs caused by the first-stage decisions and of expected future costs in the second stage. Solution methodology for this two-stage stochastic integer program rest on the dual decomposition method developed in [2]. Again a non-smooth concave dual maximization problem has

![Fig. 5. Trading activities over time](image-url)
to be solved (cf. Sect. 3) whose objective function values and subgradients are determined by solving single-scenario subproblems. The latter are of the same type as the above model (14)–(23). Research along these directions is still ongoing and will be reported elsewhere.

References

X. Optimization in Traffic and Communication

Duty Scheduling in Public Transit
   M. Grötschel, R. Borndörfer, A. Löbel

Rotation Planning for the Continental Service of a European Airline
   M. Jünger, M. Elf, V. Kaibel

Scheduling Scarce Resources in Chemical Engineering
   R. H. Möhring, M. Uetz

Computer Aided Scheduling of Switching Engines
   U. T. Zimmermann, M. E. Lübbecke

Train Schedule Optimization in Public Rail Transport
   U. T. Zimmermann, T. Lindner

Integrated Planning Approach for Cellular Radio Networks
   R. Mathar, Michael Schmeink
Scheduling Scarce Resources
in Chemical Engineering*

Rolf H. Möhring and Marc Uetz

Technische Universität Berlin, Fachbereich Mathematik, Sekr. MA 6–1, Straße des 17. Juni 136, 10623 Berlin, Germany, E-mail: {moehring,uetz}@math.tu-berlin.de

Abstract. The efficient utilization of scarce resources, such as machines or manpower, is major challenge within production planning in the chemical industry. We describe solution methods for a resource-constrained scheduling problem which arises at a production facility at BASF AG in Ludwigshafen. We have developed and implemented two different algorithms to solve this problem, an approach which is based on Lagrangian relaxation, as well as a branch-and-bound procedure. Particularly the Lagrangian approach is applicable for a whole variety of resource-constrained scheduling problems, hence it is of interest not only for the specific problem we describe, but also for many other industrial applications. In this paper, we describe both approaches, and also report on computational results, based upon practical problem instances as well as benchmark test sets.

1 Scheduling Under Limited Resources

Facing the increasing international competition, there is a growing need for planning tools in chemical engineering that allow an efficient utilization of scarce resources. Within the cooperation with BASF AG, Ludwigshafen, we focus on a scheduling problem that is typical for a production process in the chemical industry, but also occurs in many other industrial applications.

The aim is to schedule the production process for several so-called orders, or campaigns. Each order represents the demand of a certain amount of a product, which must be produced on a suited machine. Due to limited machine capacities, an order is usually split up into several identical steps, so-called jobs. In other words, the production process for each order consists of a sequence of identical jobs, each of which must be scheduled on an assigned machine. The objective is to schedule all orders, or jobs, such that the overall production time is minimized. Due dates for individual orders are given, and there may also be temporal constraints between jobs of different orders in the form of time lags, for instance if an intermediate product of an order is needed within the production process of others.

* This work is a final report on a research project with BASF AG, Ludwigshafen. The authors have been funded by the Bundesministerium für Bildung und Forschung (bmb+f) under grant 03-M07TU1-3. Most of the results are based on joint work with different colleagues, namely Andreas Fest (TU Berlin), Andreas S. Schulz (MIT, Cambridge, MA), and Frederik Stork (TU Berlin).
Apart from temporal restriction, resource constraints are imposed by limited availability of machines and scarce personnel. The availability of personnel depends on the organization of shifts, breaks, etc. Any job itself consists of several consecutive steps, so-called tasks, and each of these tasks requires a specified amount of personnel. Hence, the required personnel to operate a job is varying in time, and given by a corresponding piecewise constant function like depicted in Fig. 1. In a less restrictive model, the tasks of a job are only linked by so-called time-windows, which means that there may be idle periods of restricted time between two consecutive tasks. This leads to a model with so-called maximal time lags, which will be discussed in Sect. 6.

The objective is to find a production schedule such that temporal as well as resource constraints are satisfied, and the overall production time, the so-called makespan, is minimized. For a more detailed problem description, we refer to [KW] or [MSSU3].

1.1 Notation and Formulation of the Problem

The problem described above can be formulated as a resource-constrained project scheduling problem. In order to put up a mathematical formulation, we introduce some notation. Let us denote by $J = \{0, \ldots, n\}$ be the set of all jobs which have to be scheduled, where a job $j$ has an integral processing time $p_j$. Except for Sect. 6, we assume that jobs must be processed without interruption, and by $S = (S_0, \ldots, S_n)$ we denote a schedule, where $S_j$ is the start time of job $j$. The entity of all jobs, together with their temporal constraints and resource requirements will be called a project in what follows. Jobs 0 and $n$ are assumed to be dummy jobs with processing time 0, indicating the project start and the project completion, respectively. In other words, job $n$ indicates when the production of all orders terminates. This can be easily modeled by introducing additional temporal constraints. In general, temporal constraints are given in the form of arbitrary minimal time lags between any pair of jobs, and by $d_{ij} \geq 0$ we denote the length of the required time lag.
(i, j) between two jobs i, j \in J. L \subseteq J \times J is the set of all time lags, and by 
\[ m := |L| \] we denote the number of given time lags. We assume without loss of 
generality that the time lags always refer to the start times of jobs, thus every 
schedule S has to fulfill S_j \geq S_i + d_{ij} for all (i, j) \in L. Ordinary precedence constraints can be represented by letting d_{ij} = p_i if job i must precede job j. Additionally, we suppose that a time horizon T as an upper bound on the project makespan (the time to produce all orders) is given. It can be checked in polynomial time by longest path calculations if such a system of temporal constraints has a feasible solution.

As indicated before, jobs need resources while they are operated. We assume that we are given a finite set R of different, so-called renewable resources, and the availability of resource \( k \in R \) at time \( t \) is denoted by \( R_{kt} \). These resources can represent machines, manpower, or any other device which is required to operate a job. We assume that a job \( j \) requires an amount of \( r_{jkt} \) units of resource \( k, k \in R \), during the \( t \)-th period of its execution. Note that a job may require several resources at the same time.

The objective is to find a schedule which respects all constraints, and has minimal total processing time, or in other words, which minimizes the project makespan \( S_n \). It is well known that this problem is \( NP \)-hard. It is even \( NP \)-hard to approximate within any constant factor (see [Sch]). That means that there is virtually no hope to find an efficient, that is, polynomial time algorithm which computes an optimum or near-optimum solution for all problem instances.

### 1.2 Integer Programming Formulation

In order to model the problem mathematically as an integer linear program, so-called time-indexed formulations are quite common. Pritsker, Watters, and Wolfe [PWW] were presumably the first to give an integer programming formulation in time-indexed 0-1-variables of the type

\[
x_{jt} = \begin{cases} 
1 & \text{if job } j \text{ starts at time } t, \\
0 & \text{otherwise}, 
\end{cases}
\]

where \( j \in J \), and \( t \in \{0, \ldots, T\} \). we then obtain the following integer linear programming formulation, which is crucial for our subsequent approach

\[
\begin{align*}
\text{minimize} & \quad \sum_t t \cdot x_{nt} \\
\text{subject to} & \quad \sum_t x_{jt} = 1, \quad j \in J, \\
& \quad \sum_{s=t}^{t+d_{ij}-1} x_{is} + \sum_{s=0}^{t} x_{js} \leq 1, \quad (i, j) \in L, \ t = 0, \ldots, T,
\end{align*}
\]
Note that \( S_j = \sum_t t x_{jt} \), thus the objective (1) is to minimize the project completion time \( S_n \). Constraints (2) indicate that each job is started exactly once, and inequalities (3) represent the temporal constraints given by the time lags \( L \). Inequalities (4) assure that the jobs simultaneously processed at time \( t \) do not consume more resources than available. Given the temporal constraints and the time horizon \( T \), it is easy to compute earliest and latest starting times for each job \( j \in J \). For convenience of notation, however, we simply assume (without stating explicitly) that variables with time indices outside these boundaries are fixed at values zero, such that no job is started before its earliest start or after its latest start, respectively.

2 Relaxations and Lower Bounds

Recall that we are dealing with a minimization problem which is even \( NP \)-hard to approximate within any constant factor. Hence, we cannot expect to find an efficient algorithm which computes a provably good solution for any given problem instance. Is such a situation, heuristic solutions are usually sought. However, in order to judge the quality of a given solution, lower bounds on the optimal objective value are crucial. In order to compute lower bounds, a variety of different methods are available. Usually, one tries to solve relaxations of the problem, preferably in such a way that the relaxed problem is solvable efficiently, that is, in polynomial time.

Based upon the above integer programming relaxation, it can be shown that a Lagrangian relaxation of the problem can be reduced to a minimum-cut problem, and thus can be solved quite efficiently [MSSU3]. The idea is to dualize the resource constraints (4), and introduce nonnegative Lagrangian multipliers \( \lambda = (\lambda_{kt}), k \in R, t \in \{0,...,T\} \). Doing this, one obtains the Lagrangian subproblem

\[
\begin{align*}
\text{minimize} & \quad \sum_t t x_{nt} + \sum_j \sum_{t} \left( \sum_{k \in R} \sum_{s=t}^{t+p_j-1} r_{jk,s-t} \lambda_{ks} \right) x_{jt} - \sum_{t} \sum_{k \in R} \lambda_{kt} R_{kt} \\
\text{subject to} & \quad (2), (3), (5), \text{and} (6).
\end{align*}
\]

It is well known that for any set of nonnegative Lagrangian multipliers, the optimum solution of the Lagrangian subproblem is a lower bound on the optimum objective value of problem (1)–(6). Already Christofides, Alvarez-Valdes, and Tamarit [CAT] have proposed to use this Lagrangian relaxation, however, they solved the Lagrangian subproblems by branch-and-bound. If
one omits the constant term $\sum_t \sum_{k \in R} \lambda_{kt} R_{kt}$ and introduces weights

$$w_{jt} = \begin{cases} \sum_{k \in R} \sum_{s=t}^{t+p_j-1} r_{jk,s-t} \lambda_{ks} & \text{if } j \neq n, \\ \sum_{k \in R} \sum_{s=t}^{t+p_j-1} r_{jk,s-t} \lambda_{ks} & \text{if } j = n, \end{cases}$$

problem (7) can be rewritten as

$$\text{minimize } c_\lambda(x) := \sum_j \sum_t w_{jt} x_{jt} \quad \text{subject to (2), (3), (5), and (6).}$$

(9)

Formulation (9) specifies the problem of finding a minimum-cost schedule for a set of jobs which must respect a set of temporal constraints, and where each job $j \in J$ incurs a cost of $w_{jt}$ if it is started at time $t$. We refer to this problem as project scheduling problem with start-time dependent costs. It will be discussed in Sect. 3 below.

3 Scheduling with Start-Time Dependent Costs

Not only due the fact that is arises as Lagrangian subproblem for various resource-constrained scheduling problems, the scheduling problem with start-time dependent costs as formulated in (9) is of interest in its own. For instance, it generalizes the well-known net-present-value problem [Rus] to arbitrary costs $w_{jt}$. In the paper by Christofides et al. [CAT], this problem is solved by a branch-and-bound algorithm.

However, it is in fact solvable in polynomial time, which follows, among others, from results by Chang and Edmonds [CE]. In [MSSU2], we have given an overview of the corresponding results. We next describe a direct reduction of the scheduling problem with start-time dependent costs given in (9) to a minimum-cut problem in a directed graph. This is a crucial result with respect to the efficiency of the Lagrangian approach, and we refer to [MSSU2] and [MSSU3] for more details.

Define a digraph $D$ by introducing a vertex $u_{jt}$ for every job $j \in J$ and every $t = 0, \ldots, T + 1$. Now, define directed chains $(u_{j0}, u_{j1}), (u_{j1}, u_{j2}), \ldots, (u_{jT}, u_{j,T+1})$ for any $j$. The corresponding arcs $(u_{jt}, u_{jt+1})$ are called assignment arcs. Furthermore, define arcs between those chains which correspond to the given temporal constraints by $(u_{it}, u_{j,t+d_{ij}})$ for all temporal constraints $(i, j) \in L$. The cost coefficients $w_{jt}$ are interpreted as the arc capacities of arc $(u_{jt}, u_{jt+1})$, for all $j$ and $t$. The capacity of the remaining arcs is set to infinity. Then, after the introduction of a dummy source $a$ and sink $b$, a solution of the original scheduling problem can be computed as a minimum $a-b$-cut in that digraph. Fig. 2 gives an example for the construction of $D$. 
Fig. 2. The left digraph represents the data of the sample instance: Each vertex represents an job, each arc represents a temporal constraint. The values for the time lags are $d_{12} = 1$, $d_{23} = -2$, $d_{34} = 2$, and $d_{54} = 3$. The job processing times are $p_1 = p_4 = 1$, $p_2 = p_5 = 2$, and $p_3 = 3$. $T = 6$ is a given upper bound on the project makespan. The right digraph $D$ is obtained by the above transformation. Each assignment arc $(u_{jt}, u_{j,t+1})$ corresponds to a binary variable $x_{jt}$ of formulation (9).

**Theorem 1 ([MSSU3]).** A minimum $a$-$b$-cut $(X, \bar{X})$ of the digraph $D$ described above corresponds to an optimal solution of the project scheduling problem with start time dependent costs (9) by virtue of

$$x_{jt} = \begin{cases} 1 & \text{if } (u_{jt}, u_{j,t+1}) \text{ is a forward arc of the cut } (X, \bar{X}), \\ 0 & \text{otherwise}. \end{cases}$$

Moreover, the value $c_\lambda(x)$ of that solution equals the capacity $c(X, \bar{X})$ of any minimum cut $(X, \bar{X})$ of $D$.

Using a push-relabel maximum-flow algorithm [GT], this results in an algorithm for solving the project scheduling problem with start-time dependent costs with running time $O(n m T^2 \log(n^2 T/m))$. Note that this holds for arbitrary (also maximal) time lags.

4 Linear Programming or Lagrangian Relaxation?

Based on this insight, the computation of lower bounds via Lagrangian relaxation can be realized quite efficiently: Using a standard subgradient optimization procedure to compute a near-optimal set of Lagrangian multipliers, the computation of lower bounds reduces to a series of minimum-cut computations in the above defined digraph. In fact it is well known that, since the polytope defined by inequalities (2), (3), and (5) is integral, the optimum solution value of the Lagrangian dual

$$c^* := \sup_{\lambda \geq 0} c_\lambda,$$
### Table 1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average CPU</th>
<th>Above Crit. Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP</td>
<td>38 min. (max. 23 h)</td>
<td>20.5%</td>
</tr>
<tr>
<td>LR</td>
<td>37 sec. (max. 8 min.)</td>
<td>19.6%</td>
</tr>
</tbody>
</table>

Comparison of computation times to solve the linear programming (LP) relaxation (1)–(5) and the multiplier problem for the Lagrangian relaxation (LR). Computation times were obtained on a Sun Ultra 2 with 200MHz, operating under Solaris. The linear programs have been solved with CPLEX version 6.5.3.

equals the value of an optimal solution of the linear programming relaxation (1)–(5). Here, $c_\lambda$ denotes the value of an optimum solution of (9) for a fixed value of Lagrangian multipliers $\lambda = (\lambda_{kt}), k \in R, t = 0, \ldots, T$.

The benefit of the new insights described in the preceding section is perhaps best documented by the comparison of the corresponding computation times in Table 1. There, we compare the computation times when solving linear programming relaxation (1)–(5) to the computation times to solve the Lagrangian multiplier problem as described above. The figures in Table 1 are based upon the well-established ProGen benchmark instances for resource constrained project scheduling [KS]. We used 600 ProGen instances, each with 120 jobs. The instances are available under [Pro1]. The table shows computation times as well as the average quality of the lower bounds in terms of improvement over the so-called critical path lower bound, which is the length of a longest path in the project network.

Obviously, the computation times are drastically reduced when using the Lagrangian approach instead of solving the linear programming relaxation. The average computation times for the Lagrangian approach are based upon an average number of 70 iterations. The quality of the corresponding lower bounds deviate only marginally, due to the fact that the subgradient optimization is not exact, but rather converges to the optimum value $c_\lambda^*$.

### 5 From Minimum Cuts to Feasible Solutions

Let us now turn to the question of how to obtain feasible solutions for the scheduling problem at BASF AG. So far, we have computed a lower bound on the best possible solution as described before. In fact, we not only have computed a lower bound, but in every iteration of the above described subgradient procedure, we have computed a schedule which respects all temporal restrictions, but might be infeasible with respect to the resource constraints. The intuition behind our approach to obtain good feasible solutions is to exploit some of the information held in these resource-infeasible schedules. The simple idea is to use list scheduling algorithms which are based upon so-called $\alpha$-completion times of the jobs.

Such relaxation-based ordering heuristics, combined with $\alpha$-completion time variables, have been used previously to obtain worst case performance
bounds for certain *machine scheduling* problems. The papers [PSW] and [HSW] are two of the early references in this direction. An application of *linear programming* based approaches to resource-constrained scheduling problems has been previously analyzed in [CDS+] and [SUW].

Let us now briefly sketch the framework for a combined algorithm which computes both, lower bounds and feasible solutions. First, to obtain an initial valid upper bound to start with, we use list scheduling algorithms fed with some standard priority rules. (Remember that we use a time-indexed formulation and require a time horizon $T$.) Then, in each iteration of the subgradient optimization algorithm, a time-feasible (but likely resource-infeasible) schedule is computed by solving the Lagrangian subproblem (9) as described in Sect. 3. The cost of this time-feasible schedule, in terms of the $w_{jt}$ defined in (8), is now a valid lower bound for the original problem (1)–(6). Using orderings according to $\alpha$-completion times of jobs, we then compute feasible solutions by means of different list scheduling algorithms, schematically depicted in Fig. 3. In fact, we have developed new list scheduling algorithms which showed to be particularly suited for our approach. See [SU] and [MSSU3] for more details.

With this Lagrangian approach, we could solve a practical instance provided by BASF AG to optimality within a couple of seconds by computing matching lower and upper bounds. The instance originally consists of 21 orders which are split up into 2121 jobs, resulting in a total number of 4545 tasks. The resulting number of time lags is 2100. Although this is a very large problem instance, it turned out to be comparatively easy to solve, since a simple preprocessing reduces its size to 129 jobs and 557 tasks.

Also on the well established ProGen benchmark instances [KS], the performance of the Lagrangian approach was very encouraging. In fact, the solutions compare to results obtained with state-of-the-art local search al-
Scheduling Scarce Resources in Chemical Engineering

<table>
<thead>
<tr>
<th>Average Deviations in %</th>
<th>CPU</th>
<th>best known lower bound</th>
<th>best known solution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>88 sec. (max. 11 min.)</td>
<td>2.7 (max. 23.0)</td>
<td>2.1 (max. 9.3)</td>
</tr>
</tbody>
</table>

Table 2. Deviation of lower bounds and feasible solutions from best known lower and upper bounds. Computation times were obtained on a Sun Ultra 2 with 200MHz, operating under Solaris

Additional algorithms. The figures in Table 2 are again averaged over 600 ProGen instances [KS], each with 120 jobs. The values for best known lower bounds and feasible solutions for these problem instances are maintained and available at [Pro1]. As Table 2 suggests, the average deviations from both best known lower bounds and feasible solutions are below 3%, at an average computation time of 88 seconds per instance. Note that the best known solutions have been obtained by various researchers using different local-search, as well as branch-and-bound algorithms over the years, thus computation times cannot be given. The best known lower bounds are due to Brucker and Knust [BK]. They, however, require up to 72 h computation time per problem instance, whereas the Lagrangian approach only requires a maximum computation time of 11 minutes. In view of these figures, the Lagrangian approach seems to offer an excellent tradeoff between the quality of the bounds and the computational effort to get them.

Additionally, we have tested the algorithm using a benchmark set of 25 instances of the BASF-type which are available at [Cav]. These instances are known to be notoriously difficult. Table 3 shows the results obtained with four of these sample instances of different size. The Table shows the number of jobs and tasks as well as the critical path lower bound. T denotes the time horizon, and #it is the number of iterations in the subgradient optimization. The columns LB and UB show the lower and upper bounds obtained with the Lagrangian approach described above. Although the Lagrangian approach was not able to close the still existing gaps, are the results encouraging. Given that the computation time to solve linear programming relaxations is usually prohibitive for large instances, the computation times shown in Table 3 can

<table>
<thead>
<tr>
<th>Instance</th>
<th>#jobs</th>
<th>#tasks</th>
<th>crit. path</th>
<th>T</th>
<th>LB</th>
<th>UB</th>
<th>CPU</th>
<th>#it</th>
</tr>
</thead>
<tbody>
<tr>
<td>4o.24j.B.18</td>
<td>24</td>
<td>109</td>
<td>54</td>
<td>75</td>
<td>55</td>
<td>72</td>
<td>1 sec.</td>
<td>81</td>
</tr>
<tr>
<td>6o.44j.A.18</td>
<td>44</td>
<td>224</td>
<td>75</td>
<td>128</td>
<td>88</td>
<td>121</td>
<td>6 sec.</td>
<td>169</td>
</tr>
<tr>
<td>10o.87j.A.18</td>
<td>87</td>
<td>1001</td>
<td>194</td>
<td>633</td>
<td>412</td>
<td>595</td>
<td>8 min.</td>
<td>198</td>
</tr>
<tr>
<td>10o.106j.A.18</td>
<td>106</td>
<td>1653</td>
<td>383</td>
<td>1224</td>
<td>748</td>
<td>1126</td>
<td>33 min.</td>
<td>239</td>
</tr>
</tbody>
</table>

Table 3. Results for sample instances of BASF-type with one resource type (manpower) and variable resource requirements of jobs (see Fig. 1). Computation times have been obtained on a Sun Ultra 10 with 330 MHz, operating under Solaris
still be ranked as moderate. (We refer to the comparison of computation times in Table 1, and also to [CDS+], who could solve only small problem instances due to enormous computation times.) Moreover, the feasible solutions given in Table 3 improve upon results obtained with constraint propagation by Heipcke [Hei], and compare to results with a tabu search algorithm by Cavalcante and de Souza [CD,CDS+]. It is worth to note, however, that in comparison to tabu search, the Lagrangian approach provides both a lower bound and a feasible solution at the same time.

6 Time-Critical Tasks and Maximal Time Lags

Until now, we have assumed that the temporal restrictions are given by arbitrary minimal time lags $d_{ij} \geq 0$ between any two jobs $i$ and $j$, and the sequence of tasks of a job had to be processed without interruption. An additional feature which would be 'nice to have' in some practical applications is the possibility of modelling also maximal time distances between certain jobs or tasks. The motivation is that the execution of certain tasks may be postponed in time, but must not be postponed too much, for instance because the temperature must not fall below a certain threshold. Such constraints can be easily anticipated by introducing negative time lags between jobs, or tasks, where a time lag $d_{ji} < 0$ now implies a maximal time lag of $S_j$ relative to $S_i$. Hence, so-called time windows of the form $S_i + d_{ij} \leq S_j \leq S_i - d_{ji}$ between any two jobs (or tasks) can be modeled.

Unfortunately, this has the effect that the problem of finding a feasible solution already is $NP$-complete (see, e.g., [BMR]). Although the computation of lower bounds via Lagrangian relaxation as described in Sects. 2 and 3 remains valid also if maximal time lags are present, the approach described in Sect. 5 to compute feasible solutions can no longer be applied. This is due to the fact that this approach is based upon list-scheduling algorithms which generally fail to find feasible solutions if maximal time lags exist. Hence, we have implemented a branch-and-bound algorithm for this problem (see [FMSU]). The underlying idea is that the problem is easy solvable via longest path calculations if resource-constraints are absent. Thus, the resource-constraints are relaxed, and within an enumeration tree so-called resource conflicts are resolved by introducing additional temporal restrictions. A resource conflict is a time interval where a schedule consumes more resources than available, thus violating inequalities (4). In contrast to previous approaches by Bartusch, Möhring, and Radermacher [BMR], de Reyck and Herroelen [dRH], or Schwindt [Sch2], we implemented a branch-and-bound algorithm which uses earliest start times (or release dates) to resolve a resource conflict. This has the effect that the necessary computations in every node of the enumeration tree can be realized very efficiently. We do not go into details here, but refer to [FMSU].

Also with this branch-and-bound algorithm, we could optimally solve the above described instance provided by BASF AG within less than a minute.
### Table 4. Comparison of the performance of branch-and-bound algorithm [FMSU] with previous branch-and-bound algorithms, based upon 1059 ProGen/max instances [Sch1]. Computations for [FMSU] have been obtained on a Sun Ultra 2 with 200 MHz, operating under Solaris. Computations for [dRH] and [Sch2], however, have been conducted in a different environment.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time limit</th>
<th>Optimized</th>
<th>Feasible</th>
<th>Av. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Reyck, Herroelen [dRH]</td>
<td>30 s*</td>
<td>57.5%</td>
<td>93.6%</td>
<td>10.0%**</td>
</tr>
<tr>
<td>Schwindt [Sch2]</td>
<td>30 s</td>
<td>63.7%</td>
<td>100%</td>
<td>7.0%</td>
</tr>
<tr>
<td></td>
<td>100 s</td>
<td>64.7%</td>
<td>100%</td>
<td>6.9%</td>
</tr>
<tr>
<td>Fest, Möh ring, Stork, Uetz [FMSU]</td>
<td>30 s</td>
<td>70.6%</td>
<td>100%</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>100 s</td>
<td>72.8%</td>
<td>100%</td>
<td>6.8%</td>
</tr>
</tbody>
</table>

*Corresponds to 100s on a 60MHz personal computer.

**Average deviation based upon different lower bounds.

Concerning lower bounds for the ProGen/max instances, it is worth to mention that the Lagrangian approach could improve upon the best known lower bounds for 91 instances. We refer to [MSSU3] for more details.

### 7 Final Remarks and Outlook

We have implemented and tested two different approaches to solve resource-constrained scheduling problems which arise in a typical production process at BASG AG, Ludwigshafen. On the one hand, this is a Lagrangian approach which is suited for a whole variety of different resource-constrained scheduling
Fig. 4. Screen-shot of the user interface (based on instance 100.87j.A; see Table 3). Jobs are red rectangles on the time axis, and the blue bottom line visualizes resource utilization, where dark blue indicates heavy utilization. The two curves show the lower and upper bound improvements during the course of the algorithm.

Apart from the algorithmic side, we have also implemented a user interface which allows to display both solutions as well as problem instances, and which enables the user to manipulate the problem instances on-line. Fig. 4 shows a screen-shot of the schedule visualization. In the meantime, several software companies indicated their interest to integrate the Lagrangian-based algorithm into their products, and on the account of the present project a cooperation on a related topic was established with ATOSS Software AG.

Acknowledgments. We would like to thank Josef Kallrath and Anna Schreieck from BASF AG for their cooperation, and Lars Stolletz and Carola Schaad for their contribution to our implementation of the algorithms.
References


Duty Scheduling in Public Transit*

Martin Grötschel, Ralf Borndörfer, and Andreas Löbel


Abstract. This article is about adaptive column generation techniques for the solution of duty scheduling problems in public transit. The current optimization status is exploited in an adaptive approach to guide the subroutines for duty generation, LP resolution, and schedule construction toward relevant parts of a large problem. Computational results for three European scenarios are reported.

1 Introduction

Duty scheduling is the activity of operational planning in public transport that deals with the construction of the daily shifts of driving work. Duty scheduling is a central operational issue. It has also economic significance: The average German bus company spends half of its operating budget on driver salaries, see [41,42, both articles in German]. The duty scheduling problem has been studied extensively in the OR literature, see the proceedings of the last four CASPT conferences ([56], [18], [26], and [19]) for an overview.

This article deals with column generation methods for duty scheduling. Column generation is one of the best established optimization approaches to duty scheduling. The methodology is based on techniques such as

- duty generation by constrained shortest path computations ([32], [24], [27–29]),
- large scale techniques such as sifting, core, and active set strategies ([12], [16], [15], [51]),
- LP and other acceleration techniques ([33], [23]),
- heuristics ([5], [8], [36], [54], [16], [15]),
- and other contributions (see [7], [30], [52], [21] for surveys).

Likewise, column generation modules form the optimization cores of several commercial duty scheduling systems, among them (in alphabetical order) the GENCOL optimizer of the HASTUS CREW-OPT system (see [25], [31], [22]), the PROBl solver in the CARMEN system (see [55], [2]), the TURNI system (see [34]), and our optimizer DS-OPT which is available in the BERTA and MICROBUS systems (see [14, in German] and [40, in German]).

This article discusses the three main algorithmic modules of our column generation system for duty scheduling. These modules use relaxations to

* This research has been supported with funds of the German Ministry for Research and Education, Grant No. 03-GR7ZI1. Responsibility for the contents of this article is with the authors.
speed up the identification of "improving" duties, the computation of shadow prices, and the construction of feasible schedules. The techniques are "adaptive" in the sense that they resort to the current optimization status in order to identify relevant parts of a problem. This helps to tackle large instances without ad hoc eliminations of degrees of freedom.

The article is subdivided into five parts. Section 2 lists notation and concepts on column generation and introduces an approximation algorithm that provides the basis for our duty scheduling system. Section 3 studies the constrained shortest path problem that comes up in the pricing subroutine of the algorithm. We propose an integer linear programming formulation for this problem that we solve using Lagrangean relaxation techniques. Section 4 investigates the solution of the master LP using coordinate ascent methods. Section 5 deals with the construction of feasible schedules. We describe a variable fixing heuristic that is based on scoring and probing techniques. Computational experience for three European duty scheduling scenarios is reported in Sect. 6.

2 Duty Scheduling

We list in this section notation, terminology, and concepts on duty scheduling that we will use in this paper. This notation is quite extensive. We have therefore tried to present a structured listing that is easy to reference. We do not discuss application and modelling aspects here; such information can be found in [50], [30], and [14, in German].

2.1 Terminology

The mathematical terminology follows [38]. We use the following extensions:

- We often scale by a quantity \( 1 + \delta \) and denote that by
  \[
  \tilde{x} = (1 + \delta)x. \tag{1}
  \]
- \([\cdot]_a^b : \mathbb{R} \to \mathbb{R}, \ x \mapsto \max \{a, \min \{x, b\}\}, \ [\cdot]_0^\infty = [\cdot]_0^\infty.
- \( \min_i S \) denotes the \( i \)-th minimum of a multiset \( S = \{s_1, \ldots, s_k\} \), i.e., given some ordering \( s_{j_1} \leq s_{j_2} \leq \cdots \leq s_{j_k} \), we define \( \min_i S = s_{j_i}, \ i = 1, \ldots, k \), and \( \min_i S = \infty, \ i > k \).
- \( \arg\min_{x \in D} f(x) \) denotes some optimal solution of the optimization problem \( \min_{x \in D} f(x) \), ties are broken arbitrarily. \( \arg\max \) is analogous.
- We denote arcs in digraphs by \( uv \) and \( (u, v) \); the latter notation is used in cases such as \( (m, m + 1) \). We write \( \text{tail}(uv) = u, \ \text{head}(uv) = v \).
- We sometimes interpret matrices as sets of column vectors and use notation such as \( (a_{ij}) \cup b \) to add a column \( b \) to a matrix \( (a_{ij}) \) if an identical column is not already present.

We use Latin symbols in a network context, and Greek ones in a set partitioning context.
2.2 Network Model

The duty scheduling problem can be formulated in terms of a network model. This model involves an acyclic digraph \(D = (V, A)\), the _duty scheduling network_.

We number the nodes \(V = \{0, 1, 2, \ldots, m, m + 1\}\). Nodes \(M = \{1, \ldots, m\}\) are the _duty elements_ or _tasks_. They correspond to indivisible units of work that arise from cutting vehicle blocks at relief points. Two tasks \(u\) and \(v\) are joined by an arc or _link_ if the same driver can conduct \(v\) immediately after \(u\). Nodes 0 and \(m + 1\) serve as universal _source_ and _sink_ nodes in \(D\). They represent the start and the end of every duty. Source and sink are linked with _sign-on_ and _sign-off_ arcs \(\delta^+(0) = \{(0, 1), (0, 2), \ldots, (0, m + 1)\}\) and \(\delta^-(m + 1) = \{(0, m + 1), (1, m + 1), \ldots, (m, m + 1)\}\), respectively. Note that \(\delta^-(0) = \emptyset = \delta^+(m + 1)\) because \(D\) is acyclic.

_Duties_ correspond to directed \((0, m + 1)\)-paths in this duty network, but not every path is a duty. Legality and cost of a duty are determined by the rules of so-called _duty types_. Typical duty types are straight, split, and part time duties. A special duty type are the _trippers_, duties that contain exactly one duty element. A precise model of a duty type will be given in Sect. 3.

The _duty scheduling problem_ (DSP) is to find a minimum cost set of duties (each duty valid for at least one given duty type) that covers all tasks exactly once.

2.3 Set Partitioning Model

Let \(\hat{J} = \{1, \ldots, \hat{n}\}\) denote the set of all feasible duties, let \(\hat{\xi}_j, \ j \in \hat{J}\), be a 0/1 decision variable for the inclusion/exclusion of duty \(j\), let \(\hat{\omega} \in \mathbb{R}_{\hat{n}}^+\) be a vector of _nonnegative_ duty costs, and let finally \(\hat{\Phi} \in \{0, 1\}^{m \times \hat{n}}\) be the task-duty incidence matrix. The duty scheduling problem can be formulated as a _set partitioning problem_

\[
\text{(SPP)} \quad \min \ \hat{\omega}^T \hat{\xi}, \quad \hat{\Phi} \hat{\xi} = \textbf{1}, \quad \hat{\xi} \in \{0, 1\}^\hat{n}. \tag{2}
\]

The constraints \(\hat{\Phi} \hat{\xi} = \textbf{1}\) are the _task partitioning constraints_; they stipulate that each task is covered by exactly one duty. We associate with SPP the following LPs:

\[
\text{(MLP)} \quad \min \ \hat{\omega}^T \hat{\xi}, \quad \hat{\Phi} \hat{\xi} = \textbf{1}, \quad \hat{\ell} \leq \hat{\xi} \tag{3}
\]
\[
\text{(RMLP)} \quad \min \ \omega^T \xi, \quad \Phi \xi = \textbf{1}, \quad \ell \leq \xi \tag{4}
\]
\[
\text{(RDLP)} \quad \max \ \pi^T \eta + \omega^T \ell, \quad \pi^T \Phi \leq \omega^T. \tag{5}
\]

We further associate with any vector \(\pi \in \mathbb{R}^m\) the _pricing problem_

\[
\text{(PRICE)} \quad \exists \ j \in \hat{J} : \quad \overline{\omega}_j = \hat{\omega}_j - \pi^T \hat{\Phi}_{j, \cdot} < 0. \tag{6}
\]

MLP is an _LP relaxation_ of (2), called the _master LP_. Its special form allows fixings of variables by a 0/1 vector of lower bounds \(\hat{\ell} \in \{0, 1\}^\hat{n}\). A _restricted_
primal master LP (RMLP) and its dual (RDLP) are obtained from MLP by restriction to some subset of variables $J = \{1, \ldots, n\} \subseteq \{1, \ldots, \hat{n}\}$. We shall use the following notation and conventions:

- $\hat{\cdot}$ is supposed to indicate the “closure” $\hat{J} = \{1, \ldots, n\} \subseteq \{1, \ldots, \hat{n}\}$.
- $\hat{\Phi}_\ell \leq 1$, $\hat{\Phi}_\ell \leq 1$, i.e., we always assume pairwise compatible fixings.
- $\hat{\eta} = 1 - \hat{\Phi}_\ell$ and $\eta = 1 - \Phi_\ell$ denote the incidence vector of the residuum of rows not covered by fixed variables.
- An MLP does never contain a zero column $\Phi_j = 0$.
- Any MLP contains a tripper duty of (high) cost $L \in \mathbb{R}_+$ for each task $i$, i.e.,
  \[ \forall i \in M : \exists j \in \hat{J} : \hat{\omega}_j = L \quad \text{and} \quad \hat{\Phi}_j = e_i. \quad (7) \]
- $\hat{\nu} = \nu(\hat{\omega}, \hat{\Phi}, \hat{\ell})$ and $\nu(\omega, \Phi, \ell)$ denote the optima of MLP and RMLP (or RDLP), respectively. Note that assumption (7) guarantees finiteness and well-definition.
- $\hat{\omega}_j = \omega_j - \pi^T \hat{\Phi}_j$ is the reduced cost associated with duty $j$ (and $\pi$).

### 2.4 Pseudocode

We list in this subsection the data types of a C type pseudocode that we use to describe algorithms:

- `typedef double pos;`  
  Holds a nonnegative real number.
- `typedef struct { int |A|; int * x } path;`  
  Holds a vector $x \in \{0, 1\}^A$.
- `typedef struct {int m; double * \pi; } dual;`  
  Holds a vector $\pi \in \mathbb{R}^m$ (and later $\pi \in \mathbb{R}^R$).
- `typedef struct {int n; pos * \xi; } primal;`  
  Holds a vector $\xi \in \mathbb{R}^n$.
- `typedef struct {int m; int n; int * \Phi.cnt; int * \Phi.ind; } matrix;`  
  Holds a 0/1 matrix $\Phi \in \mathbb{R}^{m \times n}$ in column major format.
- `typedef struct { pos \gamma; dual \phi; } col;`  
  Holds objective $\gamma$ and task incidence vector $\phi$ of a duty.
- `typedef struct { primal \omega; matrix \Phi; primal \ell; dual \pi; } rdlp;`  
  Holds objective $\omega$, task incidence matrix $\Phi$, lower bounds $\ell$, and dual multipliers $\pi$ of an RDLP.

### 2.5 Column Generation Method

Figure 1 describes a column generation algorithm for the solution of the MLP (3). This algorithm, denoted by cgen, is approximative with quality control parameters $\varepsilon > 0$ and $\delta > 0$. Three subroutines, `aug`, `price`, and `test`, are called in a loop. `aug` is an approximate RDLP solver, `price` is an approximate pricing routine, and `test` is an approximation control predicate. We stipulate the following properties for these subroutines:
Fig. 1. cgen column generation algorithm

- dual aug( rdlp \{\omega, \Phi, \ell, \pi\} );
  aug must produce for an arbitrary, but fixed input RDLP \{\omega, \Phi, \ell, \pi\} an
  iteration sequence \((\pi_k)_{k \in \mathbb{N}_0}\) by means of the recursion
  \[
  \pi_0 = \pi \quad \text{and} \quad \pi_k = \text{aug}\{\omega, \Phi, \ell, \pi_{k-1}\}, \quad k \in \mathbb{N}.
  \] (8)
  This sequence is required to have the following properties:
  \[
  (\text{AUG1}) \quad \limsup_{k \to \infty} \pi_k^T \Phi \leq \omega^T \]
  \[
  (\text{AUG2}) \quad \lim_{k \to \infty} \pi_k^T \eta + \omega^T \ell = \nu(\omega, \Phi, \ell).
  \] (9, 10)
  These conditions formulate feasibility and optimality of the sequence \((\pi_k)\)
  in the limit. Note that \(\nu(\omega, \Phi, \ell) < \infty\) implies that every limit point of
  \((\pi_k)\) is optimal.
  Conditions (9) and (10) allow aug implementations as subgradient algo-
  rithms, as well as spacer step derivatives of such algorithms, in addition
  to simplex and interior point implementations.
- col price( rdlp \{\omega, \Phi, \ell, \pi\}; pos \delta );
  If the \(\delta\)-pricing problem
  \[
  (\delta\text{-PRICE}) \quad \exists \left(\frac{\gamma}{\phi}\right) \in \left(\frac{\omega^T}{\Phi}\right): \bar{\gamma} = (1 + \delta)\gamma - \pi^T \phi < 0,
  \] (11)
  has a solution, price must return such a column; if no such \(\delta\)-negative
  (reduced cost)column exists, price can return an arbitrary column. \(\delta\)-
  PRICE requires only the identification of duties of “significantly” nega-
  tive reduced cost; the problem is hence a weakening of the “strict” pricing
  problem (6).
  Section 3 discusses such an approximate pricing algorithm.
- \textbf{int} \texttt{test} ( \textup{rdlp} \{\omega, \Phi, \ell, \pi\}; \textup{pos} \varepsilon; \textup{pos} \delta; );

\texttt{test} returns 1 if the following two conditions, together referred to as \(\varepsilon\delta\)-\textit{optimality}, hold, and 0 otherwise (recall (1) for the definition of \(\tilde{\omega}\)):

\begin{align*}
(\delta\text{-feasibility}) & \quad \pi^T\Phi \leq \tilde{\omega} \\
(\varepsilon\text{-optimality}) & \quad (1 + \varepsilon)\pi^T\eta \geq \nu(\omega, \Phi, \ell) - \omega^T\ell
\end{align*}

\texttt{test} can be implemented by LP techniques.

It is not hard to see that the \(\varepsilon\) and \(\delta\)-accuracies of the subroutines carry over to the entire column generation algorithm.

\textbf{Proposition 1.} Let \(\eta_k = 1 - \Phi_k\ell_k\). Algorithm \texttt{cgen} has these properties:

i) \texttt{cgen} terminates after a finite number of iterations \(\kappa\).

ii) \(\hat{\nu} - \tilde{\omega}^T\hat{\ell} \leq (1 + \varepsilon)(\pi_k^T\eta_k)\).

iii) \(\hat{\nu} - \tilde{\omega}^T\hat{\ell} \geq (1 + \delta)^{-1}\pi_k^T\eta_k\).

\textbf{Proof.} i) RDLP is a sub-LP of an LP of finite dimension, namely, the dual of MLP. Hence, line 10 can only see a finite number of genuine additions of columns. From that point on, \texttt{cgen} merely iterates \texttt{aug}. This produces a sequence of multipliers that approaches the set of optimal solutions of RDLP, see (9) and (10). When (12) and (13) are satisfied, \texttt{test} breaks the main loop and \texttt{cgen} terminates at some iteration \(k = \kappa\).

ii) \((1 + \varepsilon)(\pi_k^T\eta_k) \geq \nu(\omega_k, \Phi_k, \ell_k) - \omega_k^T\ell_k \geq \nu(\tilde{\omega}, \tilde{\Phi}, \tilde{\ell}) - \tilde{\omega}^T\tilde{\ell} = \hat{\nu} - \tilde{\omega}^T\hat{\ell}\).

The first inequality follows from (13), the second holds because each column is a constraint in RDLP.

iii) \((1 + \delta)(\hat{\nu} - \tilde{\omega}^T\hat{\ell}) = (1 + \delta) \max_{\pi^T\Phi \leq \omega^T} \pi^T\eta = \max_{\pi^T\Phi \leq (1 + \delta)\omega^T} \pi^T\eta \geq \pi_k^T\eta_k\).

The final inequality holds because \texttt{test} guarantees \(\pi_k^T\Phi_k \leq (1 + \delta)\omega_k^T\) for all RDLP columns at iteration \(\kappa\), see (12), while \texttt{price} takes responsibility for \(\pi_k^T\Phi \leq (1 + \delta)\omega^T\) for all remaining columns, see (11).

Proposition 1 justifies the use of approximation algorithms in the LP resolution and pricing subroutines of a column generation algorithm. Such approximate procedures can help to bypass algorithmic bottlenecks in the subproblems.

Our system DS-OPT implements a \textit{heuristic} \texttt{cgen} derivative:

\begin{itemize}
  \item \texttt{aug} is implemented as a dual ascent heuristic as described in Sect. 4.
  \item \texttt{price} heuristically limits the search space, see Sect. 3.
  \item \texttt{test} terminates \texttt{cgen} if the relative progress \((\pi_{k+l}^T\eta_{k+l} - \pi_k^T\eta_k)/\pi_k^T\eta_k\) made in some last \(l\) iterations falls below \(\varepsilon\).
\end{itemize}

We close this subsection with the remark that Proposition 1 shows a way to compute a \textit{global lower bound} for the entire master LP, and hence for the duty scheduling problem itself. The algorithmic challenge in the computation of such a bound is the solution of the \(\delta\)-pricing problem (11). The potential of our methods with respect to the computation of lower bounds will be the subject of future research.
2.6 Duty Scheduling Method

The approximate solution of MLP is only a first step in duty scheduling. What we really want is an "acceptable" feasible integer solution.

Figure 2 gives pseudocode for an LP plunging heuristic to produce such a solution. This \texttt{dsospt} procedure is the top level routine of our duty scheduling system DS-OPT.

\begin{verbatim}
1 rdlp dsospt( rdlp \{\omega,Phi,ell,pi\}, pos \epsilon, pos \delta, pos \tau )
2 {
3   int k ← 0;
4   int j_k;
5   rdlp \{\omega_k,Phi_k,ell_k,pi_k\} ← \{\omega,Phi,ell,pi\};
6   dual \eta_k ← 1 - Phi_k*ell_k;
7   double \Theta ← -\infty;
8   while (Phi_k^T*ell_k < 1) do {
9      k ← k + 1;
10     if (\pi_k^T*eta_k + \omega_k^T*ell_k > \Theta) {
11        \{\omega_k,Phi_k,ell_k,pi_k\} ← cgen(\{\omega_{k-1},Phi_{k-1},ell_{k-1},pi_{k-1}\},\epsilon,\delta);
12        \Theta ← (1 + \tau)(\pi_k^T*eta_k + \omega_k^T*ell_k);
13     };
14     j_k ← chuzc(\{\omega_k,Phi_k,ell_k,pi_k\});
15     ell_k ← ell_{k-1} + \epsilon*j_k;
16     eta_k ← eta_{k-1} - (Phi_k)*j_k;
17     pi_k ← aug(\{\omega_k,Phi_k,ell_k,pi_k\});
18   };
19   return \{\omega_k,Phi_k,ell_k,pi_k\};
20 }
\end{verbatim}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{dsospt.png}
\caption{\texttt{dsospt} duty scheduling heuristic}
\end{figure}

d\texttt{dsospt} calls the subroutines \texttt{cgen} and \texttt{chuzc} in a loop. \texttt{chuzc} is a routine to select a variable that is subsequently fixed to 1 in lines 15 to 16 of \texttt{dsospt}. The calls to \texttt{cgen} are controlled by a branch-and-generate scheme involving parameters \( \tau > 0 \) and \( \Theta \). These subroutines work as follows:

- \texttt{int chuzc( rdlp \{\omega,Phi,ell,pi\} );}
  \texttt{chuzc} returns an \( \ell \)-compatible column \( j \), i.e., \( \Phi(\ell + e_j) \leq 1 \).
  An implementation will be described in Sect. 5.
- Branch-and-Generate Scheme.
  \texttt{dsospt} monitors a trust region
  \[ \pi_k^T*eta_k + \omega_k^T*ell_k \leq \Theta \] (14)
  for the RDLP objective \( \pi_k^T*eta_k + \omega_k^T*ell_k \) at hand. As long as this value stays below \( \Theta \), RDLP is considered as an acceptable approximation of the
global MLP and \texttt{cgen} is bypassed. Such a “fast iteration” merely fixes a variable and adjusts the RDLP solution by a call to \texttt{aug} in line 17. When the objective goes above \( \Theta \), this is taken as an indication that the variable fixes have \textit{potentially} changed the RDLP to such an extent that it does no longer approximate the (also changed) global MLP well. In this situation, \texttt{cgen} is rerun to update the RDLP and a new trust region is established up to \( \tau \) percent above the updated objective value.

We have learned about such a control scheme from a talk of Marsten, who termed it \textit{branch-and-generate} (BANG) and implemented it in the DOC (Delta Optimizer for Crews) system. BANG does, however, not seem to have been documented in the literature.

We remark that an efficient \texttt{dsopt} implementation will safeguard against the generation of incompatible columns \( (\gamma_k, \phi_k^T) \), i.e., \( \phi_k^T \delta_k \ell_k \neq 0 \) in the \texttt{cgen} column generation routine. A convenient, but notationally not always elegant way is to set \( \pi_i = -\infty \) for \( \eta_i = 0 \).

Our system DS-OPT is not designed to produce optimal solutions to duty scheduling problems. The focus is on the ability to process large scenarios in a reasonable way. In particular, we aim to exploit all degrees of freedom in the duty network and in the duty types. DS-OPT implements this strategy using approximative and, where we think it expedient, heuristic techniques.

\section{Constrained Shortest Paths}

This section deals with the solution of the \( \delta \)-pricing problem (\ref{eq:delta-pricing}). We propose an integer linear programming formulation for this problem that we tackle with Lagrangean relaxation techniques. It turns out that such a relaxation gives rise to lower bounds that can be used to speed up enumerative pricing algorithms.

We model the pricing problem in terms of acyclic constrained shortest path problems (ACSPs). The different constraints of different duty types are handled using separate ACSPs. We will discuss in the remainder of this section the treatment of a single duty type.

\subsection{Integer Programming Model}

Denote by \( x \in \{0,1\}^A \) the incidence vector of a directed \((0,m+1)\)-path in the duty scheduling network \( D = (V,A) \) representing a duty, by \( c \in \mathbb{R}^A \) a vector of costs associated with the arcs, and by \( N \in \{-1,0,1\}^{V\times A} \) the node-arc incidence matrix of \( D \). Let further \( W \in \mathbb{R}^{R\times A} \) be a matrix that records in each entry \( w_{ra} \) the consumption a so-called resource \( r \) on traversal of the arc \( a \), let \( b_r \) be a \textit{goal} or best resource consumption for a duty, let \( s_r = (s_r^+, s_r^-) \) be a pair of slack and surplus variables, respectively, that gather deviations from this goal value, let \( u_r = (u_r^+, u_r^-) \in (\mathbb{R}_+ \cup \{\infty\})^2 \) be upper limits for such deviations, and let \( d_r = (d_r^+, d_r^-) \in \mathbb{R}_+^2 \) be nonnegative
objective penalties associated with the slack and surplus variables. (We do not discuss the treatment of bonuses.) We remark that we use this resource constraint model also to express “infeasible path constraints”, see below. Let finally

\[ c^T x + d^T s = c^T x + d^+ T s^+ + d^- T s^- \]  

be the genuine cost associated with a vector \((x, s) \in \mathbb{R}^m \times \mathbb{R}^R \times \mathbb{R}^R\), let \(\pi \in \mathbb{R}^m\) be some vector of shadow prices, and consider a transformed cost

\[ \tilde{c}^T x + \tilde{d}^T s = (1 + \delta)(c^T x + d^T s) - \sum_{j=1}^{m} \sum_{ij \in A} \pi_j x_{ij} \]  

that arises from the genuine one by a scaling operation (recall (1)) and a subsequent subtraction of certain shadow prices.

We propose to model the single duty type pricing problem as an ACSP subject to linear side constraints and linear objective penalties:

\[
\begin{align*}
(ACSP) & \quad \min \tilde{c}^T x + \tilde{d}^T s \\
& \quad \text{(ACSP)} \quad \min \tilde{c}^T x + \tilde{d}^T s \quad (17) \\
N x & = e_{m+1} - e_0 \quad (18) \\
W x + s^+ - s^- & = b \quad (19) \\
0 & \leq s \leq u \quad (20) \\
x & \in \{0, 1\}^A. \quad (21)
\end{align*}
\]

We list two conventions that will be used in the following discussion:

- \(s(x) = (s^+(x), s^-(x)) = ([b - W x]^+, [W x - b]^+)\).
- \(M(x) = \{1 \leq i \leq m : x(\delta^+(i)) = 1\}\).

(18) are flow conservation constraints that define an \((0, m + 1)\)-path. (19) and (20) are called resource constraints. Note that our assumption \(d \geq 0\) sets the penalties automatically to the right values \(s(x)\). Let finally \((x, s(x))\) be an ACSP solution of genuine cost \(\gamma = c^T x + d^T s(x)\), and let \(\phi = \chi^{M(x)}\) be the task incidence vector of the visited duty elements. Consideration of the transformed objective (16)/(17)

\[ \tilde{c}^T x + \tilde{d}^T s = (1 + \delta)(c^T x + d^T s) - \sum_{j=1}^{m} \sum_{ij \in A} \pi_j x_{ij} = (1 + \delta)\gamma - \pi^T \phi \]

shows that ACSP is equivalent to the \(\delta\)-pricing problem (11) for a duty type that is given by the constraints (19) and (20) and the genuine objective (15).

Linear resource constraints are not only versatile modelling tools. Perhaps even more important, such constraints arise naturally in duty scheduling applications. We list some important types in a simplified format that can easily be transformed into the general form (19) and (20):
- Max constraints $W_r \cdot x \leq b_r$
  
  *Max constraints* can be used to stipulate upper limits on the consumption of resources such as driving time, working time, duty time, shift time, number of pieces of work, number of parts of work, etc.

- Min constraints $W_r \cdot x \geq b_r$
  
  *Min constraints* can set lower limits on the same quantities. They can also be used to enforce absolute and, in the form $W_r \cdot x \geq q \cdot W_r \cdot x$, $0 \leq q \leq 1$, relative minimum break times in a duty.

- Opt constraints $W_r \cdot x + s_r^+ - s_r^- = b_r$ with penalties $d_r^+ s_r^+ + d_r^- s_r^-$
  
  *Opt constraints* and the associated objective penalties can model linear overtime bonuses and the like costs.

- Infeasible path constraints $x^T \cdot \chi = |P| - 1$
  
  Some difficulties come up in the treatment of "nonlinear" rules that govern, e.g., break positions in a duty. Such constraints can be fitted within a linear framework using *infeasible path constraints* (IPCs) that rule out individual infeasible paths. We call a set of IPCs that represent a nonlinear rule an *IPC rule complex*. Such constraints must be treated implicitly. IPCs were introduced in [4] in the context of asymmetric TSPs with time windows.

Typical duty scheduling problems feature duty types with about 10 different explicit linear resources and half a dozen implicit IPC rule complexes on break positions, sign-on and sign-off issues, the construction of composite duties with several parts, and certain compensations.

The ACSP is well studied in the literature, see [46] for a survey. The problem is $\mathcal{NP}$-hard already for a single resource constraint, see [37, A2.3, ND30]. For any fixed number of resources, fully polynomial approximation schemes exist, see [53]. Pseudopolynomial algorithms have been developed and successfully used in practical applications, see [24] and others, including penalty treatment, see [25]. Enumerative approaches using Lagrangean lower bounding techniques have been studied in [39], [48,49], and [9]. A computational comparison of different methods can be found in [46]; this reference gives also a combinatorial algorithm of low complexity to solve the LP relaxation associated with ACSP for the case of a single resource constraint.

However, as far as we know, the model ACSP has not played a role in the context of duty scheduling. The duty scheduling literature focuses on alternative and well-known *time window formulations*, see [32], [24], [27–29].

### 3.2 Lagrangean Lower Bound

We follow now [39] and [9] in using a Lagrangean relaxation of the ACSP to derive a class of lower bounds that will be used to establish an efficient backtracking criterion in an enumerative ACSP algorithm; this idea has also been used in [46] to develop essentially the same algorithm.
The **Lagrangean shortest path relaxation** (LSP) that we use is obtained by transferring the resource constraints (19) into the objective:

\[
\begin{align*}
\text{(LSP)} \quad \max_{y \in \mathbb{R}^2} & \quad P(y) + Q(y) + y^T b \\
\text{(SP)} \quad P(y) &= \min \ (\tilde{c}^T - y^T W) x, \quad Nx = e_{m+1} - e_0, \quad x \geq 0 \\
\text{(BP)} \quad Q(y) &= \min \ (\tilde{d}^+ - y)^T s^+ + (\tilde{d}^- + y)^T s^-, \quad 0 \leq s \leq u.
\end{align*}
\]

LSP decomposes into an acyclic shortest path problem SP and a simple linear program BP over a box. BP has a closed formula solution, e.g.,

\[
s^* = s^*(y) = \frac{1}{2} \left( I - \text{diag} \left( \text{sign}(\tilde{d}^+ - y), \text{sign}(\tilde{d}^- + y) \right) \right) u.
\]

An optimal solution \( x^* = x^*(y) \) of SP can be obtained combinatorially using, e.g., the reaching algorithm as proposed in [1]. We are also interested in an optimal solution \( h^* = h^*(y) \) of the dual

\[
\text{(DP)} \quad \max \ h_{m+1} - h_0, \quad h^T N \leq \tilde{c}^T - y^T W.
\]

Such a solution can be interpreted in terms of distance labels \( h^*_v \) that give, for each node \( v \), the minimum distance \( h^*_v - h^*_0 \) from the source with respect to the objective \( \tilde{c}^T - y^T W \). These values underestimate, in particular, the distances along all constrained paths. Combining this bound with the solution of BP suggests to consider the following **Lagrangean distance labels**:

\[
g^*_v(y) = g^*_v(y) = h^*_v(y) - h^*_0(y) + Q(y) + y^T b, \quad v \in V.
\]

**Lemma 2.** Let \( x^1 \) and \( x^2 \) be the incidence vector of a \((0,v)\)- and a \((v,m+1)\)-path in \( D \), respectively. Suppose that \( x^1 + x^2 \) is feasible for ACSP. Let \( y \in \mathbb{R}^R \) be any vector of Lagrangean multipliers for LSP and \( g^*(y) \) be the associated Lagrangean distance labels (27). Then:

\[
g^*_v(y) + (\tilde{c}^T - y^T W)x^2 \leq \tilde{c}^T (x^1 + x^2) + d^T s(x^1 + x^2).
\]

Lemma 2 is useful in an algorithm that constructs \( \delta \)-negative paths “backwards” starting from the sink, working toward the source. (28) can be used in such an algorithm to prune the search whenever \( g^*_v(y) + (\tilde{c}^T - y^T W)x^2 \geq 0 \).

### 3.3 Lagrangean Distance Computation

The quality of the distance labels \( g^*(y) \) depends on the identification of suitable multipliers \( y \). This task can be tackled by a subgradient algorithm. The work per iteration consists in solving SP and BP.

We state on this occasion a characterization of the subdifferential \( \partial(y) \), which is readily available in this application, see [10, Prop. 6.1.2].
Corollary 3. Let \( y \in \mathbb{R}^R \) be a vector of Lagrangean multipliers for LSP. Let \( h^* \) be the solution of the associated shortest path problem DP. Denote by \( \overline{A} = \{a \in A : \overline{c}_a - y^T \overline{W}_a > h^* T N_a \} \) the set of all arcs that can not be contained in an optimal solution to SP. Then:

\[
\partial(y) = \text{conv}\{b - Wx - s^*^+ + s^*- : N_x = e_{m+1} - e_0, \quad x \geq 0, \quad x(\overline{A}) = 0\}.
\]

1 dual lsp ( dual y )
2 {
3     primal x;
4     dual s^+, s^-;
5     int g;
6     for (;;) {
7         g ← [select \( r \in R \) with \( W_r x + s^+_r - s^-_r - b_r \neq 0\)];
8         x ← x^*(y);
9         s ← s^*(y);
10        yo ← yo + \theta([\text{stepsize control}]) \text{sign}(W_r x + s^+_r - s^-_r - b_r);
11        if [\text{convergence}] return y;
12     }
13 }

Fig. 3. lsp Lagrangean distance computation

It turns out, however, that a coordinate ascent method is good enough in practice. Pseudocode for such an algorithm, which adjusts a single multiplier in each iteration, is given in Fig. 3.

3.4 Constrained Shortest Path Algorithm

Figure 4 gives pseudocode for a depth first search enumeration algorithm for the ACSP that uses the Lagrangean distance labels (27) as a backtracking criterion. The distance labels are derived from a relaxation of the ACSP that one obtains by simply ignoring all IPCs. Starting from the sink, the algorithm computes recursively all possible \((0, m+1)\)-paths. A backtrack is done in the following cases:

- Line 10.
  The path \( x \) violates an IPC. Checking an entire duty for IPC violations is in general not a problem, because the rules are made for this purpose. Detecting inevitable violation as early as possible in a partial duty is, however, certainly difficult. We use heuristic criteria that would take a detailed description of data structures in order to be discussed.

- Line 11.
  The source is reached, an improving duty is found (line 12), or not.
col price( rdlp \{\omega,T,\ell,\pi\}, dual y )
{
  dual g* \leftarrow g^*(y);
  return dfs ( 0, g^*, (+\infty,0) );
}

col dfs( path x, dual g^*, col (\gamma,\phi) )
{
  int v \leftarrow tail(x);
  if (x [violates an IPC]) return (\gamma,\phi);
  if (v = 0) {
    if (\tilde{c}^T x + \tilde{d}^T s(x) < \gamma)
    
    \begin{align*}
      (\gamma,\phi) & \leftarrow (\tilde{c}^T x + \tilde{d}^T s(x), \chi^M(x));
    \end{align*}
  
    return (\gamma,\phi);
  }
  if ([search limit exceeded]) return (\gamma,\phi);
  forall (u \in \gamma^{-}(v)) {
    if (\tilde{c}^T x + \tilde{c}_{uv} + g^*_u \geq \max\{\gamma,0\}) continue;
    \begin{align*}
      (\gamma,\phi) & \leftarrow dfs(x + e_{uv}, g^*, (\gamma,\phi));
    \end{align*}
  }
  return (\gamma,\phi);
}

Fig. 4. price column generator (single duty type)

- Line 16.
  Some search limit is exceeded. Such search limits can help to adjust the performance of an implementation to a particular scenario. There is, however, another use that is even more important. Our dsOpt implementation differs from the one in Fig. 4 in generating not only one \(\delta\)-negative duty, but bunches of several thousands of such duties. We use search limits as a heuristic means to produced "diverse" paths in an attempt to speed up the overall convergence of the algorithm. Diversification techniques of this type have been mentioned in the literature, e.g., in [23].

- Line 18.
  The Lagrangean lower bounding criterion shows that the path \(x\) can not be extended into a \(\delta\)-negative reduced cost duty.

We remark that we also use the Lagrangean distance labels for preprocessing purposes, see [9] for details.

4 Coordinate Ascent

We study in this section the \(\varepsilon\delta\)-optimal solution (recall (13) and (12)) of RDLP. Such a solution can be obtained using sophisticated LP codes. Large
RDLPs beyond 1,000 rows are, however, often hard to solve leading to LP performance problems as frequently reported in the literature, see, [2, p. 241/242], see also [45], [43], [13], and [33].

The search for alternatives focuses on subgradient algorithms and on dual ascent methods that are both based on Lagrangean relaxations. Subgradient algorithms have been used in [16], [15], and [34]. Dual ascent methods have been studied, among others, in [5], [8], [36], and [54]. See also [35, p. 9] for a survey on dual ascent applications in general.

4.1 Coordinate Ascent for RDLPs

We recall in this subsection the basics of coordinate ascent in an RDLP context. Starting point is the following Lagrangean relaxation of the RDLP:

\[(LRDLP) \quad \max_{\pi \in \mathbb{R}^m} \min_{\ell} A(\pi, \xi), \quad \ell \leq \xi \leq 1 \]\n
where

\[A(\pi, \xi) = \omega^T \xi - \pi^T (\Phi \xi - 1) = \omega^T \xi + \pi^T 1. \]  

For fixed \( \pi \), LRDLP has the closed formula solution(s)

\[\xi^*_j = \xi^*_j(\pi) = \begin{cases} 1, & \text{if } \omega_j < 0 \\ \ell_j, & \text{if } \omega_j > 0 \end{cases} \quad \text{if } \xi^*_j \leq 1 \]

\[\xi^*_j = \xi^*_j(\pi) = \begin{cases} \ell_j, & \text{if } \xi^*_j = 0 \end{cases} \quad \text{if } \xi^*_j = 0. \]

For given \( \pi \), the \textit{line search problem} along the \( i \)-th coordinate is

\[(LS) \quad \max_{\alpha \in \mathbb{R}} \min_{\xi} A(\pi + \alpha \cdot e_j, \xi), \quad \ell \leq \xi \leq 1 \]

\[= \max_{\alpha \in \mathbb{R}} \pi^T 1 + \alpha + \min_{\xi} (\omega^T - \alpha \cdot \Phi_i) \xi, \quad \ell \leq \xi \leq 1. \]

LS has also a closed formula solution

\[\alpha^* \in \begin{cases} [\min_{j \in \Phi_i} \omega_j, \min_{j \in \Phi_i} \omega_j], & \text{if } \eta_i = 1 \\ (-\infty, \min_{j \in \Phi_i} \omega_j], & \text{if } \eta_i = 0. \end{cases} \]

\(\alpha^*\) can be determined easily by a sweep over the nonzeros of the \( i \)-th row of the matrix \( \Phi \).

The basic coordinate ascent method simply iterates such line searches along the coordinates. It is well known that this method is not globally convergent, because it can get stuck at a nonoptimal point from which it is impossible to ascend along any coordinate axis, see, e.g., [10, § 6.3.3].
4.2 Boxstep

A problem that we have observed in coordinate ascent computations is that some of the multipliers tend to take unreasonably large values in early iterations that are hard to readjust later on. This phenomenon has been documented before in a simplex context in [33]. We use boxstep, a framework due to [44], to counterbalance this effect.

Boxstep is a method to optimize a concave and upper semicontinuous function $\nu : D \to \mathbb{R}$ on a compact domain $D \subseteq \mathbb{R}^m$

\[
\text{(GLOBAL)} \quad \max_{\pi \in D} \nu(\pi).
\]

The method associates with each point $\pi_0 \in D$ a local problem

\[
\text{(LOCAL)} \quad \max \nu(\pi), \quad \pi \in D, \quad \|\pi - \pi_0\|_\infty \leq \beta
\]

and iterates the recursion

\[
\pi_{k+1} = \arg\max \nu(\pi), \quad \pi \in D, \quad \|\pi - \pi_k\|_\infty \leq \beta,
\]

until $\nu(\pi_{k+1}) \leq \nu(\pi_k) + \varepsilon$. $\beta > 0$ is a parameter that controls the size of the domain in the local problem, and $\varepsilon > 0$ is a termination parameter. It has been proven in [44] that boxstep terminates in a finite number of iterations $\kappa$ with a solution $\pi_\kappa$ such that $\nu(\pi_\kappa) \geq \max \nu(D) - 2\varepsilon \Delta / \beta$, where $\Delta = \text{diam} D = \max_{x,y \in D} \|x - y\|_2$ denotes the diameter of $D$. This result carries over to our situation as follows.

**Corollary 4.** Suppose RDLP contains the “tripper” duties $\left(\frac{L}{e_i}, \Phi\right)$, $i = 1, \ldots, m$, of cost $L \in \mathbb{R}_+$. Then:

i) RDLP has an optimal solution $\pi^* \in \left[-(m-1)L, L\right]^m$.

ii) Boxstep finds, for any starting point $\pi_0 \in \left[-(m-1)L, L\right]^m$, in finite time an RDLP solution $\pi_\kappa$ such that $\pi^T_\kappa \eta \geq \nu(\omega, \Phi, \ell) - 2\varepsilon m^{3/2}L/\beta$.

**Proof.** i) Note that $\pi^T I \leq L \mathbb{1}^T$ because of the trippers. The further restriction $\pi \geq -(m-1)L \mathbb{1}$ does not change the optimum:

\[
\max \pi^T \mathbb{1} + \lambda \ell, \quad \pi^T \Phi + \lambda \ell = \omega^T, \quad -\pi^T \leq -(m-1)L \mathbb{1}^T, \quad \lambda \geq 0
\]

\[
= \min \omega^T \xi, \quad (m-1)L \mathbb{1}^T \mu, \quad \Phi \xi - \mu = \mathbb{1}, \quad \xi \geq \ell, \quad \mu \geq 0
\]

\[
= \min \omega^T \xi, \quad \Phi \xi = \mathbb{1}, \quad \ell \leq \xi \leq \ell.
\]

The second equation follows by replacing overcovers with trippers.

ii) $\text{diam}\left[-(m-1)L, L\right]^m = \sqrt{m \cdot m^2 L^2} = m^{3/2}L$. 

Fig. 5. **aug** coordinate ascent algorithm

### 4.3 Coordinate Ascent Algorithm

Figure 5 gives pseudocode for a coordinate ascent heuristic that combines a coordinate search algorithm with a boxstep steplength control in line 9. The solution formula for the line search problem in line 8 contains a control parameter $0 < \vartheta < 1$ that is used for performance adjustments.

### 5 Variable Fixing

We discuss in this section the heuristic fixing of duties in the branch-and-generate phase of our algorithm. The fixings are derived in two stages from the Lagrangean relaxation LRDLP associated with the current restricted dual master LP. The first stage selects a set of *candidates* using scoring techniques. The second stage selects a variable from this candidate set by Lagrangean probing. Figure 6 gives the pseudocode for such a variable selection scheme.

To explain the routine, recall the Lagrangean relaxation

$$(LRDLP) \quad \max_{\pi \in \mathbb{R}^m} \min_{\xi} \omega^T\xi - \pi^T(\Phi\xi - \mathbb{1}), \quad \ell \leq \xi \leq \mathbb{1}. \quad (37)$$

Let $\pi \in \mathbb{R}^m$ be some vector of multipliers and $\xi^* = \xi^*(\pi)$ (cf. (31)) be the associated primal solution. We want to identify in this situation a variable that is likely to be contained in an “acceptable” feasible solution for SPP.

#### 5.1 Stage 1: Scoring

The literature suggests the computation of a *score* value function

$$\sigma_j = \sigma_j(\omega_j, \pi, \Phi, \eta), \quad j = 1, \ldots, n \quad (38)$$
Fig. 6. chuzc duty fixing

for each variable as a heuristic measure for the attractiveness of fixing this variable. A variety of rules have been studied, see [5], [8], [36], [16], and [15]. Among the more popular scores is the “average reduced cost per uncovered row”

$$\sigma_j = \frac{\bar{w}_j}{\Phi_j^T \eta},$$  \hspace{1cm} (39)

which is a major component of the score that we use. Selecting a set $J$ of some 20 compatible candidates of smallest score concludes stage 1 of our fixing heuristic.

5.2 Stage 2: Lagrangean Probing

Stage 2 tries to improve the scoring information by tentatively fixing variables and exploring the consequences in an LRDLP reoptimization. This technique is called probing and is discussed, e.g., in [47] and, in a strong branching context, in [3]. We compute for each candidate variable $j \in J$ with our aug routine an up-penalty

$$\theta_j = \max_{\pi \in \mathbb{R}^m} \omega^T \pi + \pi^T (\Phi \xi - 1), \quad \ell + e_j \leq \xi \leq 1, \quad j \in J,$$

and fix the one with the smallest penalty.

6 Computation

We report in this section computational results obtained with our duty scheduling system dsopt. All tests have been performed within the MICROBUS II system on commonplace PCs.
|
|----------------|----------------|----------------|
|                | prob1 | prob2 | prob3  |
| # vehicle blocks | 63    | 56    | 59    |
| incl. spares     | yes   | no    | no    |
| # duty elements  | 727   | 1065  | 1968  |
| # tasks (nodes)  | 7037  | 24284 | 51488 |
| # links          | 93725 | 261971| 277681|
| duty type categories | straights | straights | straights |
|                  | splits | splits | splits |
|                  | shorts | shorts | shorts |
| opt working time | 7:42/7:42 | 8/8/5 | 5:00–5:45 |

manual DS in MICROBUS II

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>status</td>
<td>100%</td>
</tr>
<tr>
<td># duties</td>
<td>88</td>
</tr>
<tr>
<td>Ø working time</td>
<td>7:27</td>
</tr>
</tbody>
</table>

DS-OPT

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>status</td>
<td>100%</td>
</tr>
<tr>
<td># duties</td>
<td>85</td>
</tr>
<tr>
<td>Ø working time</td>
<td>7:06</td>
</tr>
<tr>
<td>running time</td>
<td>0:45*</td>
</tr>
<tr>
<td>required MB</td>
<td>150</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td># duties</td>
<td>63</td>
</tr>
<tr>
<td>Ø working time</td>
<td>7:50</td>
</tr>
<tr>
<td>running time</td>
<td>1:20a</td>
</tr>
<tr>
<td>required MB</td>
<td>230</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td># duties</td>
<td>111</td>
</tr>
<tr>
<td>Ø working time</td>
<td>5:41</td>
</tr>
<tr>
<td>running time</td>
<td>5:30b</td>
</tr>
<tr>
<td>required MB</td>
<td>230</td>
</tr>
</tbody>
</table>

\(^a\) Pentium III 650 MHz, 578 MB memory, Linux 2.2.14, gcc-2.95.2
\(^b\) Pentium III 500 MHz, 1 GB memory, Windows NT, VC++ 6.0

Table 1. Solving duty scheduling problems with dsopt

Table 1 lists characteristics and results for three European scenarios. prob1 and prob2 are urban bus crew scheduling problems from German operations, prob3 is a subway driver scheduling problem from a European metropolis.

The top of the table gives some information about the characteristics of these problems. The rows “# vehicle blocks” (“#” means “number of”) and “including spares” are self-explanatory and show that we deal here with relatively, albeit not extremely large problems on multiple lines, i.e., we allow changeovers and do not plan on a line-by-line basis. Cutting the vehicle blocks from line “# vehicle blocks” at the relief points, we obtain the “# duty elements” as listed in the next line. The large number in prob3 results from
this company’s willingness to relieve basically anytime at any subway station. The duty scheduling network is set up from this data in two steps. First, nodes for possible extension elements (extension elements take care, e.g., of sign-on and sign-off activities) are added. (We enumerate all possible extensions and add a node for every single one, see [14, in German] for details.) This results in a large “# nodes”, but it encapsulates the treatment of extensions in the network data. It also turns out that such an a priori enumeration has no negative effect on the running time. The network is completed by adding the given “# links” for locally feasible transitions among duty and extension elements. We attach importance to the inclusion of all possible links in this step, because each link is a potentially valuable degree of freedom. The next three lines give information on “duty type categories”. Each listed category represents several duty types. The last line reports, as one example of the duty type parameters, the desired average working time.

The second part of the table reports statistics on the best solutions that have been obtained by the companies. The third part lists the DS-OPT solutions. The numbers show that the optimized solutions contain sometimes significantly less duties, while the scheduling quality in terms of rule compliance is at least comparable (“status” is the percentage of duty elements that have been scheduled into duties).

Time and memory requirements for these runs are listed in the fourth part of the table. It is apparent that, also from this point of view, optimization systems such as DS-OPT definitely qualify as production tools for operational planning in public transit.

References


Rotation Planning for the Continental Service of a European Airline

M. Jünger¹, M. Elf¹, and V. Kaibel²

¹ Universität zu Köln, Institut für Informatik, Pohligstraße 1, 50169 Köln, Germany
² Technische Universität Berlin, Fachbereich Mathematik, Straße des 17. Juni 136, 10623 Berlin, Germany

Abstract. We consider a version of the aircraft rotation problem where the objective is to minimize delay risks. Given a set of flights to be flown by a subfleet the rotation problem is to find a specific route for each aircraft of the subfleet such that each flight is flown by exactly one aircraft. Additionally, the sequence of flights defining a route must satisfy certain requirements mainly to avoid delays. We present a mathematical model for the problem of minimizing the delay risk according to special requirements of a major airline. An efficient Lagrangian heuristic is proposed that uses subgradient optimization and linear assignments as subproblems. Computational results on real data are given and compared to actual aircraft rotations of that airline.

1 Introduction

The aircraft rotation problem we consider in this article is one of the first steps in the planning process of an airline. Usually, two other problems have to be solved before. First, the airline must decide which flights should be offered. This depends, for instance, on the expected numbers of passengers and profits of the flights. Because an airline has different types of aircraft (subfleets) each with different characteristics like seating capacities, crew size, and operational costs it must also decide which aircraft type should be used to establish a connection. The result of this fleet assignment is a flight schedule for each subfleet which is the input for the rotation planning problem.

A flight schedule for a given subfleet consists of different legs (i.e., non-stop flights) each defined by a departure and arrival airport, a date, and departure and arrival times. An example for a leg between Frankfurt and Amsterdam could be


The aircraft rotation problem is to determine routes, i.e., sequences of consecutive legs flown by the same aircraft.

In contrast to other problems arising within the planning process of an airline (like the fleet-assignment problem mentioned above or the crew-scheduling problem) the aircraft rotation problem has seldomly been a subject
in Combinatorial Optimization. Some models, solution methods, and results for the aircraft rotation problem of major U.S. airlines can be found, e.g., in the papers of Clarke et al. ([1997]) and Bohland et al. ([1998]). The main objective of their models is the maximization of the through value. Because airline networks in northern America have a distinctive hub and spoke structure, passengers often have to change airplanes at the hubs during their trip. Airlines try to make their connections attractive by choosing rotations which allow passengers to travel without changing airplanes too often. The trough value is the revenue the airline expects to gain by additional passengers attracted to a service because of the possibility to stay on the same airplane rather than to change it during a stop-over. Besides the through value maximization these models include maintenance constraints which guarantee regular visits of sufficient lengths at maintenance stations for every aircraft.

The rotation problem we consider is of different structure because the objective is not to optimize through values and maintenance conditions are negligible. This is due to a different handling of maintenance issues and due to a different weighting of the goals "maximizing through value" and "minimizing delay risk" by our industrial partner Lufthansa Systems GmbH.

The rest of this article is organized as follows. In the next parts of this section we define our problem in detail and describe special requirements and properties. Section 2 describes a mathematical model based on a graph formulation. This is used to formulate an algorithm for the solution of the rotation problem in Sect. 3. Computational results are presented in Sect. 4.

1.1 The Rotation Problem

The input for our rotation problem is a feasible schedule. This schedule is the result of a fleet assignment and consists of all legs to be flown by a subfleet within a period of about six months. Feasibility means that there exist routes covering all legs with at most as many routes as there are airplanes available. Each leg is flown by exactly one airplane. This rotation planning problem can also be interpreted as the problem of assigning each incoming leg at an airport to one outgoing leg. By defining the successor of each leg routes are determined.

In contrast to U.S. airlines continental services of European airlines show a different structure. There is no hub and spoke network but a small number of airports in the home country and many destinations in other parts of Europe. Figure 1 shows an example of all connections within one day of a large subfleet of a German airline. We can see that the arrival airports, resp. the departure airports, of most connections are located in the home country. That means, most of the connections affect a relatively small number of airports. Later we describe how this can be used to reduce the problem size.

Because usually there are only few stop-overs, the through value is of minor interest. The objective of our optimization is the reduction of delay risks. In order to avoid delays we are given several rules for connecting legs.
These rules had been developed by our industrial partner and represent its criteria for building rotations. Some of the rules forbid connections, others penalize them because of an increasing delay risk. There are two kinds of penalizing rules which we describe below. Each violation of a rule results in a penalty. We are interested in rotations with a minimum sum of penalties which means small delay risks.

One set of rules (local rules) are easy to handle. They affect the connection of exactly two legs. Later we will see how to model them with linear assignment problems. An example of a local rule is that between two connected legs there must be a certain positive minimum ground time between the arrival time of the first and the departure time of the second leg. This time is needed, e.g., for filling-up fuel, cleaning-up, and inspecting the aircraft. It depends on the airport of the connection, the departure airport of the first, and the arrival airport of the second leg. Sufficiently large ground times should furthermore guarantee that airplanes can start at their scheduled departure times even if they arrive late. Under certain conditions the actual ground time is allowed to be slightly less than the minimum ground time, although
this is not appreciated due to an increasing delay risk. The number of such minimum ground time (MGT) violations should be kept small.

Some airports in Europe have a very large volume of traffic. They are called air traffic critical (ATC) airports. Routes not visiting such airports too often should be preferred because large traffic may cause delay. This is taken into account by another local rule saying that the departure airport and the arrival airport of two connected legs must not be air traffic critical. A violation is called direct ATC violation. Overall there are six local rules.

The second class of rules affect the consecutive connection of more than two legs, i.e., subpaths of routes. They are called non-local rules and are much more difficult to handle. One rule restricts the number of ATC airports an airplane visits within one day. Another rule demands not to connect a certain number of consecutive legs with only the minimum ground time. The most general rule is the following. For each subpath of a route beginning in the morning a propagated minimum ground time (PMGT) is defined which depends on the structure of the subpath and the possibility of delay accumulation. The ground time of the last connection in such a subpath should not exceed the propagated minimum groundtime.

1.2 Special Properties

In our problem we do not need to worry about maintenance requirements because maintenance events are implicitly modeled in the schedule. The given schedule contains maintenance reserves (operational legs) which are virtual flights with the same arrival and departure airport like

\[(\text{FRA} / 04. \text{Aug} / 00:00 \rightarrow \text{FRA} / 04. \text{Aug} / 23:59)\].

This represents an 24 hours maintenance check at Frankfurt. Operational legs are treated like usual legs in the rotation planning process. The strategy of the airline we have cooperated with is to exchange an aircraft that requires a certain maintenance service with an aircraft scheduled for a maintenance reserve when necessary.

![Fig. 2. Average number of airborne machines over the period of one day. Each graph corresponds to a subfleet. Time is given in hours](image)
The European air traffic has a special structure. There are very few overnight flights. Figure 2 shows the average number of airborne machines for several subfleets of a German airline over the period of one day. We can see that there is only one subfleet with an average of one airborne airplane during the night. For all other subfleets there is a period of about three hours with no flights. It follows that delays cannot accumulate overnight. This means we can decompose the whole problem into optimization for each day.

2 Modelling the Problem

In this section we formulate the rotation planning problem as a graph problem and derive an integer linear program (ILP) for which we will develop an algorithm in Sect. 3.

2.1 A Graph Problem

Let \( n \in \mathbb{N} \) be the number of airports and \( \mathcal{L} \) be the set of legs for which arrival or departure time lie within the optimization period (usually one day). For each airport \( i \in \{1, \ldots, n\} \) we define a set of arrival nodes \( A_i \) and departure nodes \( D_i \). For each leg arriving at an airport \( i \) within the period there is exactly one arrival node \( a \in A_i \) and for each leg leaving an airport \( i \) within the period there is exactly one departure node \( d \in D_i \). Thus, usual legs (i.e., those completely lying within the period) yield two nodes in the model, while (the rare) “overnight-legs” give rise to only one node. Now we define edges representing feasible connections of two legs. Connectable legs correspond to a pair of arrival and departure nodes at the same airport. We define an edge set \( E_i \) at each airport \( i \). If it is possible to connect the legs corresponding to the arrival node \( a \in A_i \) and the departure node \( d \in D_i \) we introduce an edge \( e_{ad} \in E_i \). This yields a bipartite graph \( G_i = (A_i \cup D_i, E_i) \) for each airport \( i \in \{1, \ldots, n\} \). Let \( G = (A \cup D, E) = \bigcup_{i=1}^{n} G_i \) with \( A = \bigcup_{i=1}^{n} A_i \), \( D = \bigcup_{i=1}^{n} D_i \), and \( E = \bigcup_{i=1}^{n} E_i \) be the collection of these graphs.

Figure 3 shows an example with three airports. Nodes are drawn top down according to the arrival resp. departure times of their legs. Solid edges represent possible connections. Arrival and departure nodes belonging to the same leg are connected with a dotted arc. They are not edges of \( G \) but represent the flight defined by their end nodes. Some of the arrival nodes without a corresponding departure node are drawn red. These nodes are called initial nodes. They represent the aircraft being located at the respective airport at the start of the optimization period (e.g., those that have spent the night at that airport). The initial nodes mark the beginning of the different routes in the rotation schedule.

In order to find a rotation we have to assign an arriving leg to each departure leg. In the graph model this results in D-perfect matchings, i.e., subsets of the edges \( E \) such that no two edges in the subset share a common
endnode, but every node in $D$ is covered. Conversely, each $D$-perfect matching defines a rotation by assigning an arrival leg to each departure leg. Figure 4 shows a $D$-perfect matching for the graph model of Fig. 3. Bold edges are part of the $D$-perfect matching. The marked path starting from the first arrival node at airport $A_1$ represents one of five routes in the solution.

![Graph Model with Three Airports](image1)

**Fig. 3.** The graph model with three airports. Solid edges represent possible connections. Dotted edges correspond to flights. Initial nodes are drawn red.

![Solution of the Rotation Problem](image2)

**Fig. 4.** One possible solution of the rotation problem. Bold edges represent chosen connections. Red edges show a resulting route.
As mentioned above, local rules only affect two connected legs. If we weight the edges \((a, d) \in E\) of our graph \(G\) by \(c_{(a,d)} \in \mathbb{R}_{\geq 0}\) according to the penalty caused by the violation of local rules and neglect non-local rules, the rotation problem can be solved by computing a minimum weighted \(D\)-perfect matching of \(G\). This problem can be solved efficiently (see, e.g., Ahuja et al. ([1993])). Furthermore, the problem decomposes into \(D_i\)-perfect matching problems for each airport. In order to speed up the computation they should be processed separately.

2.2 An ILP-Formulation

Non-local rules affect more than one consecutive connection. They implicitly define certain possible paths in the route of an aircraft that should be avoided. These rules cannot be expressed in the matching formulation. Based on the graph model, we formulate the problem with non-local rules as an integer linear program.

Let \(P\) be the set of paths defined by the non-local rules. Every \(P \in \mathcal{P}\) consists of some edges \(P = e_1 e_2 \ldots e_k\) in \(E\). Let \(|P| := k\) be the length of the path \(P\). For every path \(P \in \mathcal{P}\) we define a weight \(\gamma_P \in \mathbb{R}_{\geq 0}\) according to its penalty. For each edge \((a, d) \in E\) we define a binary variable \(x_{(a,d)} \in \{0, 1\}\) with the meaning that \(x_{(a,b)} = 1\) holds if the edge \((a, b) \in E\) belongs to the solution and \(x_{(a,b)} = 0\) otherwise. For each path \(P \in \mathcal{P}\) we introduce a binary variable \(y_P \in \{0, 1\}\) with \(y_P = 1\) if and only if each edge \(e\) in the path \(P = e_1 e_2 \ldots e_k\) is part of the solution. The optimum solution of the integer linear program

\[
Z = \min \sum_{(a,d) \in E} c_{(a,d)} x_{(a,d)} + \sum_{P \in \mathcal{P}} \gamma_P y_P
\]

\[
\sum_{d \in D_i} x_{(a,d)} \leq 1 \quad (a \in A_i, 1 \leq i \leq n) \quad (1)
\]

\[
\sum_{a \in A_i} x_{(a,d)} = 1 \quad (d \in D_i, 1 \leq i \leq n) \quad (2)
\]

\[
\sum_{(a,d) \in P} x_{(a,d)} - y_P \leq |P| - 1 \quad (P \in \mathcal{P}) \quad (3)
\]

\[
x_{(a,d)} \in \{0, 1\} \quad ((a, d) \in E)
\]

\[
y_P \in \{0, 1\} \quad (P \in \mathcal{P})
\]

is the optimum solution of the rotation problem. The constraints (1) and (2) together with the binary condition for the \(x\)-variables guarantee that the \(x\)-part of the solution always defines a \(D\)-perfect matching in \(G\) because (1) implies that each arrival node is covered by at most one edge and (2) implies that each departure node is covered exactly once. The \(x\)-part in the objective
function sums up the penalties for local rules (matching edges). By setting \( y_P = 1 \) for all \( P \in \mathcal{P} \) the constraints (3) are always feasible and do not influence the \( x \)-part of a solution. Thus, for every rotation there is a feasible solution of the ILP and vice versa. Because the sense of the optimization is to minimize and the objective function coefficients \( \gamma_P \) are non-negative a variable \( y_P \) has value zero unless

\[
\sum_{(a,d) \in P} x_{(a,d)} = |P|,
\]

i.e., unless the \( x \)-part of the solution defines a route containing the path \( P \). It follows that the objective function value is the sum of all penalties.

3 Solving the Problem

Using the ILP formulation of the previous section we developed a Lagrangean heuristic which is able to find good solutions in reasonable time. Running time plays an important role since most of the schedule and rotation planning is still an interactive process. Schedules are manually fine-tuned by forbidding existing connections, creating new connections, and by changing legs. Afterwards new rotations are determined by the computer and again modified by a schedule designer. This process is repeated several times.

Another possibility for solving the ILP formulation is a branch & cut approach using the LP-bound of the ILP-formulation. The LP-bound can be computed with a cutting plane procedure. Starting with the linear program containing only the constraint sets (1) and (2), constraints of set (3) are added if they are violated. After that the linear program is reoptimized and the process is iterated until no constraint is violated. We tried this approach but it turned out not to be competitive with the approach we describe in the following section. This has basically two reasons. First, in general this method does not yield feasible solutions for the rotation problem and it is not easy to construct rotations from the fractional LP-solutions. The second reason is that in each iteration new violated constraints must be detected. This separation problem is a shortest path problem with difficult side constraints. Its solution is very time consuming.

3.1 Lagrangean Relaxation

Besides the constraints (1) and (2) for the \( x \)-part of the solution our ILP-formulation contains a constraint and an additional variable \( y_P \) for each subpath \( P \) associated with a non-local rule. The flight schedules we are dealing with contain thousands of such subpaths which result in integer linear programs with a very large number of constraints and variables. They are too large for state of the art mixed integer solvers. In order to solve the rotation problem, we apply a Lagrangean relaxation to our ILP-formulation and use
subgradient optimization to solve the dual problem. In general, this approach does not yield optimal solutions. But compared to actually flown rotations we can substantially decrease the number of violations. The lower bound provided by the Lagrangean relaxation is of minor quality.

By relaxing the constraint set (3) for non-local rules with Lagrangean multipliers \( \lambda_P \in \mathbb{R}_{\geq 0} \) for each subpath \( P \in \mathcal{P} \) we obtain a Lagrangean relaxation

\[
Z_D(\lambda) = \min \sum_{(a,d) \in E} c_{(a,d)} x_{(a,d)} + \sum_{P \in \mathcal{P}} \gamma_P y_P \\
+ \sum_{P \in \mathcal{P}} \lambda_P \left( \sum_{(a,d) \in P} x_{(a,d)} - y_P - |P| + 1 \right)
\]

\[
\sum_{d \in D_i} x_{(a,d)} \leq 1 \quad (a \in A_i, 1 \leq i \leq n)
\]

\[
\sum_{a \in A_i} x_{(a,d)} = 1 \quad (d \in D_i, 1 \leq i \leq n)
\]

\[
x_{(a,d)} \in \{0, 1\} \quad ((a,d) \in E)
\]

\[
y_P \in \{0, 1\} \quad (P \in \mathcal{P})
\]

The Lagrangean dual is

\[
Z_D = \max_{\lambda \in \mathbb{R}_{\geq 0}} Z_D(\lambda)
\]

which yields a lower bound on the optimum objective function value of the ILP-formulation. Our Lagrangean relaxation has the “integrality property” which means that there exists always an integer optimum solution even if the integrality requirements are ignored. Obviously the \( y \)-part of any optimum solution is integer and does not influence the \( x \)-part. By a classical theorem due to Birkhoff ([1946]) (see also von Neumann, [1953]) the \( x \)-part of an optimum solution is also integer since the feasible set is an integer polyhedron. The integrality property implies that the bound obtained by the Lagrangean dual equals the LP-bound of the ILP-formulation (see, e.g., Schrijver ([1993])). The independence of the \( x \)-part and the \( y \)-part also guarantees that the Lagrangean relaxation with a fixed \( \lambda \) can be solved by determining \( D \)-perfect matchings in the model graph like in the case without non-local rules. The optimum values of \( y_P \) can be obtained by setting

\[
y_P = \begin{cases} 
0 & \text{if } \gamma_P \geq \lambda_P \\
1 & \text{if } \gamma_P < \lambda_P 
\end{cases} \quad \text{for all } P \in \mathcal{P}
\]

### 3.2 Subgradient Optimization

Because the Lagrangean dual is a piecewise linear convex optimization problem it can be solved by a subgradient method as described in Fischer ([1985]).
Given an initial value \( \lambda^{(0)} \), a sequence \( \{\lambda^{(k)}\} \) of Lagrangean multipliers is iteratively generated by the rule

\[
\lambda^{(k+1)}_P = \max \left\{ \lambda^{(k)}_P + t_k \xi^{(k)}_P, 0 \right\} \quad (P \in \mathcal{P}, k \geq 1)
\]

(4)

where \( \xi^{(k)}_P \) is a subgradient of the dual function \( Z_D(\lambda) \) and \( t_k \) is a positive scalar step size. The subgradient of the \( k \)-th iteration is obtained as

\[
\xi^{(k)}_P = \sum_{(a,d) \in P} x^{(k)}_{(a,d)} - y^{(k)}_{P(a,d)} - |P| + 1 \quad (P \in \mathcal{P}, k \geq 1)
\]

with the optimum solution \( (x^{(k)}, y^{(k)}) \) of the Lagrangean relaxation \( Z_D(\lambda^{(k)}) \).

Theoretically, the sequence \( \{Z_D(\lambda^{(k)})\} \) converges to \( Z_D \) if the step length \( t_k \) is chosen such that \( t_k \to 0 \) and \( \sum_{i=0}^k t_i \to \infty \). In practice it is more suitable to choose the step length according to

\[
t_k = \alpha \frac{\hat{Z} - Z_D(\lambda^{(k)})}{\|\xi^{(k)}\|^2}
\]

with an upper bound \( \hat{Z} \geq Z_D \) and a periodically decreasing \( 0 < \alpha < 2 \).

In our application the initial values are \( \lambda_P = 0 \) for the Lagrangean multipliers and \( \alpha = 1.9 \). If the lower bound cannot be improved within 5 iterations of the subgradient method \( \alpha \), is divided by 2. In each iteration of the subgradient optimization, the Lagrangean subproblem is solved and yields a feasible solution of the rotation problem, since the \( x \)-part of the solution defines \( D \)-perfect matchings in the model graph \( G \). The best solution is stored and the upper bound \( \hat{Z} \) is set to the corresponding objective function value. If \( \hat{Z} \) is updated, \( \alpha \) is set to 1.9 again. Because of the large number of Lagrangean multipliers we generate them only if they become positive, i.e. influence the relaxation. This happens only if the corresponding subpath is contained in a previously computed \( D \)-perfect matching. Multipliers are removed as soon as their value becomes 0. By this strategy, the relaxation can be kept small.

In our experiments with real data the number of active multipliers never exceeded 200. We run the subgradient procedure for at most 250 iterations but stop if the length of the search direction becomes too small.

Usually subgradient optimization does not produce sequences of monotonously decreasing lower bounds. Often a phenomenon called “zig-zagging” occurs which has very bad influence on the convergence and the quality of the solution. This can be avoided by using an improved version of the subgradient method proposed by Crowder ([1976]). In this variant the search strategy (4) is replaced by

\[
\lambda^{(k+1)}_P = \max \left\{ \lambda^{(k)}_P + t_k d^{(k)}_P, 0 \right\} \quad (P \in \mathcal{P}, k \geq 1)
\]

and the search direction

\[
d^{(k)} = \xi^{(k)} + \theta d^{(k-1)} \quad (k \geq 1)
\]
is chosen as a linear combination of the current subgradient and the previous direction. In our application a damping factor of $\theta = 0.7$ turned out to be suitable.

### 3.3 Solving the Subproblems

During the subgradient procedure most of the work is spent in solving the Lagrangean subproblems, i.e., for the solution of the bipartite matching problems in the model graph $G$. In our software the time for solving the matching problems is about 98% of the overall running time. Thus, these problems should be solved as efficiently as possible. One possibility to gain a faster algorithm is to reduce the problem size. In Section 1.1 we already mentioned the special structure of our flight schedules. There are many airports with a very small number (e.g., $\leq 4$) of connections during the day. This often results in bipartite matchings with exactly one solution or in problems where most of the edges cannot be part of any solution. Due to structural reasons, edges that are not part of any solution even occur at airports with larger number of connections. We can prune all these edges from the model graph in a preprocessing step. This is done as follows.

Consider a bipartite graph $G_i = (A_i \cup D_i, E_i)$ of our model and let $M$ be a $D_i$-perfect matching. An $M$-alternating path is a path such that each non-matching edge is followed by a matching edge and vice versa. Alternating cycles are defined similarly. Since the symmetric difference of two matchings is a disjoint union of alternating paths and cycles, it follows that an edge belongs to no $D_i$-perfect matching if it is not contained in any $M$-alternating cycle and not contained in an alternating path with one end covered by $M$ and one not. If we direct all edges in $M$ from $D_i$ to $A_i$ and all other edges from $A_i$ to $D_i$, the first kind of edges are edges between two strongly connected components of the directed graph. The second kind of edges are edges which are not part of a directed path starting at a free node in $A_i$. Both can be computed in linear time.

Another crucial point for keeping the running time as small as possible is the method for solving the bipartite matching problems. Usually this is done by the Hungarian method (see Kuhn ([1955])), but for our problem primal methods are more suitable. During the subgradient optimization we have to solve a sequence of matching problems at the same graph but with different objective functions. Especially in later iterations of the subgradient method when the step size has become small there are only little changes of edge weights. It follows that the optimum objective function values of the matching problems change only slightly and that the solution of an iteration is near to the optimum of the next iteration.

This is taken into account by primal methods. They are able to perform a "warm start" at a feasible solution, e.g., the near optimum solution of the previous iteration of the subgradient method. Often it is possible to compute the optimum matching with only few iterations. In our application we
formulated the bipartite matching problems as uncapacitated transportation problems and solved them with a specialized variant of the network simplex method (see Ahuja et al. ([1993])).

4 Results

We tested our algorithm on three different data sets of real flight schedules. For each of these schedules we performed experiments with six subfleets of different sizes and characteristics. It turned out that the schedules do not differ in terms of the complexity of the rotation problems. Table 1 shows the problem characteristics of all subfleets for a typical day. There are four large subfleets (321, 320, 733, and 735) and three smaller ones (314, 319, and AB6). For each subfleet the table contains a row with the number of edges in the model graph $G$, the number of chosen edges (connections) in the solution, the number of dualized constraints for non-local rules, the maximum number of non-zero Lagrangean multipliers during the subgradient method, and the time for the Lagrangean heuristic.

Because our data shows a distinctive weekly periodicity we do not present results of the whole period of about six month but concentrate on a small number of weeks. The presented results are generalizable. Figure 5 shows the ratios of the number of violated rules in solutions found with the Lagrangean heuristic and the number of violations in actually flown routes. Each graph corresponds to one of the subfleets under consideration. Except for the subfleets 314, 319, and AB6 the ratio is between 0.4 and 0.7 which means that the computed routes contain 30%-60% less violations than the flown routes. The number of violations for the subfleets 314 and AB6 is at least 20% better than in the flown routes. The ratio for subfleet AB6 is often less than 0.4. For the subfleet 319 it is sometimes possible to obtain reductions of violations better than 50% but on some days we only get the same number of violations. The time used to compute the solutions is always less than 5 seconds. Table 1 contains typical running times for all subfleets at a SUN Ultra 4 workstation with 300 MHz.

Until now we only considered the total amount of violations. Table 2 shows a detailed comparison of all violations in computed and flown routes.

<table>
<thead>
<tr>
<th>subfleets</th>
<th>314</th>
<th>319</th>
<th>320</th>
<th>321</th>
<th>733</th>
<th>735</th>
<th>AB6</th>
</tr>
</thead>
<tbody>
<tr>
<td>edges in $G$</td>
<td>150</td>
<td>518</td>
<td>3064</td>
<td>1729</td>
<td>2334</td>
<td>1507</td>
<td>373</td>
</tr>
<tr>
<td>connections</td>
<td>33</td>
<td>108</td>
<td>207</td>
<td>236</td>
<td>207</td>
<td>129</td>
<td>48</td>
</tr>
<tr>
<td>dualized constraints</td>
<td>85</td>
<td>142</td>
<td>1035</td>
<td>1132</td>
<td>1265</td>
<td>800</td>
<td>27</td>
</tr>
<tr>
<td>non-zero multipliers</td>
<td>34</td>
<td>57</td>
<td>104</td>
<td>175</td>
<td>130</td>
<td>113</td>
<td>19</td>
</tr>
<tr>
<td>time in seconds</td>
<td>0.2</td>
<td>0.7</td>
<td>4.9</td>
<td>3.0</td>
<td>4.1</td>
<td>2.4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 1. Typical problem characteristics for all subfleets
Fig. 5. Ratios of the total number of violated rules for computed and actually flown rotations. Each of the six graphs corresponds to a subfleet for one week of subfleet 733. Also these results are similar for different weeks and subfleets. We can see that in both cases the number of violations is dominated by the number of violations for the rule “I” which is the PMGT rule. With the Lagrangean heuristic it is possible to reduce these violations drastically. Although we minimize the total amount of violations, the overall decrease of violations is obtained by reducing violations for all but one rule. Only the number of violations of rule “E” (MGT violation) in the computed routes marginally exceeds the violations in the flown routes three times.

The last column of Table 2 contains the lower bound obtained by optimizing the Lagrangean dual. There is a large gap between the best solution and this bound. This is not caused by an insufficient convergence of the subgradient method but the poor LP-bound of our ILP-formulation. Often a Lagrangean heuristic is embedded in an branch and bound scheme to find

<table>
<thead>
<tr>
<th>Day</th>
<th>local rules</th>
<th>non-local rules</th>
<th>bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>1.</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>2.</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>3.</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>4.</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>5.</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>6.</td>
<td>11/11</td>
<td>2/3</td>
<td>0/0</td>
</tr>
<tr>
<td>7.</td>
<td>7/7</td>
<td>1/2</td>
<td>0/0</td>
</tr>
</tbody>
</table>

Table 2. Detailed comparison of computed and actual routes for the subfleet 733. The left value in a column is the number of violations for the computed route the right value is the number of violations in the actually flown routes.
optimum solutions. Because such a scheme strongly depends on the quality of the bounding procedure this is not reasonable in our case.

In order to make our algorithms and tools for the rotation problem suitable for practical use we bundled them in the software package *RPF*. This package contains a C++ class library, a command line tool, and a graphical user interface. The library can be used to integrate our methods in existing software packages. With the command line tool it is possible to read/write the schedule and routing data in the standard *SSIM* format used by many airlines. With this format we can easily communicate with existing software without changing it. In addition to the computation of rotations, our tool provides a lot of functionality to analyze schedule data and routing information. With the graphical user interface the schedule and routing information can be displayed in a user-friendly way. Figure 6 shows a screenshot of our interface displaying a solution for subfleet 314. It is also possible to show differences of two rotations, e.g., rotations found by optimizing with different weighting of violations. This can help to analyse different rotations and to find an appropriate weighting strategy.
References


Computer Aided Scheduling
of Switching Engines

Uwe T. Zimmermann and Marco E. Lübbecke*

Department of Mathematical Optimization, Braunschweig
University of Technology, Pockelsstraße 14, 38106 Braunschweig, Germany

Abstract. Scheduling the switching engines of an industrial railroad is a formidable
and responsible task, closely related to the well-studied pickup and delivery problem
with time windows. Aiming at an efficient usage of resources the need arose for a
computer aided scheduling tool as support for the human dispatcher. We sketch a
set partitioning formulation of this problem to be solved via column generation. The
pricing subproblem is hard in the theoretical sense but can be solved by means of a
combination of heuristics and exact algorithms. A trade-off between mathematical
rigor and practicability becomes apparent and is extensively discussed. Our com-
putational experience with an academic prototype implementation is encouraging.
We succeed in obtaining practically acceptable solutions for instances of more than
forty customers and six vehicles.

1 Motivation

Market deregulations in the German railroad sector some years ago forced
private railroads to offer a better transportation quality and to decrease
charges. The efficient use – and desirably a reduction – of available resources
became indispensable. The research project Umlaufplanung im Güterverkehr
bei Werks- und Industriebahnen\footnote{Funded by the German Federal Ministry of Education and Research (BMBF),
grant no. 03-ZI7BR2-1} reported upon here deals with improving
the productivity of the central resource, viz. the switching engines.

In applied mathematics, in particular in combinatorial optimization, we
currently witness a change of focus towards making the large body of available
methodology \textit{utilizable} in practice. Optimal decision support for complex and
large-scale real-world applications comes in reach due to innovative numerical
and algorithmic techniques accompanied by an enormous progress made in
modern computer technology: Large amounts of data now available for the
description of a practical problem at a realistic level of detail can reveal the
limitations of an algorithm probably not detected in a laboratory environ-
ment. Moreover, advanced implementations are the reason for the success
of techniques known for a long time but used only recently. The ease of
testing may be the key feature offered by fast computers as well as the ready
availability of the most elaborated algorithms as commercial software.

* Corresponding author. E-mail: M.Luebbecke@tu-bs.de
Our main purpose here is to sketch the necessary steps for scientists and practitioners jointly attacking an industrial problem, *viz.* problem identification, model building, algorithm design, implementation, interpretation of the solution, and installation in practice.

2 Engine Scheduling

Operating a private railroad system as a subsidiary is indicated for large companies in the chemical, automobile, and steel industry in order to maintain a timely around-the-clock production process. Some run more than 100 *switching engines*, cf. Fig. 1, which differ in their technical and personal equipment, to handle the flow of up to 6000 freight cars between various terminals called customers. No train schedules exist, except for trains connecting to German Rail. That is, each transportation is performed upon *request* only, defined by origin, destination, service times and time windows, tonnage, and a list of eligible engines. Most attributes are optional. This definition is an abstract mathematical device, distinct from the one used in practice. But it is able to also incorporate all occurring kinds of special events like engineer's breaks, fueling, maintenance, and repair, to list only a few. It is the dispatcher's task to allocate engines to requests such that all requested transportation is served to the customer's satisfaction, i.e., compliant to the above constraints. The tractive power of an engine must never be exceeded. It goes without saying

Fig. 1. A switching engine transports molten iron at a steel works
that our presentation is highly simplified. In railroad traffic, requests are to be scheduled according to certain simple configurations, cf. Fig. 2, in order to avoid too much shunting; the topography of the tracks, peculiarities of single customers, weather conditions, service agreements, priorities, and many more complete the picture. For a much more detailed discussion we refer to [9,10].

Current information and disposition systems merely map reality onto the screen, offering an efficient administration e.g., by means of a graphical display of the track layout. The actual disposition, however, is done manually; the decision about how to schedule engines is completely up to the dispatcher and his or her experience and motivation. One would expect that in a more or less regular industrial operation an experienced dispatcher will anticipate most of the customer’s requests and schedule the engines accordingly. While in normal operation this is actually true, during peak workloads looking ahead is practically impossible. Requests are served on a first-come first-serve basis, or, more realistically, in a manner depictedly termed **loudest-shout first-serve**. Clearly, a human judgment of dependencies between different decisions of such complexity is only local and incomplete. The present project aims at providing a computer tool which suggests a schedule, thus supporting and relieving the dispatcher. What is more, the tool should be able to pursue operational goals, e.g., the minimization of unproductive work. The following
result [10] attests that in principle there is no better way of solving this problem other than evaluating each possible schedule and select the best.

Lemma 1. The above Engine Scheduling Problem is NP-hard.

3 Set Partitioning Formulation

An obvious mathematical formulation does not necessarily yield a good model. This simple statement conceals that considerable mathematical experience and intuition are required in order to identify the non-obvious. A natural mixed-integer formulation of the Engine Scheduling Problem (ESP), following the lines of [5,12], with explicit decisions on engine assignments, request sequences, service times, and tonnage requirements, respectively, turned out to be highly symmetric [7]. That is, in a sense, each decision carries too few information; similar solutions are reproducible in many different ways. Such circumstance embodies the danger that a reasonably directed search for an optimum be fruitless. Bearing this knowledge in mind, we follow an alternative approach.

A complete schedule for a given planning horizon splits into proposals, one for each engine. Assume for the moment that we are able to derive from a particular subset of requests and a specific engine the best possible way to serve the whole subset. Further assume that we only consider subsets which are feasible for the respective engine. Then the ESP is to select subsets of requests, at most one for each engine, such that the disjoint union of selected subsets contains each request precisely once, and the operational goal is minimized by the selection. To the adept this reads as follows.

\[
\begin{align*}
\min & \sum_{e \in E} \sum_{R \in \Omega_e} c_R^e x_R^e \\
\text{subject to} & \sum_{e \in E} \sum_{R \in \Omega_e} \delta_R x_R^e = 1, \quad r \in R \\
& \sum_{R \in \Omega_e} x_R^e \leq 1, \quad e \in E \\
& x_R^e \in \{0, 1\}, \quad e \in E, R \in \Omega_e
\end{align*}
\]

This seemingly compact, so called set partitioning formulation has been successfully applied to the generic problem, known as Multiple-Vehicle Pickup and Delivery Problem with Time Windows [5,12], but has some serious detriments. A minor difficulty is that determining the best possible cost of a
particular subset itself is an \( \mathcal{NP} \)-hard combinatorial optimization problem since it involves the optimal solution of the ESP for a single engine. Secondly, although the model has only a few constraints, the number of variables – which is the number of all feasible subsets of requests – is so large that storing the model in core memory, let alone solving it, is definitively out of reach. What more directly meets the eye is the fact that all information about time windows, engine loads, and tour structures is implicitly hidden behind some obscure construction of feasible subsets of requests. It seems that we only complicated matters, but fortunately some other advantages are ahead.

4 Decomposition and Column Generation

Relaxing the integrality constraint to \( 0 \leq x_{eR} \leq 1 \) for all \( e \in \mathcal{E}, R \in \Omega_e \), the set partitioning formulation becomes a linear program (LP) of prohibitive size. However, from the theory of the simplex method we know that in each iteration we only need access to as many variables as there are constraints present in the formulation, i.e., very few compared to the size of our model. It is thus an appealing idea to start with a very small portion of the model, and dynamically extend it in a smart way until it can be proved that all relevant information was considered, i.e., optimality is reached. The method is a two-level strategy; hence it is known as decomposition principle. A superior level determines what model extensions are locally most valuable, while the extension itself is calculated at an inferior level.

The theoretical principle underlying this idea can be best explained adopting an economic viewpoint, see e.g., [8]. Different branches of a large company independently take decisions on how much to use shared scarce resources in order to maximize the company’s overall profit. The branches inform a central coordinator about their locally optimal activities. The coordinator with a global view on the resource usage, but without direct influence on the subordinate decisions, fixes prices for each resource and returns them to the branches. These revise their decisions according to the resource prices, again without knowledge about the other decisions. The iterative process stops as soon as the global prices are such adjusted that no branch can offer a revision which contributes to an overall improvement of resource usage.
A translation to the linear programming framework goes as follows. An arbitrary feasible solution to (SP) induces initial sets \( \Omega'_e \subseteq \Omega_e \), \( e \in \mathcal{E} \). Already determining such a feasible solution is \( \mathcal{NP} \)-complete. At the beginning, only the variables corresponding to the elements in \( \Omega'_e \), \( e \in \mathcal{E} \), are incorporated in the linear program, the so-called restricted master program (RMP) – or the coordinator in the above wording. This latter is optimally solved based on the given tiny detail of the total model information, yielding an optimal dual solution \((u^*, v^*)\), partitioned according to the constraints of (LP). These prices – fixed by the coordinator – enable us to decide whether to enlarge the RMP by a certain, still unconsidered variable \( x_{R}^e \) or not by looking at the respective reduced cost coefficient

\[
c^R_e - \sum_{r \in R} u^*_r - v^*_e.
\]  

In the case that this entity is negative, adding the feasible subset \( R \) to \( \Omega'_e \) is promising from a global point of view. This gives new dual prices and we iterate. Otherwise we know that the optimal primal solution to the RMP solves the huge (LP) as well. The crucial task in this scheme is that of the company’s branches which have to propose favorable feasible subsets without explicitly considering them all. The strength of the described decomposition approach is that this is possible by solving an optimization subproblem, the so-called pricing problem. Since variables and their associated matrix coefficients are successively adjoined to the RMP one also refers to column generation.

The subproblem is to construct a feasible subset of requests with (minimal) negative associated reduced cost (1), or to prove that none such exist. Omitting the details, one derives from the above that this amounts to finding a (shortest) feasible tour for an engine, respecting all the constraints mentioned in Sect. 2, with the additional cost or gain, respectively, of the dual price \( u_r \) incurred for each visited request \( r \in \mathcal{R} \). By now we see where all the information vanished at the end of the last section. Sadly, the elegance of this algorithm is slightly afflicted by the following result [10].

**Theorem 2.** The pricing problem is \( \mathcal{NP} \)-complete.

**Solution of the Pricing Problem**

An explicit evaluation of (1) for all feasible tours and all engines would be justified by theory, cf. Theorem 2, but is out of the question from a practical point of view. Instead, an intelligent implicit enumeration is performed, in the hope – but without guarantee – that many solutions need not be considered.

The general strategy is, starting from empty tours, to successively append yet unvisited locations to partial tours, always preserving feasibility with respect to tonnage and time window constraints [3–5]. In our algorithm, we combine this technique with the fact that feasible tours are severely restricted in their combinatorial structure, cf. Fig. 2. Indeed, we proceed by iteratively
concatenating entire request configurations as depicted in Fig. 4. This gives a quickly growing tree like structure. It happens in the course of this dynamic programming algorithm that two tours are constructed that visit the same set of requests and end in the same location. Then the more expensive one is disregarded, pruning the respective branch of the tree. Such dominance criteria help to reduce the number of solutions to be considered and are already stated in [4]. However, for the use of such a criterion it is necessary to have the tours to be compared already constructed, and it requires a good overview over the search tree. Rather, it would be desirable to anticipate unpromising tours before their construction.

To this end, we exploit the following observation. There is no use further considering a particular tour when we have evidence that the best possible extension of this tour cannot have cost smaller than the cost of the cheapest tour constructed so far. In view of (1), the most advantageous extension of a tour $T$ is to visit exactly those yet unvisited requests $r \in R$ with $u_r$ positive. Whenever the so calculated lower bound $LB$ is larger than the cost $C$ of the currently best known tour, i.e.,

$$
LB = \text{cost}(T) - \sum_{r \in R \setminus R: u_r > 0} u_r \geq \min\{0, C\}
$$

we disregard $T$ and need not generate any of its extensions in the tree. Since the pricing problem must return a column with negative reduced cost coefficient (1) we also check whether $LB$ is non-negative. The use of this technique led to a considerable speedup of the pricing algorithm. In [9,10] we discuss possible refinements of $LB$. For an exhaustive presentation of the application of such bounds in implicit enumeration strategies we refer to [6].
Integer Solutions

It must be stated clearly that all our efforts so far only lead to an optimal solution of the linear relaxation of (SP) which may be fractional. Even worse, there is no guarantee that the final RMP is integer soluble at all; columns may be missing. That is, branch-and-bound, an implicit enumeration scheme to solve integer programs, must be enabled to generate additional columns if required. The resulting algorithm is known as branch-and-price [1].

The general proceeding is to iteratively fix the binary variables to zero or one in a binary tree manner. Again, evaluating each solution is impracticable, hence one strives to keep the search tree as small as possible. This is accomplished by virtue of lower bounds, similar in principle to our use in the dynamic programming algorithm for the pricing problem. Having fixed a set of variables, a lower bound on the integral objective function value is derived by solving the linear programming relaxation of the resulting (mixed) integer program. It goes without saying that the quality of the bound is of vital importance for the effectiveness of the procedure. Additionally, different model formulations may differ in the quality of their linear relaxations. This supplies a criterion for evaluating a formulation, and (SP) performs very well in this respect.

Our computational experiments indicate, under conditions to be explained in the next section, that the generated ("root node") restricted master programs happen to be already integer feasible. Here again, we come across the phenomenon that reality is not as bad as suggested by computational complexity, or in other words, in practice the theoretically worst case rarely occurs. This was observed in many other practical settings as well where artificial data were compared to "real world" data. Although being no mathematically satisfactory explanation this is an encouragement to attack practical problems even in the presence of negative complexity results.

5 Implementation Issues

With a mature mathematical background the column generation principle is relatively easy to comprehend. However, when it comes to an implementation the generic textbook presentation, cf. Algorithm 1, is of little help. In our experience, every ingredient offers much degree of freedom, the impact of which on the overall performance cannot be overestimated, c.f. Fig. 5. Questions arise as how to provide an initial solution, heuristically or artificially ("first phase simplex")? Solve the pricing problem exactly or approximately? What pricing rule to use? How many columns are adjoined at each iteration and how to make use of columns generated in earlier iterations ("column pool")? What about the growing size of the RMP? How often to re-optimize and how to solve the RMP, and when to terminate the algorithm? This list is by far not complete and omits all problem specific issues.
Algorithm 1 Generic column generation scheme to solve a linear program

Start with initial basis
while columns with negative reduced cost exist do
    Adjoin column to restricted master program
    Re-optimize
end while

Fig. 5. Impact of implementation details on the computation time for a particular instance. The generic algorithm even with usage of a column pool and a simple pricing heuristic terminated with a memory allocation failure. When the exact pricing algorithm is completed only when the pricing heuristics fail, at least the instance gets soluble. Using the lower bound (2) and a heuristic network reduction finally drastically speeds up the solution process.

Our bottleneck procedure is the pricing subproblem, which, as a general rule, should be solved exactly only when heuristics fail. We make extensive use of a column pool and add multiple columns per iteration. To our experience, this also increases the chance of obtaining integer solutions. There is no doubt that without the concerted interaction of heuristics and exact methods a solution even of smaller instances would not have been possible. The monographs [8,11] and the recent synopsis [2] on various acceleration techniques are valuable sources of information.

Computational Experience

As indicated earlier, our concept of request is more general than the one used in practice. Hence, all practical data had to be manually post-processed, and are only scarcely available. Thus we will only report on trends, but these are encouraging. The size of the instances we are able to solve within acceptable
time, i.e., in a few seconds in order to guarantee interactivity, corresponds to a planning horizon of more than two hours. According to officers of in-plant railroads involved in our project it is seldom sensible to plan further ahead because of the increasing uncertainty of future requests. Still, if there is need for solving larger instances a decomposition of the planning horizon into overlapping, manageable time slices would be reasonable.

The objective function value cumulates the minutes during which engines wait or run without load. In our case the optimum usually ranges around half an hour. In view of practical relevance our optimality gap of less than ten minutes is definitively tolerable. On a today’s personal computer it currently takes a few minutes to obtain an integer solution for more than forty requests to be served by six engines. Depending on parameter settings, rarely more than 20,000 columns are generated, in less than 200 iterations. The ratio customers/vehicles of seven is notably large for our problem class.

6 Towards Computer Aided Scheduling

When we launched this project, we identified the necessity of three areas of contribution, for each of which different partners with their respective competence were involved. Railroad companies furnish the practical background and realistic data, mathematicians supply models, algorithms, and validation of the produced solutions as well as an academic prototype software. People from the IT business are responsible for the marketing process, including the production of the final tool. Here is some experience we have drawn from many discussions with these partners and we found useful for similar projects.

The support and feedback of a practitioner at any stage helps to identify and to incorporate additional or modified operational constraints not thought of in the initial problem formulation. As a starting point, these constraints are not forgotten but left out intentionally in order to quickly provide a first prototype implementation. The purpose of such a model that represents only a simple or simplified scenario is to supply the practitioner with an estimation of the capabilities and potency of a more elaborated model. People are usually not used to see a computer performing a task previously done by a human expert. Hence they are sceptical. We remark, however, that the planning process can be covered only partly by a model anyway. A reasonable goal is to depute routine tasks to the computer while the dispatcher concentrates on non standard work. This relieves the dispatcher especially during peak workloads.

It is by no means sufficient (but necessary!) to convince the management of the usefulness of an optimization based decision tool. The managerial decision maker is in general not identical to the person who will work with the tool in every day operation. It is therefore even more important to study and carefully evaluate the experience and demands of the end user.

A computer still is sometimes seen as an opponent. Even when an existing computer system is to be enhanced by optimization techniques in form of a
planning or decision suggestion the fear of being replaced by the software is often seizable. Objection against the new technology may be a consequence, especially when the underlying methodology is not common knowledge even for experts in the planning environment. The use of mathematical programming is in this sense a black box method and solutions are not immediately recognizable by the personnel. Note, that we do not refer to the mathematical solution, but to the solution in terms of the practical setting formulated in the language of the practitioner. The structure of such a solution may differ considerably from a manually generated one the planner is used to see. When it does not feel right, it is not right. Thus, it is helpful to accompany the introduction of the system not only by explaining how it works but also by showing that it works and what practical ideas motivate the differences in the appearance of the solution. A graphical simulation clarifying the system’s suggestions may be of great help.

One point to particularly focus on is the quality of a solution. In mathematical terms, this notion is easily explained via bounds on the objective function value. In practice however, questions of stability and reliability arise. The former corresponds to the desired property that small changes in the input data shall result in small changes in the output, especially when the input changes frequently. Reliability means that the system is failsafe and provides a solution in any circumstance. A good strategy is to first compute a heuristic solution. Time permitting, an improvement of this solution based on new data is performed without giving notice and delivered only when needed.

Last, but not least, the notion of a computer suggestion should be taken literally. The proposal by the system serves as a default only, always subject to the final decision of the human dispatcher. Partial solutions fixed manually must never be overruled by the computer.

7 Discussion and Outlook

We have seen that the trade-off between mathematical rigor and practical needs is often decided in favor of the latter. One may have the impression that the benefits gained from the use of mathematical methods in industry are then compensated. Indeed, shortly after the introduction of a computer aided scheduling system, productivity may decrease significantly and the overall effect may even be negative. Of course, this is only one part of the truth. Once a planner judges the tool useful and reasonable, the default suggestion of the system will be accepted in most cases, not least for the reason of convenience. Building on this basis further components can be added, making the introduction of new technologies a dynamic process. We believe that in the long run remarkable savings offered by mathematical optimization approaches are definitely worth enduring possibly occurring intermediate embarrassments.

Industrial in-plant railroads usually operate a surplus of engines in order to handle peaks in the demand for transportation. However, it is believed
that a more regular and steady-going operation, also planning a little ahead, could help cutting back this overhead considerably. We have good reason to hope that in the end precisely this will happen due to the incorporation of the presented or revised algorithms in a dispatching software.

A shortcoming of our model is the assumption of independent engines, which needs not hold in general. In fact, individual transportation requests may be precedence related, a matter one can take care of if the respective requests are to be served by the same engine. If they are possibly incorporated in different tours, the presented set partitioning formulation is not suited to control feasibility of precedence constraints. Since this is a problem relevant to many practical vehicle routing settings lively research efforts are to be expected in this area.

Of course, we have seen a detail of the operational planning tasks of an industrial railroad only. When customers request for transportation capacity or a specified material, the flow of freight cars has to be managed. Only a small portion of the cars is owned by the railroad company; the majority is rent from other railroads. Cars are not identical, and for a customer request there is a type that fits best. It is roughly known which cars will be needed in particular working shifts, but manually it is hardly possible to anticipate the exact stock. Lacking cars have to be ordered timely. Since paying the cars is on an hourly basis the dispatcher tries to return them as soon as possible. Emptied cars from one terminal may be assigned to requests of another terminal when

**Fig. 6.** Similar output by our implementation leads to a readable schedule
appropriate. The goal is to minimize both the paid rent and the light running of cars while guaranteeing service without type mismatches. This problem of a distinct combinatorial nature also offers significant financial savings, and will be dealt with in an industrial project we are about to kick off.

Acknowledgments We are grateful to the in-plant railroads Eisenbahn und Häfen GmbH, EKO Trans GmbH, and Verkehrsbetriebe Peine-Salzgitter GmbH for many discussions and critical remarks, and for providing us with real world data.

References

Train Schedule Optimization in Public Rail Transport

U. T. Zimmermann and T. Lindner*

Department of Mathematical Optimization, Braunschweig University of Technology, Pockelsstraße 14, 38106 Braunschweig, Germany

Abstract. Attractive train schedules are quite important for the success of public rail transport. We consider schedules which are repeated after some fixed time period. Periodicity is a well accepted, convenient attribute of all major railroad systems. Since trains share the public railroad network, we have to handle many conflicting demands. Increasing competition from transportation alternatives as well as the politically motivated process of deregulation forces railroad companies to reduce cost and resource utilization. In this situation, better decision tools are quite welcome. Here, we discuss models and corresponding methods for cost optimal train scheduling in real world railroad networks.

1 Introduction

The train schedule summarizes the arrival and departure times of trains at certain points of the railroad network. These data form the backbone of public rail transport planning. Different purposes require information on different levels of aggregation. In particular, data points may include stations (high degree of aggregation) or switches and important signal points (low degree of aggregation). For the German railroad network depending on the degree of aggregation, we have to handle 8000 to 27000 data points.

Former schedules for long distance trains were highly irregular. Since usually only one train per day was scheduled for a specific connection, a periodical schedule made no sense. However, in highly congested urban areas, periodical schedules were used almost from the very beginning, e.g. subway trains in London 1863, Paris 1900, Berlin 1902.

Customers prefer periodical schedules that are much easier to memorize. When introducing periodical or, as they were called, fixed interval schedules for long distance traffic in 1939, the Dutch railroad company marked a new epoch. Other European countries followed much later: Denmark in 1974, Switzerland in 1984, Belgium and Austria in 1991. In Germany, an hourly schedule for InterCity trains was started in 1979. Beginning in 1985/86, schedules for InterRegio trains were based on a two hour period. Finally, from 1992/93, regional trains were scheduled with some fixed time intervals. For a more detailed historical introduction, we refer to [11].

* Corresponding author. E-mail: T.Lindner@tu-bs.de

W. Jäger et al. (eds.), Mathematics - Key Technology for the Future © Springer-Verlag Berlin Heidelberg 2003
Traditionally, schedules are visualized by time space diagrams, cf. Fig. 1. For a particular route of the network, a time space diagram contains lines representing trains serving that route. Critical points or potential resource conflicts are quite obvious: the train’s speed corresponds to the gradient of its line, a crossing of two lines shows that one train meets or passes another train.

Nowadays, software tools offer effective support for the construction of schedules. Information on track topology, engine and coach properties as well as available crews are stored in databases. Graphical user interfaces enable schedule planners to build and edit schedules interactively based on time space diagrams, e.g. Fig. 1. Conflicts are automatically indicated on the screen. However, the computer-automated generation and optimization of practicable schedules still remains quite time-consuming. In particular, currently implemented algorithms are simply too slow for networks of real world size.

Different departments of the railroad company formulate various goals. The sales and marketing departments give priority to travelers’ requests, e.g. short travel time, direct connections and, if necessary, switching of trains on the same platform with short waiting time. The controlling, management and logistics departments pay much more attention to cost related aspects and ask for efficient management of rolling stock and personnel resources. Rolling stock is a scarce resource, rules implied by contracts and legal requirements set tight limitations for planning. High traffic load at critical points in the network and security requirements add a lot of operational constraints. Goals

---

**Fig. 1.** Time space diagram for a schedule on the route Groningen–Rotterdam
and requirements, cf. Fig. 2, are obviously in conflict. Moreover, external factors like political decisions influence the planning process.

Of course, planning of train schedules is only a small part of the general planning process of traffic systems (cf. [3]) which includes in particular line planning as well as management of rolling stock and crews. Interaction between all parts of the hierarchically decomposed planning process has to be handled. For example, an optimal train schedule is based on the line plan and will be used as input to the subsequent crew planning. Therefore, a proposed train schedule may need adaptation or redesign in order to meet certain requirements from other subproblems.

2 Model

A railroad network is usually modeled by a graph $G = (V, E)$, where $V$ denotes the set of nodes and $E$ denotes the set of edges. Nodes represent stations or important network points like switches, and edges represent railroad tracks connecting these points. Train scheduling is based on a known line plan that defines the lines, i.e. the paths in the network that have to be served by trains with some fixed period $T$. This set of lines is denoted by $R$. For a fixed line plan, we have to construct a train schedule which assigns feasible departure and arrival times to the nodes of all lines, cf. Fig. 3.

Many requirements on train schedules can be modeled by so-called "periodical interval constraints". Here is a short example. At some station travelers want to change from line 1 to line 2. Therefore, in a feasible train schedule, the difference between the arrival of a train of line 1 and the departure of the corresponding train of line 2 has to stay within a certain interval. If the

![Fig. 3. Network, line plan and schedule for one line](image_url)
difference is too small, travelers may fail to reach the train of line 2. On the other hand, if the difference is too large, the waiting time at the station will be inconvenient.

Fig. 4 shows this situation for a period of $T = 60$ minutes. If 8–15 minutes are considered to be suitable, then 68–75 minutes are suitable as well, since arrivals and departures repeat every hour.

We model such a “periodical interval constraint” in the following way. If $a^S_1$ is the arrival time of some train of line 1 at station $S$ then we know that the arrival time of any train of line 1 at station $S$ is $a^S_1 + z \cdot 60$ for some $z \in \mathbb{Z}$. Similarly, the departure time $d^S_2$ of some train of line 2 at station $S$ fixes the departure times of all trains of line 2. Therefore, the time difference requirement for all trains can be modeled by

$$8 \leq d^S_2 - a^S_1 - z \cdot 60 \leq 15, \quad z \in \mathbb{Z}. \quad (1)$$

Many other requirements can be modeled in a similar way, e.g. headway times between trains on the same track and travel times of trains between stations. The problem of finding a feasible schedule, i.e. a schedule satisfying a set of periodical interval constraints is called PESP (Periodical Event Scheduling Problem) and was introduced in Serafini and Ukovich [14] in 1989. There are several different algorithmic approaches for solving PESPs, e.g. implicit enumeration methods in [13,14] and a cutting plane method in Odijk [12]. Voorhoeve’s algorithm [15] is based on a constraint propagation.

For feasible train schedules, several objectives have been discussed in the literature. Minimizing total travel time is important in order to attract travelers. If train delays are essential, robustness of connections may well be as important. An obviously different point of view asks for the maximization of profits or the minimization of costs. Cordeau et al. [6] survey optimization models for train routing and scheduling.

Here, we propose a new model combining the PESP with the minimization of certain operational costs that were introduced by Claessens [4,5]. We
suppose that the railroad company has the choice of assigning certain train types to the lines. A train type is characterized by speed, coach capacity, and bounds on the number of coaches, and cost. Here, investment cost per coach and engine as well as mileage dependent cost are included. Faster trains decrease the connection times along a line. If choosing some cheap train type for a line leads to an unavoidable conflict, a more expensive choice may be necessary to find feasible schedules. In our model, we minimize the cost of train type assignment for feasible schedules.

The operational costs of a schedule depend on the number of required engines and coaches as well as on the covered distance. We assume that a train cycles on one line only. The average speed of a train and the length of a line provide an estimate on the time required for one cycle around the line. If \( t \) denotes this cycle time, then \( \lceil t/T \rceil \) trains are required to serve the line. The number of coaches of a train assigned to a line must be large enough to carry all the travelers on the corresponding paths.

In mathematical terms, we propose a mixed integer linear programming (MIP) model containing a large number of variables, i.e.

\[
\begin{align*}
  x_{r,\tau} & \quad \text{train type } \tau \in \mathcal{T}_r \text{ is used for line } r \in \mathcal{R} \text{ (binary)} \\
  w_{r,\tau} & \quad \text{number of coaches of type } \tau \text{ for trains of line } r \in \mathcal{R} \text{ (integer)} \\
  a_{r,\mu} & \quad \text{arrival time of one train of line } r \text{, with direction } \mu \text{, at station } v \\
  d_{r,\mu} & \quad \text{departure time of one train of line } r \text{, with direction } \mu \text{, at station } v \\
  z & \quad \text{vector of integers for the PESP constraints}
\end{align*}
\]

where \( \mathcal{T}_r \) denotes the set of train types that can be assigned to a line \( r \in \mathcal{R} \). The direction \( \mu \) can either be 0 or 1, corresponding to the two directions of each line.

The complete model, referred to as minimum cost scheduling problem (MCSP), is shown in Fig. 5. We will explain its parts in some detail. The objective function sums all costs. Its left part contains the fix costs. The estimated cycle time of train type \( \tau \) along line \( r \) is denoted by \( \hat{t}_{r,\tau} \). \( C^\text{fix}_\tau \) denotes the fix cost for one engine of type \( \tau \), \( C^\text{fix}_\tau \) the fix cost for one coach. The right part of the sum contains the mileage cost (per kilometer). Here, \( d_r \) denotes the length of line \( r \).

With the first constraint class, we assure that the capacity of trains is sufficiently large to carry all travelers. The capacity of one train is the capacity of one coach, denoted by \( \mathcal{C}_\tau \), multiplied by the number of coaches \( w_{r,\tau} \) of the train. Summation over all trains describes the available capacity.

With the second constraint class, we assure that the number of coaches for a train is feasible. The respective lower and upper bounds are denoted by \( \underline{W}_r \) and by \( \overline{W}_r \). In particular, if no train of type \( \tau \) is assigned to the line \( r \), i.e. \( x_{r,\tau} = 0 \), then no coach for this train can be chosen, i.e. \( w_{r,\tau} = 0 \).

With the third class of constraints, we assure that exactly one train type is chosen for each line. All three classes of constraints involve only train types and numbers of the coaches of these trains.
minimize
\[ \sum_{r \in R} \sum_{\tau \in T_r} \left( \frac{l_{r,\tau}}{T} \cdot (x_{r,\tau} \cdot C_{r,\tau}^{\text{fix}} + w_{r,\tau} \cdot C_{r,\tau}^{\text{fixC}}) + d_{r,\tau} \cdot (x_{r,\tau} \cdot C_{r,\tau}^{\text{km}} + w_{r,\tau} \cdot C_{r,\tau}^{\text{kmC}}) \right) \]

subject to
\[ \sum_{r \in R} \sum_{\tau \in T_r} e_{\tau} \cdot w_{r,\tau} \geq N_e \quad \text{for all } e \in E \]
\[ W_{\tau} \cdot x_{r,\tau} \leq w_{r,\tau} \leq W_{\tau} \cdot x_{r,\tau} \quad \text{for all } r \in R \text{ and } \tau \in T_r \]
\[ \sum_{\tau \in T_r} x_{r,\tau} = 1 \quad \text{for all } r \in R \]

and to PESP (many periodical interval constraints),

\[ \text{and to } x_{r,\tau} \in \{0, 1\} \quad \text{for all } r \in R \text{ and } \tau \in T_r \]
\[ w_{r,\tau} \in \mathbb{Z} \quad \text{for all } r \in R \text{ and } \tau \in T_r \]
\[ a_{r,\mu}^v \in \mathbb{R} \quad \text{for all } r \in \mathcal{R}, v \in r, \mu \]
\[ d_{r,\mu}^v \in \mathbb{R} \quad \text{for all } r \in \mathcal{R}, v \in r, \mu \]

\[ z \text{ vector of integer variables for PESP} \]

Fig. 5. The minimum cost scheduling problem (MCSP)

For fixed assignment of train types to lines, all other constraints are periodical interval constraints that we already mentioned before (cf. Fig. 4) describing a PESP. In particular, the arrival and departure times of all trains assigned to lines are subject to constraints similar to inequality (1). For each such constraint an integer variable models the periodicity of the schedule as in inequality (1). All these and all further integer variables in the remaining PESP constraints (e.g. waiting times at stations, headway etc.) are collected in a huge vector \( z \).

The MCSP contains subproblems that are very difficult to solve. The optimal choice of train types and the optimal choice of numbers of coaches are both NP-hard problems, and the generation of a feasible schedule (the PESP) is NP-complete (cf. [7]). On the other hand, the MIP model precisely states which train schedule we consider as feasible and which cost we want to minimize. Moreover, the MIP formulation enables easy addition of other requirements, e.g. available number of engines and coaches of a certain type, without destroying the structure of the model. The high flexibility of MIP models is very helpful when modelling and solving real world applications.
3 Solution Algorithm

The MCSP model is intended for strategic and tactical railroad planning, i.e. for long or medium term decisions rather than for day-by-day operations. In order to provide a helpful evaluation tool for different network or train type scenarios, only short computation times, say some minutes, are acceptable when solving the model for real world MCSP instances. Furthermore, the planner prefers to use standard computer equipment. For our test data like the InterCity or the InterRegio network of Germany and the Netherlands, the direct application of some commercial MIP solver resulted in several hours or even days of computation time on a 400 MHz Pentium II PC. For some instances, its main memory of 256 MB was not sufficient. Obviously, one has to find a more sophisticated approach for solving MCSP.

Our approach is based on decomposition. We consider the matrix of all objective function coefficients and all constraint coefficients of the MCSP in Fig. 6. Gray colored blocks will contain nonzero entries for the respective variables and the corresponding class of constraints.

Variables

\[
\begin{array}{cccc}
  w & x & z & a & d \\
\end{array}
\]

- objective function
- traveler capacity
- number of coaches
- exactly one train type
- travel time data
- other PESP constraints

\[
\begin{array}{c}
\{ \text{MCTP} \} \\
\{ \text{(for fixed train types) FSP} \}
\end{array}
\]

Fig. 6. Structure of objective function and constraints of the MCSP

With the exception of the \( x \)-variables in the travel time data constraints, the matrix can be divided into a two-block diagonal matrix. The first block represents the problem of minimizing costs subject to requirements on the trains that can be assigned to lines. This subproblem will be called *minimum cost train problem (MCTP)*. Its optimal solution assigns train types to the lines and defines the number of coaches in each assigned train or, shortly, it assigns trains to lines. At this point, it is not clear whether a feasible train schedule for these trains exists at all. For its existence, the number of coaches of the train plays no role. Therefore, the \( w \)-variables do not occur in the constraints describing feasible schedules. However, *for some fixed train type assignment*, these constraints form the second subproblem. This *feasible schedule problem (FSP)* is a PESP.
For solving MCSP, we propose a branch-and-bound method based on these subproblems. In particular, the MCTP is used as relaxation of the MCSP.

**Branch-and-Bound Method**

The method is shown in Fig. 7. Each branch-and-bound node represents an MCSP for which only a certain subset of train type assignments is allowed. In each node, we solve the corresponding MCTP relaxation. If the costs for the optimal MCTP train assignment are less than the costs of the currently best known MCSP train assignment, we try to find a feasible schedule of the MCTP train assignment. In other words, we solve the FSP for the fixed train type assignment defined by the optimal train assignment of the MCTP. If we find such a feasible train schedule, then we have a better MCSP train assignment. Otherwise, we branch on conflicting lines. For infeasible FSP, we determine a set of lines that cause infeasibility. For each of these lines, we generate a new MCSP for which the train type assigned to this line in the optimal solution of the previous MCTP is forbidden. For details, we refer to [8] and [10].

**Solving MCTPs**

Unfortunately, when directly applied to MCTP of real world size, commercial solvers may still need many hours of computation time. For its repeated use

---

**Fig. 7.** Branch-and-bound method for MCSP solution
Train Schedule Optimization

\[
\begin{align*}
\min \sum_{r \in \mathcal{R}} \sum_{\tau \in \mathcal{T}_r} \sum_{c=W_{\tau}} \left(\frac{[t_{r,\tau}]}{T} \cdot (C_{r}^{\text{fix}} + c \cdot C_{r}^{\text{fix}C}) + d_{\tau} \cdot (C_{\tau}^{\text{km}} + c \cdot C_{\tau}^{\text{km}C})\right) w_{r,\tau,c}
\end{align*}
\]

\[
\sum_{r \in \mathcal{R}, \tau \in \mathcal{T}_r} \sum_{c=W_{\tau}} w_{r,\tau,c} \geq N_e \text{ for each } e \in E
\]

\[
\sum_{\tau \in \mathcal{T}_r} w_{r,\tau,c} = 1 \quad \text{for each } r \in \mathcal{R}
\]

**Fig. 8.** Binary variable formulation of the MCTP

within a branch-and-bound scheme, we propose a reformulation with the following binary variables

\[w_{r,\tau,c}\] line \(r\) uses train type \(\tau\) with \(c\) coaches

Due to the definition of the binary variables, the constraints for the numbers of coaches are implicitly handled. Therefore the resulting binary MIP model, cf. Fig. 8, has less constraints. Though the number of variables is considerably increased in comparison with the previous model, we are able to solve this binary model much faster since the corresponding LP-relaxation is provably better.

Some further acceleration is possible using suitable classes of cutting planes. As a simple example for such a constraint, we consider a network edge with \(N_e\) travelers which is served by exactly two lines, say line 1 and line 2. Then, for each \(d \in \{1, \ldots, N_e\}\) either line 1 carries at least \(d\) travelers or line 2 carries at least \(N_e - d\) travelers. Therefore, there has to be at least one feasible “train and coach” assignment with sufficient capacity to one of these lines. In terms of the “train and coach” assignment variables \(w_{r,\tau,c}\), we may describe this implied requirement by the linear inequality

\[
\sum_{(\tau,c) \text{ feasible for line 1}} w_{1,\tau,c} + \sum_{(\tau,c) \text{ feasible for line 2}} w_{2,\tau,c} \geq 1. \quad (2)
\]

**Solving FSPs**

Solving an FSP is equivalent to finding an integral solution to a set of linear inequalities. Again, the straightforward application of some commercial MIP solver leads to unacceptably high computation times. On the other hand, we already observed that FSPs are PESPs. There are several combinatorial algorithms for solving a PESP, e.g. the algorithm of Serafini and Ukovich. An enumeration scheme for all possible values for the \(z\)-variables forms the
crucial part of this algorithm. The enumeration scheme proposed by Serafini and Ukovich is only suitable for small PESPs. We have developed a more effective scheme so that instances of the size of our FSPs can be solved in an acceptable amount of time. For details and more information on PESP algorithms, we refer to [9].

In view of the use of FSPs in the branch-and-bound method, it is important to analyze infeasible FSPs. Infeasibility leads to a branching step with the generation of new subproblems. In order to keep the size of the branch-and-bound tree as small as possible, it seems to be preferable to generate only few subproblems. In the algorithm of Serafini and Ukovich, one can sometimes easily identify a small set of train assignments to lines which imply infeasibility.

4 Results

Test data for our algorithms are real networks from the German railroad company Deutsche Bahn (DB) and from the railroad company of the Netherlands Nederlandse Spoorwegen (NS). For the German railroad, we used network data, line plans, origin destination matrices and cost data for the InterCity (IC) and InterRegio (IR) networks. For the Dutch railroad, we obtained the respective data for the InterRegio (IR), InterCity (IC) and AggloRegio (AR) supply networks.

Some characteristics of these instances are displayed in Table 1. For all instances, 4 different train types have been considered.

As an example, the InterCity network of the Netherlands is shown in Fig. 9. Cost optimal lines for this and other Dutch networks were generated in [3].

In order to guarantee suitable arrival and departure times for travelers changing trains, we studied the influence of respective periodical interval constraints at important stations, cf. Table 2.

With the exception of the Dutch AggloRegio network, which is a very large network, our branch-and-bound approach is very fast. Even when we enforce arrivals and departures suitable for travelers changing trains (by adding 40 periodical interval constraints), we generate nearly optimal train schedules within 5 min.

<table>
<thead>
<tr>
<th></th>
<th>DB-IC</th>
<th>DB-IR</th>
<th>NS-IC</th>
<th>NS-IR</th>
<th>NS-AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>90</td>
<td>297</td>
<td>36</td>
<td>38</td>
<td>122</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>107</td>
<td>384</td>
<td>48</td>
<td>40</td>
<td>134</td>
</tr>
<tr>
<td>Number of Lines</td>
<td>31</td>
<td>89</td>
<td>25</td>
<td>21</td>
<td>117</td>
</tr>
<tr>
<td>Average number of edges per line</td>
<td>7.5</td>
<td>5.9</td>
<td>5.0</td>
<td>5.8</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 1. Some characteristics of the railroad networks
**Fig. 9.** InterCity network of the Netherlands

<table>
<thead>
<tr>
<th>Instance</th>
<th>DB-IC</th>
<th>DB-IR</th>
<th>NS-IC</th>
<th>NS-IR</th>
<th>NS-AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of added constraints</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Verified optimum found in</td>
<td>219 s</td>
<td>4 s</td>
<td>30 s</td>
<td>33 s</td>
<td>0:47 h</td>
</tr>
<tr>
<td>Optimality gap after 5 minutes</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>-*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instance</th>
<th>DB-IC</th>
<th>DB-IR</th>
<th>NS-IC</th>
<th>NS-IR</th>
<th>NS-AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of added constraints</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Verified optimum found in</td>
<td>9:31 h</td>
<td>122 s</td>
<td>14:00 h</td>
<td>1:20 h</td>
<td>1:24 h</td>
</tr>
<tr>
<td>Optimality gap after 5 minutes</td>
<td>0.10%</td>
<td>0%</td>
<td>0.39%</td>
<td>0.27%</td>
<td>-*</td>
</tr>
</tbody>
</table>

*no solution in 5 minutes

**Table 2.** Results for test instances without and with added constraints
For larger networks, regional decomposition may be promising. At first, the network is partitioned into several regional networks which allow fast optimization. Then, the corresponding partial train schedules are combined to a train schedule of the complete network. The combination step will require some interactive corrections which may be supported by computerized tools. In manual schedule planning, regional decomposition is a well established approach.

5 Conclusions

With our method for minimizing costs of train schedules, traffic planners may interactively generate or evaluate different scenarios. We implemented a prototype that visualizes trains in the network, cf. Fig. 10. In particular, we display the current positions of all trains in the network. Trains are repre-

![Fig. 10. Visualization of a schedule (screenshot)](image-url)
presented by squares numbered with the corresponding line. Different colors of squares indicate different train types.

As already mentioned, schedule planning is only one subproblem of traffic planning. Traditionally and with good reasons, the highly complex planning process is hierarchically decomposed, cf. Fig. 11. An survey on the application of mathematical programming methods for these subproblems is given in [1]. The solution of the line planning problem has already been the subject of a former BMBF-funded project (see [2]).

Of course, it is well known that decisions on different levels are not at all independent. For example, any change in the transportation service, e.g. a new line plan or a new schedule, will influence the travelers' behavior. Travel demand data used in the development of a schedule will not match travel demand after the introduction of that schedule. In order to handle such difficulties, schedule planners try to simulate the travelers' behavior and to estimate the actual effect of a new schedule. They iteratively generate a schedule based on demand data, then estimate the new demand resulting from the schedule etc. until changes in schedule and demand data become insignificant. Unfortunately, no mathematical argument is known that guarantees some kind of convergence for such an approach.

Acknowledgements. We would like to express our thanks for encouragement, support and data made available by Adtranz Signal (Braunschweig), Dr. M. Krista (Lineas, Braunschweig) and Prof. Dr. L. Kroon (Railned, the Netherlands).

References


An Integrated Planning Approach for Cellular Radio Networks

Rudolf Mathar and Michael Schmeink

Aachen University of Technology, 52056 Aachen, Germany

Abstract. Two major planning problems must be solved in the design phase of a cellular radio network. The initial question is where to locate the base transmitter stations such that full coverage is achieved at low interference. This is relevant for frequency division (FDMA) as well as code division multiple access (CDMA) technology. To tackle this problem, precise knowledge of the radio wave propagation from each candidate base station location is necessary, particularly in urban areas. A corresponding software tool has been developed in the present project. Its main advantages are high precision and fast computation times, which is achieved by a sophisticated preprocessing phase. Once the base station positions are fixed, for FDMA systems frequencies have to be assigned such that in each cell there is a sufficient number of channels available at a low overall interference in the network. Since cell site selection and frequency allocation have mutual influence on each other, the ultimate goal is to deal with both problems in a single design step. In this paper, the above planning issues are modelled as linear integer programs and approximate solution methods for the corresponding NP-hard problems are discussed. It turns out that optimal solutions can be found for realistic problem sizes, which makes the developed software a versatile tool for third generation radio network planning.

1 Introduction

Radio waves are used for mobile communication in cellular networks. Because of the scarce available radio spectrum, spatial frequency reuse is necessary to accommodate large numbers of subscribers. This is achieved by putting up a large number of so called base transmitter stations (BTS) with relatively low transmission power. To cover the area of Germany, e.g., presently about 10,000 base stations are used in each of the existing GSM (global system for mobile communication) networks. Each active mobile station is connected to one of the BTS, usually the one with the least attenuated path loss.

Mainly two different radio transmission technologies are presently used to further share the available radio spectrum between users: frequency division multiple access (FDMA) with a time division (TDMA) component, and code division multiple access (CDMA). In CDMA, all mobiles in a cell use the same frequency spectrum for transmission, the signals of different users are separated by orthogonal code sequences, which spread the signal in a controlled way over the available bandwidth. FD/TDMA is used for the presently very successful GSM networks. Here, the available spectrum is splitted up into
Fig. 1. The radio network planning process

separate frequencies. Furthermore, each frequency is divided into time slots (eight slots per frame in GSM) which are periodically assigned and used for transmission. Because of its higher data rate and flexibility, CDMA technology will be the main component in forthcoming third generation mobile networks. To achieve very high data rates of 2 MB per second a mixed system of FD/TDMA with CDMA on top will be used.

The radio network planning process, either for frequency/time or code division multiple access, is a mathematically challenging problem. Figure 1 gives an overview of what the main input data are, and which dependencies occur in the planning hierarchy. The basic input is formed by terrain, land-use, and population data of an area to be supplied. On a numerical tool level these data are used to predict the radio wave propagation and to obtain a rough estimate of the expected traffic per area unit. The major planning steps now consist in selecting appropriate cell sites, to allocate channels for FDMA systems, whenever the positions of the base stations are fixed, or otherwise, to integrate both steps in a single setup.

For GSM networks, cell site selection and channel assignment are usually carried out in separate steps. First cell sites are chosen according to certain coverage criteria, hereafter channels are allocated. This of course reduces the complexity of the planning procedure, however, some degrees of freedom are lost concerning the optimality of a simultaneous integrated setup.

In this project, we have mainly contributed to radio wave propagation prediction, the cell site selection problem, and the integrated setup. According results will be presented in the following. Section 2 deals with a sophisticated algorithm which uses a preprocessing step to allow for an extremely fast radio wave propagation prediction. Computational efficiency of the prediction is important to tackle the cell site selection problem, since for each candidate
base station configuration a complete propagation prediction is necessary to obtain the input data for the corresponding optimization problem.

In Sect. 3, the simplest form of channel allocation with fixed base station locations is treated in order to introduce the basic methodology for what follows. The cell site selection problem is treated in Sect. 4. Numerical results are derived for some selected realistic scenarios with a branch-and-bound setup. Finally, the integrated cell site selection and channel allocation problem is introduced in Sect. 5.

2 Preprocessed Ray Launching for Radio Wave Propagation Prediction

In this section, an algorithm for predicting the field strength of transmitters in urban areas is presented. Field strength prediction is the main prerequisite to decide whether points are covered by a sufficient radio signal, and to determine the mutual interference between base stations. Since this information is repeatedly needed in the subsequent cell site selection and frequency assignment problems, a competitive software tool has been developed in the present project.

There are two main approaches to field strength prediction in urban areas: empirical and deterministic models, the latter composed of ray-tracing or ray-launching methods [1]. Empirical models have short computation times, however, an extensive parameter calibration is necessary to obtain reliable results. Deterministic models take account of waveguiding effects at building corners and easily provide additional information like impulse responses and building penetration. Usually long computation times are needed in order to obtain accurate results. Appropriate preprocessing is a way to reduce computation times significantly [4].

In ray-launching algorithms, rays are emitted from the transmitter and are possibly reflected and diffracted. Because of the forward-analyzing structure of the algorithm, only significant ray paths are tracked. However, since rays are emitted in discrete angular steps, areas far away from the transmitter are less frequently visited than areas of the same size in the vicinity of the transmitting station. Moreover, diffraction sources might be ignored. Both effects produce misleading predictions. To avoid this, we introduce a fast and accurate new 3D ray-launching algorithm that is raster oriented.

The preprocessing consists of two steps. First, a three-dimensional grid of cubes is generated from a database containing building and terrain height information. Subsequently, all visible neighbours of particular cubes are stored. The purpose is to pass diffraction information on to diffraction sources in higher recursion depths. Both steps merely handle geographical information and are independent of particular transmitter locations.

Buildings are represented by ground plans and the corresponding height, they are hence modeled as polyhedra consisting of the outside walls with
a flat roof on top. After choosing the desired grid resolution, each cube is
either unfilled or filled according to the classes 'ground', 'building wall', and
'vertical or horizontal building edge'. Each building wall cube is also endowed
with the appropriate normal vector pointing outwards. Cubes at the fringe of
the grid are called 'bordering'.

Since diffraction at building edges is the most important non-line-of-sight
propagation path, for each horizontal building-edge cube (BEC) any other
horizontal BEC in line of sight is determined and stored. For this purpose, a
ray is launched from each BEC to every bordering cube. Rays are described by
a starting cube and a main direction. The cube with the minimum deviation
from the main direction is the one to be visited next by the ray. The same
procedure is applied to the vertical building-edge cubes. In summary, this
leads to the following preprocessing algorithm:

\textit{Preprocessing:}

For each horizontal/vertical BEC $c$: 
For each bordering cube:

Launch a ray from BEC $c$ to the bordering cube.

For each cube $d$ intersected by the ray:

If $d$ is a horizontal/vertical BEC:

BEC is in line of sight to $c$ and stored.

If $d$ is a filled or bordering cube:

Stop the ray.

End.

After the preprocessing step a transmitter location can be selected, jointly
with a corresponding prediction area. First, each cube in line of sight to
the transmitter is determined by intersecting rays launched to the bordering
cubes. The cube's field strength $P_r$ is determined according to free space
propagation as follows. Let $d$ denote the distance to the transmitter, $P_t$ the
transmission power, $G_t$ and $G_r$ respective antenna gains, and $\lambda$ the wave-
length in meters. The field strength is then given by

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2}.$$ 

If a cube in line of sight is filled, reflection occurs and a new ray is
launched. If diffraction applies at a BEC in line of sight, $n$ rays are launched
in the diffraction cone. For every neighbouring BEC it is determined whether
diffraction continues. This is done recursively up to a maximum recursion
depth, or until the field strength falls short of a certain threshold. At each
neighbour and each cube intersected by a ray the field strength is calculated
according to free space propagation along the respective path. Rays may
further be reflected at filled cubes. Any ray is stopped at bordering cubes
or whenever its field strength becomes too small. For the calculation of a
reflection direction or a diffraction cone the appropriate wall's normal vector is used. In summary, we end up with the following algorithm:

**CORLA - 3D Cube Oriented Ray-Launching Algorithm:**

For each bordering cube:
- Launch a ray from the transmitter to the bordering cube.
  - For each cube intersected by the ray:
    - Cube is in line of sight to the transmitter.
    - If the cube is filled or a bordering cube is reached:
      - Stop the ray.
  - For each cube in line of sight to the transmitter:
    - If the cube is in the receiving plane:
      - Calculate line of sight field strength at the cube.
    - If the cube is filled (reflection):
      - Launch a ray.
    - If the cube is a BEC and diffraction occurs:
      - Do **diffraction**:
        - Determine the diffraction cone.
        - Launch \( n \) diffraction rays into this cone.
        - For every neighbour of the BEC inside the diffraction cone:
          - If diffraction at the neighbour occurs:
            - Do **diffraction** ... until a certain recursion depth is reached or field strength is too low.

For each ray:
- For each cube intersected by the ray:
  - If the cube is in the receiving plane:
    - Add up the field strength of the ray at the cube.
  - If the cube is filled (reflection):
    - Launch a ray.

End.

The following figures and tables refer to predictions in the city of Munich with a transmitter placed well below rooftops. Table 1 lists computation times for the preprocessing of an area of 3.5 km\(^2\) for different cube sizes. Table 2 displays running times on a standard PC for the prediction of different areas.

<table>
<thead>
<tr>
<th>Cube Size</th>
<th>1 m</th>
<th>2 m</th>
<th>3 m</th>
<th>5 m</th>
<th>10 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 m (\times) 200 m</td>
<td>0:00:31</td>
<td>0:00:09</td>
<td>0:00:03</td>
<td>0:00:02</td>
<td>0:00:02</td>
</tr>
<tr>
<td>400 m (\times) 400 m</td>
<td>0:02:48</td>
<td>0:00:51</td>
<td>0:00:23</td>
<td>0:00:19</td>
<td>0:00:12</td>
</tr>
<tr>
<td>800 m (\times) 800 m</td>
<td>0:30:20</td>
<td>0:09:12</td>
<td>0:03:20</td>
<td>0:02:12</td>
<td>0:01:10</td>
</tr>
</tbody>
</table>

**Table 1.** Computing times for different cube lengths in h:min:sec (appr. values)
Table 2. Preprocessing times for different cube lengths in h:min:sec (appr. values)

<table>
<thead>
<tr>
<th>Cube Length</th>
<th>1 m</th>
<th>2 m</th>
<th>3 m</th>
<th>5 m</th>
<th>10 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 km$^2$</td>
<td>12:37:00</td>
<td>2:22:00</td>
<td>1:10:00</td>
<td>0:35:00</td>
<td>0:14:00</td>
</tr>
</tbody>
</table>

Fig. 2. A typical prediction with a cube length of 2 m, field strength is coded by colours (red=intense, yellow=medium, green=weak)

using different cube sizes. Figure 2 shows the predicted field strength with a cube length of 2 m. The transmitter is located in the center of the area. Field strength is coded by colours from red (very intense) via yellow (medium field strength) to dark green (very low).

3 Channel Allocation

In order to introduce the methodology and basic notation we start with the formulation of channel allocation adopted from [6]. Assume that there are $N$ physical frequencies, labeled 1, \ldots, $N$, and $z$ cells, each cell identified with the corresponding base station. A channel design is described by a binary
matrix \( X = (x_{ik}) \), \( i = 1, \ldots, z \), \( k = 1, \ldots, N \), with the meaning

\[
x_{ik} = \begin{cases} 
1, & \text{if channel } k \text{ is allocated to cell } i, \\
0, & \text{otherwise.}
\end{cases} \tag{1}
\]

The amount of interference between cells \( i \) and \( j \) on the same frequency is usually known from a radio wave propagation tool (see Sect. 2) or from real field measurements. It is usually expressed as the so-called interference probability, which denotes the relative area between a pair of cells with a too low carrier to interference ratio. The analogous measure of interference is also calculated for adjacent channels. If the interference probability between two cells on the same channel or adjacent channel exceeds a certain threshold, 4% say, then the corresponding channel or pair of adjacent channels must not be used simultaneously in both cells. This leads to the compatibility matrix \( C = (c_{ij}) \), \( i, j = 1 \ldots, z \), where \( c_{ij} \) denotes the minimum channel distance between cell \( i \) and cell \( j \). Compatibilities can be easily expressed by a set of linear constraints in the binary variables \( x_{ik} \) by

\[
x_{ik} + x_{jl} \leq 1 \quad \text{for all } (i, k) \neq (j, l) \text{ with } |k - l| \leq c_{ij}.
\]

There are further problem constants that must be taken into account. \( r_i, i = 1, \ldots, c \), denotes the number of channels required for cell \( i \). \( r_i \) is determined by Erlang’s B-formula as the minimum number of channels not to exceed a certain blocking probability. Furthermore, there are sets of banned and prescribed channels per cell, expressed by the sets \( B_i, P_i \subset \{1, \ldots, N\}, i = 1, \ldots, z \).

The objective of maximizing the number of allocated channels without violating any compatibility constraints, and not allocating more than \( r_i \) channels in cell \( i \) may be written as the following integer linear program.

\[
\text{maximize} \quad \sum_{i=1}^{z} \sum_{k=1}^{N} x_{ik} \tag{2}
\]

such that

\[
x_{ik} + x_{jl} \leq 1 \quad \text{for all } (i, k) \neq (j, l) \text{ with } |k - l| < c_{ij} \tag{3}
\]

\[
\sum_{k=1}^{N} x_{ik} \leq r_i \quad \text{for all } i = 1, \ldots, z \tag{4}
\]

\[
x_{ik} \in \{0, 1\} \quad \text{for all } i = 1, \ldots, z, \quad k = 1, \ldots, N \tag{5}
\]

\[
x_{ik} = 0, \quad k \in B_i, \quad x_{ik} = 1, \quad k \in P_i \quad \text{for all } i = 1, \ldots, z \tag{6}
\]

If there is a solution for which all inequalities (4) hold with equality, then there exists an optimal solution which fulfills all requirements \( r_i \). On the other hand, for an optimal solution of (2)–(6), the difference between the objective function and \( \sum_i r_i \) indicates how many requirements cannot be satisfied under the given compatibility constraints.
Obviously, as an extension of the graph coloring problem, the above problem is NP-hard. For large problem sizes heuristics must be used to find solutions close to the optimum. A very powerful procedure which iteratively optimizes small subproblems is introduced and tested in [6].

The above linear program introduces the basic methodology used in the following. For real applications, however, further objectives must be taken into account. Usually there are two classes of channels with different quality of service requirements, namely broadcast control channels (BCCH) and traffic channels (TCH). Furthermore, coherence bandwidth constraints are of importance, particularly if frequency hopping is applied. The final objective is to minimize the total interference sum in the class of all channel designs which satisfy the above constraints. Papers [3] and [6] deal extensively with these extensions, all additional constraints and objectives are embedded in an integer linear programming setup.

4 Cell Site Selection

We employ a discrete traffic model which represents spatial service demand by marked points. This can be done at varying granularity. Having no a-priori knowledge of the spatial traffic distribution at all, a regular point grid is sufficient to represent mobile stations demanding for communication capacity. A more refined model would be to represent each active mobile by a point with a mark indicating the amount of data traffic per time unit, e.g., rates induced by voice or data traffic. Corresponding spatial point processes can be used as typical snapshots of such patterns. However, usually the number of points in this approach is too large to allow for numerical solutions of the network optimization problems. In order to decrease numerical complexity, certain clustering or agglomeration methods can be applied. This approach is pursued in [9] and [11]. In the following, a point in the area of interest with a mark indicating the corresponding amount of service requirement is called a traffic node (TN).

The cell site selection problem, particularly for CDMA, has been addressed in a number of papers. Objective functions range from simply covering a given area to including traffic requirements by endowing base stations with weights. Solution methods are based on genetic algorithms, see [2], [5], as well as on simple greedy heuristics, see [8], [12]. [7] treat different aspects of the cell site selection problem by a linear programming approach. The performance of greedy heuristics and simulated annealing is compared in [10].

We assume that the service demand in the relevant area is represented by \( q \) traffic nodes, each endowed with a weight \( w_t, t = 1, \ldots, q \), indicating the amount of traffic connected to this node. Furthermore, a number \( p \) of candidate base station configurations is given. Different BTS configurations may have the same position, but different power or antenna height.

The aim is to select at most \( K \) base stations out of this pool such that a maximum of traffic nodes can be connected. For this purpose, the coverage
of each potential base station at each traffic node is needed. Let

\[ V = (v_{it})_{i=1, \ldots, p, t=1, \ldots, q} \]  

(7)
denote the matrix of field strength of candidate base station \( i \) at traffic node \( t \) on the same frequency (co-channel field strength). Similarly, matrix

\[ V' = (v'_{it})_{i=1, \ldots, p, t=1, \ldots, q} \]  

(8)
combines the field strength of candidate base station \( i \) at traffic node \( t \) on adjacent frequencies (ad-channel field strength). \( V' \) is needed if different frequencies are used for transmission.

A connection from base station candidate \( i \) to traffic node \( j \) can be established only if the received power is greater than a certain threshold \( \beta \), say. This is expressed by the derived binary matrix \( S = (s_{it})_{i=1, \ldots, p, t=1, \ldots, q} \) with

\[ s_{it} = \begin{cases} 1, & \text{if } v_{it} \geq \beta, \\ 0, & \text{otherwise}. \end{cases} \]

Base station's \( i \) signal at traffic node \( t \) can be correctly decoded if \( s_{it} = 1 \), and moreover the total field strength of all interfering base stations does not exceed the usable signal power by a certain amount \( \gamma \). This phenomenon is called the capture effect.

Since the threshold \( \gamma \) is independent of the particular scaling of \( v_{it} \) and \( v'_{it} \), both sets of field strength values may be normalized as to satisfy

\[ \sum_{t=1}^{q} (v_{lt} + 2v'_{lt}) \leq \frac{1}{\gamma} \text{ for all } t = 1, \ldots, q. \]  

(9)
Condition (9) will be assumed throughout the paper.

A base station design \((x_1, \ldots, x_p)\) is composed of \( p \) binary variables \( x_i \), \( i = 1, \ldots, p \), with the meaning

\[ x_i = \begin{cases} 1, & \text{if BTS candidate } i \text{ is selected,} \\ 0, & \text{otherwise}. \end{cases} \]

We are now in a position to define whether a radio link can be established to some traffic node.

**Definition 1.** Traffic node \( t \) is called **captured under BS-design** \((x_1, \ldots, x_p)\), if there exists some \( i \) such that

\[ s_{it} = 1 \text{ and } v_{it}x_i > \gamma \sum_{i \neq i} v_{it}x_i. \]
Auxiliary binary variables $u_{it}$ express the fact that traffic node $t$ can be captured by some base station $i$, $i = 1, \ldots, p$, $t = 1, \ldots, q$,

$$u_{it} = \begin{cases} 
1, & \text{if traffic node } t \text{ is captured by BTS candidate } i, \\
0, & \text{otherwise.}
\end{cases} \quad (10)$$

The weights $w_t$ represent the load at traffic node $t$. Maximizing the captured traffic by selecting at most $K$ base stations out of all candidates now reads as follows.

$$\text{maximize } \sum_{i=1}^{p} \sum_{t=1}^{q} u_{it} w_t \quad (11)$$

such that

$$u_{it} \leq v_{it} x_i - \gamma \sum_{t \neq i} v_{lt} x_t + 1 \quad \text{for all } i, t \quad (12)$$

$$u_{it} \leq s_{it} x_i \quad \text{for all } i, t \quad (13)$$

$$\sum_i x_i \leq K \quad (14)$$

$$u_{it}, x_i \in \{0, 1\} \quad \text{for all } i, t \quad (15)$$

Objective function (11) sums up all captured traffic. If $\gamma > 1$, multiple capturing is prohibited by constraint (12), such that soft hand-off to more than one base station is excluded from the optimal solution. By Definition 1 and normalizing condition (9), constraints (12) and (13) ensure that only captured traffic nodes are considered in the objective function. $u_{it} = 1$ holds for the optimum solution iff traffic node $t$ is captured by some base station $i$ and base station $i$ is also selected. Otherwise $u_{ij}$ will be zero.

The above integer linear program has been solved for an example of $p = 329$ base station candidates, $q = 1106$ traffic nodes, and a maximum number of $K = 7$ base stations for a field strength data set from [9], also used in [10]. The optimal solution is depicted in Fig. 3. It shows the convex hulls of traffic nodes captured by the corresponding base stations. Nodes captured by the same BS are tagged by identical symbols (circles, triangles, squares, etc.). The underlying field strength data $V$ has been generated on the basis of a semi-empirical terrain and morphology dependent propagation law for an area surrounding the city of Würzburg (see [9]). Stars indicate the position of the selected base stations. Because of the hilly characteristic not all traffic nodes are captured by the optimal solution, see the small non-marked dots. The optimal solution contains four small cells overlayed by a large one. This effect is due to shading behind ridges. Observe that the selected base stations may of course have different transmit power.

Soft hand-off is a way to improve system capacity, and can be incorporated in the same framework. If base station design $(x_1, \ldots, x_p)$ applies, $\sum_i s_{it} x_i$ counts how many base stations supply traffic node $t$ with sufficient field
Fig. 3. Optimal solution for $p = 329$ candidate BTS, $q = 1106$ traffic nodes and at most $K = 7$ selected BTS

strength. A soft hand-off to base station $i$ with $s_{it}x_i = 1$ can be performed, if additionally the corresponding C/I is large enough. Maximizing the binary varibles $y_t \in \{0,1\}$ such that $y_t \leq \frac{1}{2} \sum_i s_{it}x_i$ means to select a base station design that maximizes the number of possible two-way soft hand-offs. In this sense the following integer linear program favours the number of applicable soft hand-off connections in the network. $\alpha_1$ and $\alpha_2$ balance the objectives of capturing traffic nodes and of allowing for two-way soft hand-off.

$$
\begin{align*}
\text{maximize} \quad & \alpha_1 \sum_{i,t} u_{ij}w_t + \alpha_2 \sum_t y_t \\
\text{such that} & \\
& u_{it} \leq v_{it}x_i - \gamma \sum_{\ell \neq i} v_{\ell t}x_\ell + 1 \quad \text{for all } i,t, \\
& u_{it} \leq s_{it}x_i \quad \text{for all } i,t, \\
& y_t \leq \frac{1}{2} \sum_i s_{it}x_i \quad \text{for all } t, \\
& \sum_i x_i \leq K, \\
& u_{it}, x_i, y_t \in \{0,1\}
\end{align*}
$$

Because of the enormous size of the above program, optimal solutions for realistic problem instances are not yet available.
5 Integrated Cell Site Selection and Channel Allocation

In the classical planning approach for GSM networks, first the positions of base stations are determined and thereafter channels are allocated. Solution methods for each of these steps have been derived in the previous sections. However, improved results can be obtained if cell sites are selected and channels assigned in a single integrated planning step. In the following, this highly complicated task is modeled as an integer linear program.

We adopt the notation $x_{ik} \in \{0, 1\}$ from (1) to denote a channel design. The matrices $V$ and $V'$ from (7) and (8), respectively, contain co- and ad-channel field strengths of candidate base stations at traffic nodes.

**Definition 2.** Traffic node $t$ is called captured under channel design $X = (x_{ik})$ if there is some $i \in \{1, \ldots, p\}$ such that $s_{ij} = 1$ and

$$v_{it} x_{ik} > \gamma \sum_{\ell \neq i} \left( v_{\ell t} x_{\ell k} + v_{\ell t}' (x_{\ell,k-1} + x_{\ell,k+1}) \right)$$

for all $k$ with $x_{ik} = 1$. $x_{t0} = x_{t,N+1} = 0$ are set to cover the boundary cases.

In Definition 2, for any channel $k$ used in cell $i$ the signal strength is compared to the sum of all interferers using the same or the next adjacent channel in surrounding cells.

Let again $w_t$ denote weights representing the amount of service load at traffic node $t$. The auxiliary binary variables $u_{it}$ have the same meaning as in (10), i.e., $u_{it} = 1$ iff traffic node $t$ is captured by base station candidate $i$. $\sum_t u_{it} w_t$ then represents the total traffic carried by candidate $i$. A linear function $f(\sum_t u_{it} r_t)$ is used to determine the number of channels required in cell $i$ for a sufficiently low blocking probability. $f$ should be chosen as to approximate the inverse Erlang B-function.

The following linear program aims at maximizing the carried traffic by simultaneously selecting optimal cell sites and channel assignments.

$$\text{maximize } \sum_{i,t} u_{it} w_t$$

such that

$$u_{it} \leq (v_{it} - 1) x_{ik}$$

$$- \gamma \sum_{\ell \neq i} \left( v_{\ell t} x_{\ell k} + v_{\ell t}' (x_{\ell,k-1} + x_{\ell,k+1}) \right) + 2 \quad \text{for all } i, t, k \quad (17)$$

$$u_{it} \leq \sum_k x_{ik} \quad \text{for all } i, t \quad (18)$$

$$u_{it} \leq s_{it} x_i \quad \text{for all } i, t \quad (19)$$
\[
\sum_{k} x_{ik} \geq f \left( \sum_{t} u_{it} w_t \right) \quad \text{for all } i \tag{20}
\]
\[
\sum_{k} x_{ik} \leq x_i \quad \text{for all } i, k \tag{21}
\]
\[
\sum_{i} x_i \leq K \tag{22}
\]

(17)--(19) impose that \( u_{it} = 1 \) only if traffic node \( t \) is captured by some base station that has at least one channel assigned. Constraint (20) requires that all traffic is served. (21) allows channels to be allocated only to selected base station configurations, and (22) admits at most \( K \) base stations.

Because of the huge number of variables and constraints we are presently not in a position to solve the above integer linear program for realistic problem sizes. Heuristic procedures to provide at least suboptimal solutions will be developed in the future.

6 Conclusions

The present project deals with two important steps in the planning phase of a cellular radio network, namely the cell site selection and the channel assignment problem in an integrated approach. Detailed and iterated information about the radio wave propagation behaviour is necessary as input data for both problems. For this purpose, a fast and accurate field strength prediction algorithm has been developed on the basis of a three-dimensional cube oriented ray launching concept. Its extraordinary speed is attained by a preprocessing phase in which repeatedly needed information is calculated in advance and stored in an easily accessible data structure. Corresponding optimization problems are formulated as integer linear programs. Selected examples of realistic problem size are solved by a branch-and-bound approach.

References


Author Index

G. Albinus 401
E. Bänsch 315
P. Böhm 365
D. Böhnning 533
S. Börm 305
U. Bandelow 377
M. Bebendorf 55
Chr. Becker 43
K. W. Benz 315
S. Bitterlich 152
H. G. Bock 3
J. Borchardt 184
R. Borndörfer 653
S. Boschert 315
T. Boskamp 483
H. Buchner 472
R. Bulirsch 84
R. Bulirsch 393
Ch. Böckmann 162
X.-N. Chen 103
C. Chucholowski 84
P. Dörr 460
C. Dârțu 503
U.-P. Dammann 533
F. Darvas 472
S. de Vries 441
P. Deuflhard 279
G. Deuschl 554
W. Dreyer 233
F. Duderstadt 233
G. Dziuk 315
M. A. Efendiev 175
K. Ehrhardt 184
M. Elf 675
J. Elschner 293
D. Estévez Schwarz 413
U. Feldmann 413
U. Feldmann 429
B. Fiedler 175
T. Friese 279
M. Fuchs 472
J. Fuhrmann 127
M. Günther 429
H. Gajewski 355,401
M. Gerds 74
H. Glaab 609
A. Glitzky 401
M. Grötschel 653
R. D. Grigorieff 252
P. Gritzmann 441
F. Grund 184
C. Gruss 460
D. Hömberg 233
R. Hünlich 401
W. Hackbusch 305
V. Heuveline 33
R. Hinder 293
K.-H. Hoffmann 331
H. Holling 533
M. Holzmann 216
J. Honerkamp 554
R. H. W. Hoppe 365
D. Horn 184
M. Hoschek 429
O. Iliev 243
R. Iza Teran 152
W. Jäger 137,503
M. Jünger 675
R. Köcher 252
V. Kaibel 675
H.-Ch. Kaiser 377
H. Chr. Kaiser 355
T. Kaiser 315
K. Kirchgässner 103
L. Klassen 22
G. W. Klau 584
O. Klein 343
A. Klimmek 22
P. Knabner 152
J. Knopke 401
M. Kočvara 573
Th. Koprucki 377
D. Kröner 22
S. Kulig 263
A. Löbel 653
P. Lang 263
H. Langmach 355
M. Lauk 554
T. Lengauer 63
L. Lerman 175
T. Lindner 703
D. Logashenko 204
R. W. Longman 3
A. K. Louis 460,472
M. E. Lübbecke 690
C.-H. Lücking 554
R. März 413
P. Maass 162
B. Maar 204
U. Malzahn 533
P. Mathé 493
R. Mathar 717
G. Mazurkevitch 365
W. Mehnert 533
A. Mekler 162
M. Metzger 331
P. Mutzel 584
R. Nürnberg 355,623
R. Neubert 393
M. Niggemann 216
M. P. Nowak 623
H.-O. Peitgen 483
H. J. Pesch 74
S. Petrova 365
H. Petry 460
M. Petzold 127
P. Philip 343
A. Pott 609
D. Prätzel-Wolter 263
A. Prechtel 152
W. Römisch 623
Author Index

W. Röpke 401
J. Rademacher 175
R. Rannacher 33, 43
J. Rehberg 377
P. Rentrop 429
R. H. Richter 355
S. Rjasanow 55

D. Saupe 520
W. Schäfer 137
J. P. Schlöder 3
P. Schlattmann 533
M. Schmeink 717
A. Schmidt 315
D. Schmidt 216
F. Schmidt 279
G. Schmidt 293
U. Schmitt 472
E. Schneid 152
R. Schultz 623
V. Schulz 204

R. Schulze 533
A. Schuppert 175
A. Schwarz 393
M. Schäfer 63
S. D. Sharma 103
K. G. Siebert 315
P. Singer 483
K. Soldner 216
J. Sprekels 343
N. Stuntz 103
S. Sturtzel 413

A. Thiele 194
J. Timmer 554
C. Tischendorf 413
L. Tobiska 194
D. Trescher 22
S. Turek 43

M. Vögel 84
K. Velten 243
A. Voigt 331

O. von Stryk 84

G. Wachutka 365
Ch. Wagner 137
M. Wagner 520
M. Westphalen 623
H. Willershausen 137
K. Wilmański 343
M. J. Winckler 3
A. Wirsen 263
G. Wittum 137, 204
E. Wolfgang 365
Th.-M. Wolter 84

J. H. Zacharias-Langhans 493
A. Zemitis 243
U. T. Zimmermann 690, 703
J. Zowe 573
L. Zschiedrich 279