Proceedings

20th Biennial Conference

on

Numerical Analysis

University of Dundee
24–27 June, 2003

Editors

David F Griffiths
G Alistair Watson

Contents

Preface

The 20th Dundee Biennial Conference on Numerical Analysis was held at the University of Dundee during the four days June 24–27, 2003. Around 150 people attended, about one half of those coming from outside the UK. The programme for the meeting contained fourteen invited talks, including the A. R. Mitchell lecture which opened the Conference, given by Stanley Osher from UCLA. In addition there were 103 submitted talks, given in parallel sessions.

We have in the past arranged for commercial publication of full versions of the invited talks. This has meant that not all the talks were published, either because sometimes speakers were not able to meet our deadlines, or they wished to use the material in some other way. These facts, and other considerations such as the timeliness of a printed version, made us decide this year to try something different which, while maintaining a permanent record of the meeting, also gave speakers a greater level of flexbility and informality. Therefore, we asked all the invited speakers if they could make available short versions of their talks, in a form most convenient for them, and these are contained in this technical report. We hope that these will give a flavour of the invited talks, and we would encourage readers who wish to find out more to contact the authors directly.

We are grateful to the invited speakers, all of whom agreed to allow us to produce versions of their talks in this way. We are also grateful to SIAM for permission to reprint a version of Stanley Osher's talk which has already appeared in SIAM Newsletter. Also included here is a list of the titles of all contributed talks given at the meeting, together with authors. The first named author in each case is the person who presented the material.

We would like to take this opportunity of thanking all the speakers, including the after–dinner speaker James Lyness of Argonne National Laboratory, all who chaired sessions, and all who contributed in any way. We are as always also grateful for help received from members of the Department of Mathematics both before and during the Conference. The Conference is also indebted to the University of Dundee for making available various University facilities throughout the week, and for the provision of a Reception for the participants in West Park Hall.

D F Griffiths

G A Watson

October 2003

Invited Talks

The A R Mitchell Lecture 2003

Professor Stanley J Osher,
UCLA
Department of Mathematics
520 Portola Plaza
Los Angeles, CA 90095-15
USA

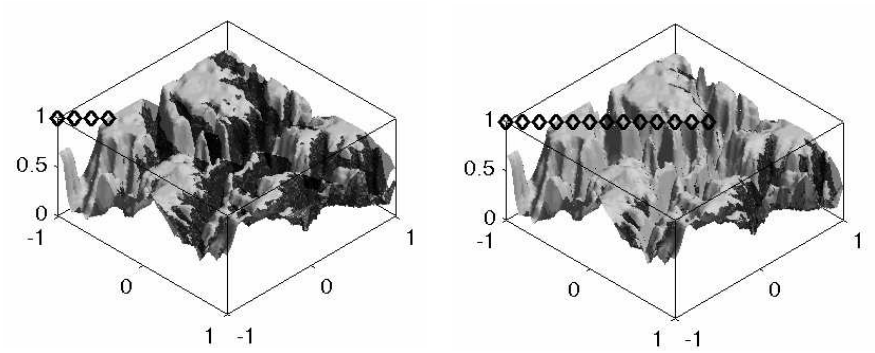
sjo@math.ucla.edu
www.math.ucla.edu/∼sjo

from *SIAM News*, Volume 35, Number 4

Dynamic Visibility and the Level Set Method

By Stanley Osher

The level set method (LSM), invented in 1987 by J.A. Sethian and me, has proved remarkably successful as a numerical (and theoretical) device in a host of applications, many of them in imaging science. The original reference [9] has been cited almost 600 times (according to the ISI Web of Science), and a recent query to Google's search engine gave nearly 2700 responses for "level set methods." The LSM is still an active research area (for recent results, see [7], a book published this year). In a plenary talk at the First SIAM Conference on Imaging Science, I gave a brief overview of the LSM and the associated level set technology and touched on a few imaging applications. Among them was dynamic visibility [11], which is briefly described here.



Simulation, using real data, of an airplane flying through the Grand Canyon. The regions on the occluder that have not yet become visible are dark; the diamond-shaped dots show the path of the airplane.

The problem is easily stated: Given a collection of closed surfaces representing objects in space, determine quickly the regions (in space or on the surfaces) that are visible to an observer. This question is crucial to applications in fields as diverse as rendering, visualization, etching, and the solution of inverse problems. In a 3D virtual-reality environment, knowing the visible region speeds up the rendering by enabling us to skip costly computations on occluded regions.

Representing a family of surfaces implicitly as the zero level set of a *single* function $\varphi(\vec{x})$, $\vec{x} = (x, y, z)$, has several advantages, particularly when things are moving. Topological changes are easily handled (without "emotional involvement"), and geometric quantities, such as normals and curvatures, are also easily computed dynamically. Most published work in computer graphics and computer vision uses explicit surfaces, usually constructed with triangles, but this is changing. The upcoming SIGGRAPH conference (in San Antonio, this July) will have many LSM-related papers and a full-day course on LSM and PDE-based methods in graphics. For a recent detailed report on the visibility problem with explicit surfaces, see [4].

In our approach to the dynamic visibility problem, as described in [10], we begin by laying down a simple Cartesian grid—this is step one in almost all level set methods. Next, we compute the signed distance function $\varphi(\vec{x})$ to the occluding region Ω (which generally has many disjoint pieces). Here we can use the optimally fast algorithm of Tsitsiklis [12]. The function φ approximately satisfies the eikonal equation:

$$\sqrt{\varphi_x^2 + \varphi_y^2} = |\nabla \varphi| = 1 \tag{1a}$$

with

$$\varphi(\vec{x}) < 0 \text{ in } \Omega,$$
$$\varphi(\vec{x}) > 0 \text{ in } \Omega^c,$$
$$\text{and } \varphi(\vec{x}) = 0 \text{ on } \partial\Omega. \tag{1b}$$

Then, for a given vantage point $\vec{x}_0$, we use a simple, easily parallelizable, multiresolution algorithm of optimal complexity to compute the visibility function $\psi_{\vec{x}_0}(\vec{x})$:

$$\psi_{\vec{x}_0}(\vec{x}) \geq 0 \Leftrightarrow \vec{x} \text{ is visible to } \vec{x}_0. \tag{2}$$

Next, we allow $\vec{x}_0$ to move with velocity $d\vec{x}_0/dt$. Along the way, we obtain fairly elegant geometric-based formulae for the motion of the horizon. This is defined to be the set of visible points $\vec{x}$ lying in $\partial\Omega$ for which $\vec{x} - \vec{x}_0$ is orthogonal to $\partial\Omega$, i.e., for which

$$\psi_{\vec{x}_0}(\vec{x}) = 0 = \varphi(\vec{x}) = (\vec{x} - \vec{x}_0) \cdot \nabla \varphi(\vec{x}). \tag{3}$$

We do the same for the motion of points on the cast horizon—that is, points that lie both on the extension of the ray connecting a horizon point $\vec{x}_0$ to $\vec{x}$ and on an occluder.

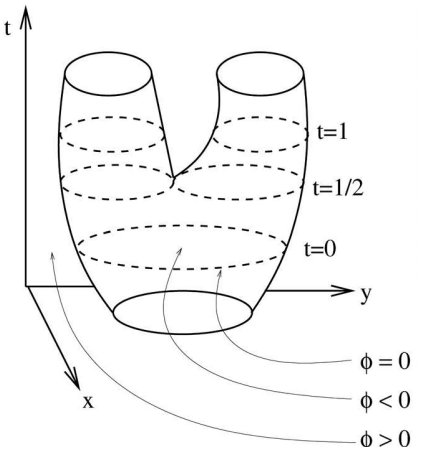
$\psi_{\vec{x}_0(t)}(\vec{x},t)$ can be found quickly at each discrete time step. As $\vec{x}_0(t)$ flies through a region, we can compute the invisible region at time *t* by computing the intersection of the sets

$$S_\tau = \left\{ \vec{x} \middle| \psi_{x_0(\tau)}(\vec{x},\tau) < 0 \right\} \tag{4}$$

for all $0 \le \tau \le t$. This is done by simple discrete Boolean operations.

In the plenary talk, I also gave a brief overview of the LSM and the associated level set technology, and touched on a few other imaging applications: (1) interpolation of unorganized points, curves, and surface patches [14,15]; (2) solution of PDEs on embedded manifolds with a fixed 3D local Cartesian grid [1, 2]; (3) the LSM-induced link [13] between (a) total variation-based image denoising [10], (b) active contours (snakes) [3], and (c) Mumford–Shah image segmentation [5] . All work mentioned was supported by the Office of Naval Research.

Looking to the future, we plan to modify two assumptions made in this work—that rays travel in straight lines and that the occluders are motionless. Another future direction emerged during my talk in Boston, when Arje Nachman of the Air Force Office of Scientific Research asked about visibility for radar signals, which travel around occluded regions and may return to the observer. We intend to use our new phase space-based level set approach to ray tracing [5] (developed with support from AFOSR) to analyze this and related problems.



Topological changes in (x,y) space are computed without "emotional involvement" by using the zero level set of the evolving level set function $\phi(x,y,t)$.

References

[1] M. Bertalmio, L.-T. Cheng, S. Osher, and G. Sapiro, *Variational problems and PDE's on implicit surfaces. The framework and examples in image processing and pattern formation*, J. Comput. Phys., 174 (2001), 759–780.

[2] M. Bertalmio, G. Sapiro, L.-T. Cheng, and S. Osher, *Variational problems and PDE's on implicit surfaces*, Proceedings of th 1st IEEE Workshop on Variational and LSM in Computer Vision, 186–193.

[3] T.F. Chan and L.A. Vese, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), 266–277.

[4] F. Durand, *3D visibility: Analysis, study and applications*, PhD thesis, MIT (1999).

[5] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure. Appl. Math., 42 (1989), 577–685.

[6] S. Osher, L.-T. Cheng, M. Kang, H. Shim, and R. Tsai, *Geometric optics in a phase space based level set and Eulerian framework*, LSS report 01–01, www.levelset.com.

[7] S. Osher and R.P. Fedkiw, *The Level Set Method and Dynamic Implicit Surfaces*, Springer-Verlag, New York, 2002.

[8] S. Osher and L.I. Rudin, *Feature-oriented image enhancement using shock filters*, SIAM J. Num. Anal., 27 (1990), 919–940.

[9] S. Osher and J.A. Sethian, *Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), 12–49.

[10] L.I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal*, Physica D, 60 (1992), 259–268.

[11] R. Tsai, P. Burchard, L.-T. Cheng, S. Osher, and G. Sapiro, *Dynamic visibility in an implicit framework*, UCLA CAM Report 02–06, 2002.

[12] J. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), 1528–1538.

[13] L. Vese and S. Osher, *The level set method links active contours, Mumford–Shah segmentation and total variation restoration*, UCLA CAM report 02–05, 2002.

[14] H.K. Zhao, S. Osher, and R. Fedkiw, *Fast surface reconstruction using the level set method*, in Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, 2001, Vancouver, 194–201.

[15] H.K. Zhao, S. Osher, B. Merriman, and M. Kang, Implicit and nonparametric shape reconstruction from unorganized points using variational level set method, Comput. Vision and Image Understanding, 80 (2000), 295–319.

Stanley Osher is a professor of mathematics at UCLA.

Backward error analysis and stopping criteria for Krylov space methods

M. Arioli
Rutherford Appleton Laboratory,
Chilton, Didcot,
Oxfordshire, OX11 0QX, UK.
m.arioli@rl.ac.uk

Abstract

We combine linear algebra techniques with finite element techniques to obtain a reliable stopping criterion for Krylov method based algorithms. The Conjugate Gradient method has for a long time been successfully used in the solution of the symmetric and positive definite systems obtained from the finite-element approximation of self-adjoint elliptic partial differential equations. Taking into account recent results [5, 8, 9, 11] which make it possible to approximate the energy norm of the error during the conjugate gradient iterative process, in [1] we introduce a stopping criterion based on an energy norm and a dual space norm linked to the continuous problem. Moreover, we show that the use of efficient preconditioners does not require us to change the energy norm used by the stopping criterion.

In [3], we extend the previous results on stopping criteria to the case of non-symmetric positive-definite problems. We show that the residual measured in the norm induced by the symmetric part of the inverse of the system matrix is relevant to measuring convergence in a finite element context. We then provide alternative ways of calculating or estimating this quantity.

1 Introduction

The finite element method approximates the weak form of a coercive elliptic partial differential equation defined within a Hilbert space by a linear system of equations

$$Au = b \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ is positive definite (not necessarily symmetric) and $b \in \mathbb{R}^N$.

It is important to observe that the previous expression of a linear system can be written by a variational formulation with an expression that is formally equal to the weak formulation of a partial differential equation. To achieve this more precise variational framework, we introduce in $\mathbb{R}^N$ a scalar product and, therefore, a norm based on a symmetric and positive definite matrix $H \in \mathbb{R}^{N\times N}$:

$$(x, y)_H = x^T Hy \qquad \text{and} \qquad \|x\|_H = \sqrt{x^T Hx}$$

It is useful to remember that, conversely, each scalar product in $\mathbb{R}^N$ is uniquely identified by a symmetric and positive definite matrix. Then, we denote by $\mathcal{H}$ the space $\mathbb{R}^N$ with such a scalar product. The topological dual space $\mathcal{H}^*$ of $\mathcal{H}$ is then the space $\mathbb{R}^N$ with the scalar product induced by H^{-1}. Let us denote by $a(x, y) = x^T Ay$. Then the variational formulation of (1) is

$$\left\{ \begin{array}{lll} \text{Find} \quad u \in \mathcal{H} & \text{such that} & \\ a(u, v) = L(v) & \forall v \in \mathcal{H} & (L(\cdot) \in \mathcal{H}^*) \end{array} \right. \tag{2}$$

Existence and uniqueness of the solution are guaranteed if the following conditions are satisfied:

$$a(w,v) \leq C_1\|w\|\mathcal{H}\|v\|\mathcal{H} \tag{3a}$$
$$a(v,v) \geq C_2\|v\|\mathcal{H}^2, \tag{3b}$$

where $C_1 \geq 0$ and $C_2 \geq 0$ are independent of N.

The above relations (3) indicate that the generalized eigenvalues of A with respect to H are strictly positive and, then, that A is nonsingular.

Finally, we note that if A is symmetric we can identify H with A.

2 Stopping criteria, backward error, and Krylov methods

The Krylov methods approximate iteratively the solution u of (2) by $u^{(k)}$ computed by minimizing either a norm of the residual $b - As$ (GMRES, MinRES are two examples) or a suitable norm of the error $u - s$ (Conjugate gradient method, SYMMLQ, and special cases of FOM where $H = (A^T + A)/2$ is used as preconditioner, are some examples) on a Krylov space [6, 9, 12]. When using an iterative method, we normally incorporate a stopping criterion based on the a posteriori component-wise or norm-wise backward error theory [2].

Owing to the special structure of the space $\mathcal{H}$, we must modify the usual backward error analysis taking into account the norm in $\mathcal{H}$. The following theorem gives us the theoretical foundation for new stopping criteria taking into account the norm in $\mathcal{H}$.

Theorem 1 *Let $\mathcal{BL}$ be the space of the bilinear forms on $\mathcal{H}$. We have:*

$$\left.\begin{array}{l} \exists b \in \mathcal{BL}(\mathcal{H}), \exists \delta L \in \mathcal{H}^* \text{ such that:} \\ a(\tilde{u},v) + b(\tilde{u},v) = (L+\delta L)(v), \\ \forall v \in \mathcal{H}, \text{ and} \\ \|b(\cdot,\cdot)\|_{\mathcal{BL}(\mathcal{H})} \leq \alpha, \|\delta L\|_{\mathcal{H}^*} \leq \beta \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} \|\rho_{\tilde{u}}\|_{\mathcal{H}^*} \leq \alpha\|\tilde{u}\|_{\mathcal{H}} + \beta \\ \text{where } \rho_{\tilde{u}} \in \mathcal{H}^* \text{ is defined by} \\ \langle \rho_{\tilde{u}}, v\rangle_{\mathcal{H}^*,\mathcal{H}} = a(\tilde{u},v) - L(v), \\ \forall v \in \mathcal{H} \end{array}\right.$$

Proof See [10] for the finite dimensional case and [4] for an extension to Banach spaces. ∎

Therefore, a good candidate stopping criterion could be

$$\text{IF } \|Au^{(k)} - b\|_{H^{-1}} \leq \eta\|b\|_{H^{-1}} \text{ THEN STOP }, \tag{4}$$

with $\eta < 1$ an a priori threshold fixed by the user. We will see how to build realistic stopping criteria using (4) as a reference.

2.1 Conjugate gradient method and symmetric matrices

If we use the conjugate gradient method, we assume that the matrix A is symmetric and positive definite, and that $H = A$. Moreover, owing to the equalities

$$\sqrt{a(x,x)} = \|x\|_H = \|x\|_A = \|Ax\|_{A^{-1}} = \|Hx\|_{H^{-1}}. \tag{5}$$

we have that (4) is equivalent to the following stopping criteria:

$$\text{IF} \quad \|u^{(k)} - u\|_H \leq \eta \|u\|_H \quad \text{THEN STOP} \ . \tag{6}$$

First of all, we need to add, within the conjugate gradient algorithm, some tool for estimating the value $e_A^{(k)} = (u-u^{(k)})^T A(u-u^{(k)}) = r^{(k)T} A^{-1} r^{(k)}$ at each step k. Moreover, we must estimate $b^T A^{-1} b = u^T A u$.

In [1], we presented several techniques that can be used to estimate $e_A^{(k)}$ and $b^T A^{-1} b$. In particular, the value of $e_A^{(k)}$ can be estimated by using the rule presented by Hestenes and Stiefel in their original paper [7]. The Hestenes and Stiefel rule computes a lower bound ξ_k for $e_A^{(k)}$ that is equal to the bound computed by the Gauss rule proposed in [5]. Moreover, Strakoš and Tichý [11] proved that the Hestenes and Stiefel rule is numerically stable when finite precision arithmetic is used.

Under the assumption that $e_A^{(k+d)} << e_A^{(k)}$, where the integer d denotes a suitable delay, the Hestenes and Stiefel estimate ξ_k will be then computed very cheaply using the information computed during the conjugate gradient method.

Moreover, in [1] we proved that, introducing a preconditioner, the energy norm of the preconditioned problem is equal to $e_A^{(k)}$.

Finally, the choice of η will depend on the properties of the problem that we want to solve, and, in the practical cases, η can be frequently much larger than ε, the roundoff unit of the computer finite precision arithmetic. In [1], we suggested that a reasonable choice for η, when (1) is obtained from a finite-element approximation of a partial differential equation, could be:

$$\eta = \left(\max_{T_j \in \mathcal{T}_h} \int_{T_j} 1 d\mathbf{x} \right)^{1/2} \approx h,$$

where $\mathcal{T}_h$ is a given mesh, T_j is a general element and h is the maximum diameter of an element in $\mathcal{T}_h$. In [1], it is also proved that using the previous choice for η within (4), the error between the exact solution of the partial differential equation and the function built using $u^{(k)}$ and the basis functions of the finite elements used to approximate the problem, measured with the continuous norm, is of order $\mathcal{O}(h)$.

2.2 The non-symmetric case

In [3], the non-symmetric case is analysed where the relation (5) is no longer true. Thus, (4) cannot be seen as a straightforward relation among the errors and their measure in the norm of $\mathcal{H}$. However, the following Lemma proves the equivalence between the norm of the symmetric part of A^{-1} and the dual norm.

Lemma 1 *Let conditions* (3) *hold. Then* $\forall r \in \mathcal{H}$

$$\frac{1}{\sqrt{C_1}}\|\mathbf{r}\|A \leq \|\mathbf{r}\|H \leq \frac{1}{\sqrt{C_2}}\|\mathbf{r}\|A$$

and

$$\frac{\sqrt{C_2}}{C_1^2}\|\mathbf{r}\|H^{-1} \leq \|\mathbf{r}\|A^{-1} \leq \frac{1}{\sqrt{C_2}}\|\mathbf{r}\|H^{-1}.$$

(see [3] for the proof)
As for the symmetric case, in [3], estimates of

$$(b - Au^{(k)})^T H^{-1}(b - Au^{(k)}) = (u - u^{(k)})^T H(u - u^{(k)})^T$$

and of

$$(b - Au^{(k)})^T A^{-1}(b - Au^{(k)}) = (u - u^{(k)})^T A(u - u^{(k)})^T$$

are produced and tested for several Krylov based algorithms. In particular, we analyze the GMRES case and the three-term GMRES and FOM introduced by Widlund [12].
The stopping criteria (4) can be then replaced with:

$$\text{IF} \quad \|b - u^{(k)}\|_{A^{-1}} \leq \eta\|u\|_H\sqrt{C_2} \quad \text{THEN STOP} \ . \tag{7}$$

In [3], it is also proved that choosing $\eta = \mathcal{O}(h)$ within (7), the error between the exact solution of a partial differential equation and the function built using $u^{(k)}$ and the basis functions of the finite elements used to approximate the continuous problem, measured with the continuous norm, is of order $\mathcal{O}(h)$.

3 Conclusions

In [1, 3], several examples of partial differential equations and their finite-element approximation are used to test the stopping criteria. The numerical results support the theory. Moreover, they also support the practical feasibility of the use of the stopping criteria introduced for both stopping the iterative processes and giving reliable estimates of the final errors.

References

[1] M. Arioli. *A stopping criterion for the conjugate gradient algorithm in a finite element method framework*, RAL-TR-2002-034, (2002), to appear on Numerische Mathematik.

[2] M. Arioli, I. S. Duff, and D. Ruiz, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144.

[3] M. Arioli, D. Loghin, and A. Wathen. *Stopping criteria for iterations in finite-element methods*, RAL-TR-2003-009, (2003).

[4] M. Arioli, E. Noulard, and A. Russo, *Stopping criteria for iterative methods: Applications to PDE's*, CALCOLO, 38 (2001), pp. 97–112.

[5] G. Golub and G. Meurant, *Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[6] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[7] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[8] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numerical Algorithms, 16 (1997), pp. 77–87.

[9] ——, *Computer Solution of Large Linear Systems*, vol. 28 of Studies in Mathematics and its Application, Elsevier/North-Holland, Amsterdam, The Netherlands, 1999.

[10] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system.*, J. Assoc. Comput. Mach., 14(3) (1967), pp. 543–548.

[11] Z. Strakoš and P. Tichý, *On error estimation by conjugate gradient method and why it works in finite precision computations*, Electronic Transactions on Numerical Analysis, 13 (2002), pp. 56–80.

[12] O. Widlund, *A Lanczos method for a class of nonsymmetric systems of linear equations.* SIAM J. Numer. Anal., 15(4) (1978), pp. 801–812.

Computational Models for Phase Transformation, Metastability, and Microstructure

Pavel Bělík and Mitchell Luskin[1]
School of Mathematics
University of Minnesota
206 Church Street SE
Minneapolis, MN 55455 U.S.A.
belik@math.umn.edu
luskin@math.umn.edu

Abstract

Phase transformation, metastability, and microstructure offer great challenges to the development and analysis of numerical methods. We present some solutions to these problems that we have developed in the context of martensitic structural phase transformations. These crystals are observed to be in metastable states (local minima), sometimes exhibit fine-scale spatial oscillation, and hysteresis is observed as the temperature or boundary forces are varied. We present computational methods and a numerical analysis for this microstructure, and we discuss several multiscale methods and the different metastable states that they compute.

1 Introduction

To model the evolution of metastable states, we have developed a computational model that nucleates the first order phase change since otherwise the crystal would remain stuck in local minima of the energy as the temperature or boundary forces are varied [4]. Our finite element model for the quasi-static evolution of the martensitic phase transformation in a thin film nucleates regions of the high temperature phase during heating and regions of the low temperature phase during cooling. A more detailed discussion of our algorithm and graphical displays of computational results for our model are given in [4].

A review of mathematical and numerical methods for martensitic phase transformation and microstructure is given in [4]. A more recent study of the numerical analysis of microstructure is given in [2]. A more extensive bibliography of papers on the numerical analysis of the martensitic phase transformation and microstructure can be found at http://www.math.umn.edu/~luskin/.

2 A Computational Model for Martensitic Phase Transformation

We have developed a computational model for the quasi-static evolution of the martensitic phase transformation of a single crystal thin film [4]. Our thin film model [3] includes surface energy, as well as sharp phase boundaries with finite energy. The model also

[1]This work was supported in part by NSF DMS-0074043, by AFOSR F49620-98-1-0433, and by the Minnesota Supercomputer Institute.

includes the nucleation of regions of the high temperature phase (austenite) as the film is heated through the transformation temperature and nucleation of regions of the low temperature phase (martensite) as the film is cooled. The nucleation step in our algorithm is needed since the film would otherwise not transform.

For our total-variation surface energy model, the bulk energy for a film of thickness $h > 0$ with reference configuration $\Omega_h \equiv \Omega \times (-h/2,\, h/2)$, where $\Omega \subset \mathbb{R}^2$ is a domain with a Lipschitz continuous boundary $\partial\Omega$, is given by the sum of the surface energy and the elastic energy

$$\kappa \int_{\Omega_h} |D(\nabla u)| + \int_{\Omega_h} \phi(\nabla u, \theta)\, dx \tag{8}$$

where $\int_{\Omega_h} |D(\nabla u)|$ is the total variation of the deformation gradient [3] and κ is a small positive constant.

We have shown in [3] that energy-minimizing deformations u of the bulk energy (8) are asymptotically of the form

$$u(x_1, x_2, x_3) = y(x_1, x_2) + b(x_1, x_2)x_3 + \mathrm{o}(x_3^2) \text{ for } (x_1, x_2) \in \Omega,\ x_3 \in (-h/2,\, h/2),$$

(which is similar to that found for a diffuse interface model [1]) where (y, b) minimizes the thin film energy

$$\mathcal{E}(y, b, \theta) = \kappa \left(\int_\Omega |D(\nabla y|b|b)| + \sqrt{2} \int_{\partial\Omega} |b - b_0| \right) + \int_\Omega \phi(\nabla y|b, \theta)\, dx \tag{9}$$

over all deformations of finite energy such that $y = y_0$ on $\partial\Omega$. The map b describes the deformation of the cross-section relative to the film [1, 3]. We denote by $(\nabla y|b) \in \mathbb{R}^{3\times 3}$ the matrix whose first two columns are given by the columns of ∇y and the last column by b. In the above equation, $\int_\Omega |D(\nabla y|b|b)|$ is the total variation of the vector valued function $(\nabla y|b|b) : \Omega \to \mathbb{R}^{3\times 4}$.

We describe our finite element approximation of (9) by letting the elements of a triangulation τ of Ω be denoted by K and the inter-element edges by e. We denote the internal edges by $e \subset \Omega$ and the boundary edges by $e \subset \partial\Omega$. We define the jump of a function ψ across an internal edge $e \subset \Omega$ shared by two elements $K_1, K_2 \in \tau$ to be

$$[\![\, \psi \,]\!]_e = \psi_{e,K_1} - \psi_{e,K_2}$$

where ψ_{e,K_i} denotes the trace on e of $\psi|_{K_i}$, and we define $\psi|_e$ to be the trace on e for a boundary edge $e \subset \partial\Omega$. Next, we denote by $\mathcal{P}_1(\tau)$ the space of continuous, piecewise linear functions on Ω which are linear on each $K \in \tau$ and by $\mathcal{P}_0(\tau)$ the space of piecewise constant functions on Ω which are constant on each $K \in \tau$. Finally, for deformations $(y, b) \in \mathcal{P}_1(\tau) \times \mathcal{P}_0(\tau)$ and temperature fields $\tilde\theta \in \mathcal{P}_0(\tau)$, the energy (9) is well-defined and we have that

$$\begin{aligned}
&\kappa \left[\int_\Omega |D(\nabla y|b|b)| + \sqrt{2} \int_{\partial\Omega} |b - b_0| \right] + \int_\Omega \phi(\nabla y|b, \tilde\theta)\, dx \\
&\qquad = \kappa \left(\sum_{e\subset\Omega} \Big| [\![\, (\nabla y|b|b) \,]\!]_e \Big|\, |e| + \sqrt{2} \sum_{e\subset\partial\Omega} \Big| b|_e - b_0|_e \Big|\, |e| \right) + \sum_{K\in\tau} \phi\big((\nabla y|b, \tilde\theta)|_K\big)\, |K|,
\end{aligned}$$

where $|\cdot|$ denotes the euclidean vector norm, $|e|$ denotes the length of the edge e, $|K|$ is the area of the element K, and

$$\left|[\![\,(\nabla y|b|b)\,]\!]_e\right| = \left(\left|[\![\,\nabla y\,]\!]_e\right|^2 + 2\left|[\![\,b\,]\!]_e\right|^2\right)^{1/2}.$$

The above term is not differentiable everywhere, so we have regularized it in our numerical simulations.

Since martensitic alloys are known to transform on a fast time scale, we model the transformation of the film from martensite to austenite during heating by assuming that the film reaches an elastic equilibrium on a faster time scale than the evolution of the temperature, so the temperature $\tilde{\theta}(x,t)$ can be obtained from a time-dependent model for thermal evolution [4]. To compute the evolution of the deformation, we partition the time interval $[0,T]$ for $T>0$ by $0=t_0<t_1<\cdots<t_{L-1}<t_L=T$ and then obtain the solution $(y(t_\ell),b(t_\ell))\in\mathcal{A}_\tau$ for $\ell=0,\ldots,L$ by computing a local minimum for the energy $\mathcal{E}(v,c,\theta(t_\ell))$ with respect to the space of approximate admissible deformations

$$\mathcal{A}_\tau = \{(v,c)\in\mathcal{P}_1(\tau)\times\mathcal{P}_0(\tau):\ v=y_0 \text{ on } \partial\Omega\}. \tag{10}$$

Since the martensitic transformation strains $\mathcal{U}\subset\mathbb{R}^{3\times3}$ are local minimizers of the energy density $\phi(F,\theta)$ for all θ near θ_T, a deformation that is in the martensitic phase will continue to be a local minimum for the bulk energy $\mathcal{E}(v,c,\theta(t))$ for $\theta>\theta_T$. Hence, our computational model will not simulate a transforming film if we compute $(y(t_\ell),b(t_\ell))\in\mathcal{A}_\tau$ by using an energy-decreasing algorithm with the initial state for the iteration at t_ℓ given by the deformation at $t_{\ell-1}$, that is, if $(y^{[0]}(t_\ell),b^{[0]}(t_\ell))=(y(t_{\ell-1}),b(t_{\ell-1}))$. We have thus developed and utilized an algorithm to nucleate regions of austenite into $(y(t_{\ell-1}),b(t_{\ell-1}))\in\mathcal{A}_\tau$ to obtain an initial iterate $(y^{[0]}(t_\ell),b^{[0]}(t_\ell))\in\mathcal{A}_\tau$ for the computation of $(y(t_\ell),b(t_\ell))\in\mathcal{A}_\tau$.

We used an "equilibrium distribution" function, $P(\theta)$, to determine the probability for which the crystal will be in the austenitic phase at temperature θ and we assume that an equilibrium distribution has been reached during the time between $t_{\ell-1}$ and t_ℓ. The distribution function $P(\theta)$ has the property that $0<P(\theta)<1$ and

$$P(\theta)\to0 \text{ as } \theta\to-\infty \quad \text{and} \quad P(\theta)\to1 \text{ as } \theta\to\infty.$$

At each time t_ℓ, we first compute a pseudo-random number $\sigma(K,\ell)\in(0,1)$ on every triangle $K\in\tau$, and we then compute $(y^{[0]}(t_\ell),b^{[0]}(t_\ell))\in\mathcal{A}_\tau$ by (x_K denotes the barycenter of K):

1. If $\sigma(K,\ell)\le P\left(\theta(x_K,t_\ell)\right)$ and $(\nabla y(x_K,t_{\ell-1})|b(x_K,t_{\ell-1}),\theta(x_K,t_\ell))$ is in austenite, then set
$$(y^{[0]}(t_\ell),b^{[0]}(t_\ell)) = (y(t_{\ell-1}),b(t_{\ell-1})) \text{ on } K.$$
2. If $\sigma(K,\ell)\le P\left(\theta(x_K,t_\ell)\right)$ and $(\nabla y(x_K,t_{\ell-1})|b(x_K,t_{\ell-1}),\theta(x_K,t_\ell))$ is in martensite, then transform to austenite on K.
3. If $\sigma(K,\ell)> P\left(\theta(x_K,t_\ell)\right)$ and $(\nabla y(x_K,t_{\ell-1})|b(x_K,t_{\ell-1}),\theta(x_K,t_\ell))$ is in austenite, then transform to martensite on K.

4. If $\sigma(K,\ell) > P\left(\theta(x_K,t_\ell)\right)$ and $(\nabla y(x_K,t_{\ell-1})|b(x_K,t_{\ell-1}),\theta(x_K,t_\ell))$ is in martensite, then set

$$(y^{[0]}(t_\ell),b^{[0]}(t_\ell)) = (y(t_{\ell-1}),b(t_{\ell-1})) \text{ on } K.$$

We then compute $(y(t_\ell),b(t_\ell)) \in \mathcal{A}_\tau$ by the Polak-Ribière conjugate gradient method with initial iterate $(y^{[0]}(t_\ell),b^{[0]}(t_\ell)) \in \mathcal{A}_\tau$. We have also experimented with several other versions of the above algorithm for the computation of $b^{[0]}(t_\ell)$. For example, the above algorithm can be modified to utilize different probability functions $P(\theta)$ in elements with increasing and decreasing temperature. We can also prohibit the transformation from austenite to martensite in an element in which the temperature is increasing or prohibit the transformation from martensite to austenite in an element for which the temperature is decreasing.

References

[1] K. Bhattacharya & R. James, A theory of thin films of martensitic materials with applications to microactuators, *J. Mech. Phys. Solids* 47 (1999), 531–576.

[2] P. Bělík & M. Luskin, Stability of microstructure for tetragonal to monoclinic martensitic transformations. *Math. Model. Numer. Anal. 34* (2000), 663–685.

[3] P. Bělík & M. Luskin, A total-variation surface energy model for thin films of martensitic crystals, *Interfaces and Free Boundaries* 4 (2002), 71–88.

[4] P. Bělík & M. Luskin, A computational model for the indentation and phase transformation of a martensitic thin film, *J.Mech. Phys. Solids* 50 (2002) 1789–1815.

[5] M. Luskin, On the computation of crystalline microstructure. *Acta Numerica* 5 (1996), 191–258.

New applications of a posteriori analysis

Christine Bernardi
Laboratoire Jacques-Louis Lions,
C.N.R.S. & Université Pierre et Marie Curie,
B.C. 187, 4 place Jussieu, 75252 Paris Cedex 05, France.
bernardi@ann.jussieu.fr

Abstract

A posteriori analysis has known a huge development in the last twenty years. Its main application is mesh adaptivity since it provides the appropriate tools for constructing a mesh which is perfectly adapted to the solution that must be approximated. However other applications have been recently brought to light and we try to give a brief idea of some of them.

1 A posteriori estimates and mesh adaptivity

We consider an elliptic, linear or nonlinear, partial differential equation set in a two- or three-dimensional bounded domain Ω and its discretization by any method relying on a triangulation or quadrangulation of Ω. Indeed, even if we are more interested in their use in finite elements, the a posteriori estimates can also be employed for a large number of other types of discretizations such as finite differences, finite volumes or spectral elements.

A priori analysis provides a bound of the error between the exact and discrete solutions, for instance in the energy norm, as a function of the discretization parameter (which is usually the maximal diameter of the elements of the triangulation or quadrangulation), the exact solution and sometimes also of the data. This estimate is very important since it proves the convergence of the family of discrete solutions towards the solution of the equation (or a solution of this equation in the non-uniqueness case) when the discretization parameter tends to zero. However it involves the regularity of the exact solution which is most often unknown, so that it cannot provide an explicit order of convergence.

In contrast, a posteriori analysis leads to a bound of the same error as a function of the discretization parameter, the discrete solution and the data or an approximation of it. Note that the maximal diameter of the elements associated with the mesh is not sufficient to characterize it, so that the discretization parameter is now a map, called mesh size field, defined on the computation domain and equal to the diameter of the element on each element of the triangulation or quadrangulation (more complicated metrics must be used in the case of anisotropic meshes but we do not consider this situation here). The key point is that, once the discrete solution is known, the a posteriori bound can be computed explicitly. The first application is of course to enforce security criteria since it allows for maximizing or minimizing the values of the energy norm or a weaker norm or linear forms applied to the exact solution, and this is rather important in a number of physical or engineering problems.

Assume that the bound for the error can be expressed as a Hilbertian sum of local terms, called error indicators, where each indicator involves only the values of the discrete solution and of the data on a "small" subdomain of Ω. Conversely, this indicator can

be bounded from above by the error on this small subdomain or a neighbourhood of it. The two bounds contain constants and they are said to be optimal if both constants are independent of the discretization parameter. In this case, it can be thought that the error indicators provide a good representation of the local error. So they are the appropriate tools for mesh adaptivity: The mesh can be refined in the parts of the computation domain where the indicators are large (more precisely larger than their mean value) and coarsened in the parts when they are smaller. Of course more sophisticated adaptivity strategies can be employed but in any case error indicators allow for increasing the accuracy of the discretization when keeping the same number of degrees of freedom as in the initial mesh. We refer to the book [16] (see also the references therein) for the detailed a posteriori analysis of several elliptic problems and its application to mesh refinement.

2 Application to multi-step discretizations

Most often, in the discretization of a partial differential equation, one or several intermediate non-discrete problems appear between the continuous and discrete ones. In this case, the discretization is called multi-step. This means that several discretization parameters are involved and, in order to optimize the choice of each of them, several families of error indicators must be introduced. We give several examples of such discretizations.

- Parabolic equations

A large part of the research work concerning the a posteriori analysis of parabolic equations deals either only with the space discretization, see [10] for instance, or with the global space-time discretization relying on discontinuous Galerkin methods, see [14], [15] and [17] among others by the same authors. However another approach consists in optimizing separately the choice of the time step and of the mesh size field, which seems to lead to a more natural adaptivity strategy. Indeed, the idea is to uncouple the errors which come from the time and space discretizations as much as possible. The intermediate problem here is the time semi-discrete problem and the two discretization parameters are the time step and the mesh size field. The a posteriori analysis in this framework has been performed first for the linear and a nonlinear heat equations [4], second for the Stokes problem [9]. Its extension to a model for unsteady flows in porous media is under consideration [8].

- Asymptotic expansion for plate and shell problems

The solution of the three-dimensional elasticity system set in a plate or in a shell admits an asymptotic expansion as a function of the thickness of the plate or the shell, and each coefficient of the expansion is the solution of a two-dimensional problem. The key idea for the discretization of the full system consists in applying a finite element method to a small number (one, two or three) of the two-dimensional problems. This fits the framework of multi-step discretizations, since the intermediate problems are the two-dimensional ones and the discretization parameters are the order of truncature of the expansion and the two-dimensional mesh size field. A first a posteriori estimate in this direction is proved in [13].

- Problems set in a three-dimensional axisymmetric domain

The solution of a partial differential equation which is set in a domain invariant by rota-

tion around an axis admits an expansion as a Fourier series of the angular variable, and each Fourier coefficient is the solution of a two-dimensional problem in the meridian domain. These problems can be coupled or not, according to whether the initial equation is linear or not and whether or not its coefficients depend on the angular variable. The main difficulty here is that their variational formulation relies on weighted Sobolev spaces since the Lebesgue measure is transformed into a weighted one by the use of cylindrical coordinates, however the definition and properties of these spaces have been fully investigated in [5]. Here also the intermediate problems are a finite number of the two-dimensional ones and the discretization parameters are the order of the Fourier truncation and the mesh size field. One of the aims is of course to optimize the first one when performing mesh adaptivity. The a posteriori analysis of such a discretization for the Stokes problem is under consideration [3].

- Regularization and penalization methods

In a large number of numerical simulations, a regularization term, sometimes called penalization term, is added to the initial system of partial differential equations in order to stabilize the problem before its simulation. This term involves a small parameter ε, and the choice of ε in the simulation is most often completely arbitrary. The intermediate problem is clearly the stabilized one and the value of the parameter ε must be optimized in the context of mesh adaptivity. We have performed the a posteriori analysis of the penalization method for the Stokes and Navier–Stokes equations [6, 7]. Numerical experiments show that the two families of error indicators uncouple the two types of errors in a perfect way, indeed the first family turns out to be completely independent of the variations of the mesh size field and the second one is completely independent of ε. Moreover the CPU time when working with an adapted value of ε is about the third of the CPU time with an arbitrary small ε. An extension to the much more complex problem of a grade-two fluid model is under consideration [1].

In all these discretizations, the aim is to uncouple as much as possible the errors due to the different steps, which are linked to independent parameters. But each indicator only requires the knowledge of the data and the fully discrete solution, since the main idea is that it can be computed explicitly. So this is not completely possible for the estimates. However the first numerical experiments show that the uncoupling is better than the estimates allow us to hope for.

It can also be observed that the idea of local representation of the error must be given up (or at least modified) for some families of indicators. But in any case the optimization of the other parameters must remain compatible with mesh adaptivity.

3 Application to the coupling of models

In real life situations, the system of partial differential equations that models the phenomenon is set in a three-dimensional bounded or unbounded domain, is highly nonlinear and can involve a large number of unknowns. So most often it is impossible to simulate it directly. Engineers or physicists or mechanists then explain that some quadratic or cubic terms or part of the unknowns can be neglected in the whole domain or at least in some subdomains. These heuristic approaches are usually very efficient for the simulation. Recently it has been observed that a posteriori analysis can provide the first mathematical

and numerical justifications for neglecting these terms and also for optimizing the choice of the subdomains where they can be forgotten. We refer to [11] for a first study of this new approach in an abstract framework with some interesting examples: to [12] for the automatic coupling of Navier–Stokes and Euler equations via the χ-method and to [2] for a first application to the automatic coupling of one-, two- and three-dimensional models in river estuaries. An open question is to decide if this approach can be inserted in the multi-step framework or not. In any case, the number of possible applications seems to be incredibly high.

References

[1] M. Amara, C. Bernardi, V. Girault, F. Hecht "Regularized finite element discretizations of a grade-two fluid model" in preparation.

[2] M. Amara, D. Capatina, D. Trujillo "Hydrodynamical modelling and multidimensional approximation of estuarian river flows" to appear in Computing and Vizualisation in Science.

[3] Z. Belhachmi, C. Bernardi, S. Deparis, F. Hecht "A truncated Fourier/finite element discretization of the Stokes equations in an axisymmetric domain" in preparation.

[4] A. Bergam, C. Bernardi, Z. Mghazli "A posteriori analysis of the finite element discretization of a nonlinear parabolic equation" submitted.

[5] C. Bernardi, M. Dauge, Y. Maday plus M. Azaïez "Spectral Methods for Axisymmetric Domains" Series in Applied Mathematics 3, Gauthier-Villars et North-Holland (1999).

[6] C. Bernardi, V. Girault, F. Hecht "A posteriori analysis of a penalty method and application to the Stokes problem" to appear in Math. Models and Methods in Applied Sciences.

[7] C. Bernardi, V. Girault, F. Hecht "Choix du paramètre de pénalisation pour la discrétisation par éléments finis des équations de Navier–Stokes" C.R. Acad. Sci. Paris 336 série I (2003) 671–676.

[8] C. Bernardi, V. Girault, K.R. Rajagopal "Discretization of an unsteady flow through a porous solid modeled by Darcy's equations" in preparation.

[9] C. Bernardi, R. Verfürth "A posteriori error analysis of the fully discretized time-dependent Stokes equations" submitted.

[10] M. Bieterman, I. Babuška "The finite element method for parabolic equations. I. A posteriori error estimation" Numer Math. 40 (1982) 339–371.

[11] M. Braack, A. Ern "A posteriori control of modeling errors and discretization errors" SIAM J. Multiscale Modeling and Simulation 1 (2003) 221–238.

[12] F. Brezzi, C. Canuto, A. Russo "A self-adaptive formulation for the Euler/ Navier–Stokes coupling" Comput. Methods Appl. Mech. Engrg. 73 (1989) 317–330.

[13] M. Dauge, E. Faou, in preparation.

[14] K. Eriksson, C. Johnson "Adaptive finite element methods for parabolic problems. I. A linear model problem" SIAM J. Numer. Anal. 28 (1991) 43–77.

[15] K. Eriksson, C. Johnson "Adaptive finite element methods for parabolic problems. IV. Nonlinear problems" SIAM J. Numer. Anal. 32 (1995) 1729–1749.

[16] R. Verfürth "A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques" Wiley and Teubner Mathematics (1996).

[17] R. Verfürth "A posteriori error estimation techniques for non-linear elliptic and parabolic pdes" Revue européenne des éléments finis 9 (2000) 377–402.

A View on Spline Prewavelets

M. D. Buhmann
Mathematical Institute,
Justus Liebig University,
35392 Giessen, Germany,
martin.buhmann@math.uni-giessen.de

Abstract

In this paper, we briefly summarize some results on spline prewavelets. Prewavelets and wavelets expansions, as well as fast algorithms for their computation, are currently at the core of modern methods for numerical analysis due to their usefulness in applications for numerical PDE solvers, also for signal processing, (stiffness) matrix and data compression and transmission, and other uses of computational harmonic analysis.

Overview

The talk at the Dundee conference 2003 presented a summary of some results on prewavelets that were designed on the basis of spline approximation spaces. A number of other contributions at that conference showed applications such as in the numerical computation of elliptic partial differential equations, where discrete wavelets (Daubechies and others) were employed for stiffness matrix compression. The uses of such methods are far-reaching within scientific and engineering applications. Seen from the theoretical side and more abstractly, wavelets and prewavelets are appealing due to their nature linked to uni- and multivariate Fourier expansions, numerical harmonic analysis, approximation and spline theory, and due to their versatility in practice ([11, 16, 15], for example).

Their applications in science and engineering are manifold and they range from transmitting pictures over the internet and sound over the telephone or filtering sound signals, *e.g.* from underwater sonar-scans or on compact discs, to the numerical solution of elliptic and parabolic partial differential equations, especially when Petrov-Galerkin or pure Galerkin approaches are used [11, 16]. As is well-known to practitioners, typical two-dimensional splines, described as finite elements in the language of PDE solvers, are important here with particular properties such as piecewise polynomial (pp) structure and small, finite support. In [7] we propose spline prewavelets of small support that can be good for such practical use. Without small support there is no sufficient sparsity in the stiffness or mass matrices of the PDE problems which is a problem in most numerical computations of large linear systems. In fact, most efficient algorithms such as preconditioned conjugate gradients for the solution of the stiffness equations rely on substantial sparsity and good conditioning of the underlying matrix.

There are many uni-variate constructions of prewavelets and wavelets. Also, in the context of splines and radial basis functions, there are suitable constructions, of which we mention [8] and [3, 4, 5] amongst very many others.

We note from the current standard literature that continuously differentiable spline-based prewavelets on $\mathbb{R}^n$ for two or more dimensions are available only as tensor product or box spline constructions. As pointed out in the contribution at Dundee, the box spline

constructions are ruled out since any stable constructions (unless we let the pp degree be larger) contain a spline space without all polynomials of corresponding total degree whereas criss-cross partition box splines are instable in this use (see, *e.g.*, [13], [10]). However, our construction in [7] is with stable pp cubic bases which generate all cubic polynomials. They give prewavelets with so-called vanishing moments [14] of total degree up to three. We recall that the criss-cross partition of $\mathbb{R}^2$ is generated by the four families of parallel lines $x_1 = k$, $x_2 = k$, $x_1 - x_2 = k$, $x_1 + x_2 = k$, $k \in \mathbb{Z}$ [9], while the three-direction mesh comes from leaving out the $x_1 + x_2 = k$ lines.

As always, we have to use for the construction the standard definition of a multiresolution analysis (MRA) which is fundamental to the development of any wavelets or prewavelets. This is an infinite nest of closed subspaces $V_j \subset L^2(\mathbb{R}^n)$, $j \in \mathbb{Z}$,

$$\{0\} \subset \cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset L^2(\mathbb{R}^n)$$

that satisfy the following properties:

(i) $f \in V_j$ if and only if $f(2\cdot) \in V_{j+1}$ for all integers j,

(ii) the union of all V_js is dense in $L^2(\mathbb{R}^n)$ and their intersection contains only O (but see [1] for conditions under which (ii) is redundant),

(iii) there is a Riesz basis $\{\phi_i : \ i \in I\}$ of V_0, *i.e.*,

$$V_0 = \mathrm{span}_{\ell^2(I)}\{\phi_i : \ i \in I\},$$

where I is a countable index set, the coefficients of the spanning functions are always square-summable, for all $c \in \ell^2(I)$

$$\|c\|_2 \sim \Big\|\sum_{i\in I} c_i\phi_i\Big\|_2.$$

The above Riesz basis (*i.e.*, basis with Riesz stability estimates) property is of particular importance with respect to the stability of the computations of the coefficients of an expansion. Unless this property is provided, instabilities can occur through cancellations of coefficients of a function's expansion in the infinite basis, which would be at the expense of numerically undesirable or indeed useless results.

We observe that the properties of MRA have many fundamental consequences. One of them is that we can find a collection of square-integrable functions named *prewavelets* in $V_1 \setminus \{0\}$, which are orthogonal to V_0, call the set of prewavelets Ψ, whose translates span a space $W_0 = V_1 \ominus V_0$. In other words, V_1 is the direct and orthogonal sum of V_0 and the space W_0, for which $W_0 = V_1 \ominus V_0$ is a short and usuful notation for

$$V_1 = V_0 \oplus W_0, \qquad V_0 \perp W_0.$$

In principle, all functions of W_0, and indeed the aforementioned spanning set, can be found by computing the error of an orthogonal (with respect to the Euclidean norm) projection of all elements from V_1 onto V_0, *i.e.*, the element of V_1 minus its projection onto V_0. The derivation relies on finding a suitable set of generating functions (in the

event, they are differentiable cubic splines of small support) in V_1 whose projection is then computed to form the prewavelets.

As a consequence of the properties of multiresolution analysis and the properties of W_0, we get the infinite decomposition of square-integrable functions

$$L^2(\mathbb{R}^n) = \bigoplus_{j=-\infty}^{\infty} W_j.$$

Here W_j denotes W_0 with the functions scaled by 2^j, see for instance [16, 1]. The above decomposition of $L^2(\mathbb{R}^2)$ then admits the decompositions of signals of finite energy into their components at different frequencies, reflected by the different scaling factors in the definition of W_j, and at different places or at different time.

In summary, we have the desired decomposition of the whole of $L^2(\mathbb{R}^n)$, because the W_j are mutually orthogonal which follows from a standard argument using the fact that the prewavelets ψ are orthogonal to V_0 and from (i). The orthogonality of the different W_j corresponds to the idea that they represent different frequencies just as the exponential functions which are orthogonal and represent different frequencies in uni- and multivariate Fourier expansions. The prewavelets are called wavelets if their translates on the same scale are also mutually orthonormal.

We also recall the use of the piecewise polynomial (pp) box-splines that are familiar from the book [2] for instance. We can provide their formal definition either in recurrence relation form, viz.

$$B(x) = \int_0^1 \tilde{B}(x - t\xi)\, dt, \qquad x \in \mathbb{R}^n,$$

where B is the box-spline that results from adding the direction ξ to the direction set of $\tilde{B}$, the initial set of directions to launch the recursion with the piecewise constant box-spline $\bar{B}$ being a set X of n linear independent directions from $\mathbb{Z}^n$, and $\bar{B}$ being the indicator function of $X[0,1]^n$ scaled by $|\det X|^{-1}$. Alternatively, we may define box-splines as the inverse Fourier transforms of certain simple entire functions, namely let $v_1, \ldots, v_\ell$ be different vectors in $\{-1,0,1\}^n = \mathbb{Z}^n \cap [-1,1]^n$ which span $\mathbb{R}^n$ (linear independence is not required and is usually, in fact, not provided). The *box-spline* B associated with these vectors called directions with multiplicities $n_1, \ldots, n_\ell \geq 1$, respectively, may be defined by its Fourier transform

$$\hat{B}(\omega) := \left(\frac{1 - z^{\xi_1}}{i\xi_1 \cdot \omega}\right)^{n_1} \cdots \left(\frac{1 - z^{\xi_\ell}}{i\xi_\ell \cdot \omega}\right)^{n_\ell}, \qquad \omega = (\omega_1, \ldots, \omega_n)^T \in \mathbb{R}^n,$$

where $z = (z_1, \ldots, z_n)^T = (e^{-i\omega_1}, \ldots, e^{-i\omega_n})^T$, $z^{\xi_k} := z_1^{\xi_{1,k}} \cdots z_n^{\xi_{d,k}}$, and $\xi_k \cdot \omega$ denotes the scalar product of the two vectors. From those box-splines, [6] construct small support prewavelets in one, two and three dimensions.

In order to compute useful prewavelets, the goal in [7] is an explicit construction of prewavelets spanned by a finite linear combination of splines from V_0 and V_1 with explicit coefficients. Incidentally, not all pp spaces of two dimensions are able to provide multiresolution analyses and are indeed refinable. However, the construction in [7] provides this property, as in, for example, the construction of splines in two dimensions with the famous Powell-Sabin split [18].

The finding of an adjoint (biorthogonal) basis is very relevant to our construction because it facilitates the computation of the aforementioned projections (see a similar approach to computing continuous prewavelets in [19, 14].

References

[1] de Boor, C., DeVore, R. A., Ron, A. On the construction of multivariate (pre) wavelets; Constr. Approx.; 9; 1993; 123–166;

[2] de Boor, C., Höllig, K., Riemenschneider, S. Box Splines; Springer (New York); 1993;

[3] Buhmann, M.D. Radial Basis Functions; Cambridge University Press (Cambridge); 2003;

[4] Buhmann, M. D. Discrete least squares approximation and pre-wavelets from radial function spaces; Mathematical Proceedings of the Cambridge Philosophical Society; 114; 1993; 533–558;

[5] Buhmann, M. D. Multiquadric pre-wavelets on non-equally spaced knots in one dimension; Mathematics of Computation; 64; 1995; 1611–1625;

[6] Buhmann, M. D., Davydov, O., Goodman, T. N. T. Box-spline prewavelets with compact support; J. Approx. Theory; 112; 2001; 16–27;

[7] Buhmann, M. D., Davydov, O., Goodman, T. N. T. Cubic spline prewavelets on the four direction mesh; Foundations of Computational Mathematics; 3; 2003; 113–133;

[8] Buhmann, M. D., Micchelli, C. A. Spline prewavelets on non-uniform knots; Numer. Math.; 61; 1992; 455–474;

[9] Chui, C. K. Multivariate Splines; CBMS-NSF Reg. Conf. Series in Appl. Math., vol. 54, SIAM (Philadelphia); 1988;

[10] Chui, C. K., Jetter, K., Stöckler, J. Wavelets and frames on the four-directional mesh, in *Wavelets: Theory, Algorithms, And Applications*, C. Chui, L. Montefusco and L. Puccio (eds), Academic Press, New York, 1994, 213–230;

[11] Dahmen, W. Wavelet and multiscale methods for operator equations; Acta Numerica; 6; 1997; 55–228;

[12] Dahmen, W., Han, B., Jia, R.-Q., Kunoth, A.; Biorthogonal multiwavelets on the interval: Cubic Hermite spines; Constr. Approx.; 16; 2000; 221–259;

[13] Dahmen, W., Micchelli, C. A. Translates of multivariate splines; Linear Algebra Appl.; 52; 1983; 217–234;

[14] Dahmen, W., Stevenson, R. Element-by-element construction of wavelets satisfying stability and moment conditions; SIAM J. Numer. Anal.; 37; 1999; 319–352;

[15] Daubechies, I. Ten Lectures on Wavelets; SIAM (Philadelphia); 1992;

[16] DeVore, R. A., Lucier, B. Wavelets; Acta Numerica; 1; 1992; 1–56;

[17] Goodman, T. N. T. Properties of bivariate refinable spline pairs; in Multivariate Approximation, Recent Trends and Results, W. Haussmann, K. Jetter, M. Reimer (eds), Akademie Verlag (Berlin); 1997, 63–82;

[18] Powell, M. J. D., Sabin, M. A. Piecewise quadratic approximations on triangles; ACM Trans. Math. Software; 3; 1977; 316–325;

[19] Stevenson, R. Piecewise linear (pre-)wavelets on non-uniform meshes; in *Multigrid Methods V*, W. Hackbusch and G. Wittum (eds), Springer, Berlin, 1998; 306–319;

Reduced order modeling of complex systems

John Burkardt[†], Qiang Du[‡], Max Gunzburger[†] & Hyung-Chun Lee[⋆1]

[†]*School for Computational Science and Information Technology,*
Florida State University,
Tallahassee FL 32306-4120, USA
burkardt@csit.fsu.edu

[‡]*Department of Mathematics,*
Penn State University,
University Park PA 16802, USA
qdu@math.psu.edu

gunzburg@csit.fsu.edu
&
[⋆]*Department of Mathematics,*
Ajou University,
Suwon 442-749, Korea
hclee@ajou.ac.kr

1 Reduced-order modeling

Solutions of (nonlinear) complex systems are expensive with respect to both storage and CPU costs. As a result, it is difficult if not impossible to deal with a number of situations such as: continuation or homotopy methods for computing state solutions; parametric studies of state solutions; optimization and control problems (multiple state solutions); and feedback control settings (real-time state solutions). Not surprisingly, a lot of attention has been paid to *reducing the costs of the nonlinear state solutions by using reduced-order models for the state*; these are *low-dimensional* approximations to the state. Reduced-order modeling has been and remains a very active research direction in many seemingly disparate fields. We will focus on three approaches to reduced-order modeling: reduced basis methods; proper orthogonal decomposition (POD); andcentroidal Voronoi tessellations (CVT). Before describing the three approaches, we first discuss what we exactly mean by reduced-ordering modeling and make some general comments that apply to all reduced-order models. For a *state simulation*, a reduced-order method would proceed as follows. One first chooses a reduced basis $\mathbf{u}_i$, $i = 1, \ldots, n$, where n is hopefully very small compared to the usual number of functions used in a finite element approximation or the number of grid points used in a finite difference approximation. Next, one seeks an approximation $\widetilde{\mathbf{u}}$ to the state of the form $\widetilde{\mathbf{u}} = \sum_{i=1}^{n} c_i \mathbf{u}_i \in V \equiv \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Then, one determines the coefficients c_i, $i = 1, \ldots, n$, by solving the state equations in the set V, e.g., one could find a Galerkin solution of the state equations in a standard way, using V for the space of approximations. The cost of such a computation would be very small if n is small (ignoring the cost of the off-line determination of the reduced basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$). In *control or optimization settings*, one is faced with multiple state solves

[1]Supported by KOSEF R01-2000-000-00008-0.

or real-time state solves. If one approximates the state in the reduced, low-dimensional set V, then state solutions will be relatively very cheap. In an adjoint or sensitivity equation-based optimization method, one could also employ the adjoint equations for the low-dimensional discrete state equations; thus, if n is small, the cost of each iteration of the optimizer would be very small relative to that using full, high-fidelity state solutions. In a feedback control setting, the approximate state equations in the low-dimensional space could possibly be solved in real time. Does reduced-order modeling work? It is clear that reduced-order methods should work in an *interpolatory setting*. In a simulation setting, if the state can be approximated well in the reduced basis V, then one should expect that things will work well. If the optimal solution and the path to the optimal solution can be well approximated in the reduced basis V, then one should expect that things will work well in an optimal control or design setting. If all the states determined by the feedback process can be well approximated in the reduced basis V, then again one should expect that things will work well in a feedback control setting.Thus, the reduced basis V should be chosen so that it *contains all the features, e.g., the dynamics, of the states encountered during the simulation or the control process*. This, of course, requires some intuition about the states to be simulated or about where in parameter space the optimal set of parameters are located.

What happens in an *extrapolatory setting* is not so clear. Most reduced-order control computations have been done in an interpolatory regime. It is obvious that if the reduced set V does not contain a good approximation to the solution one is trying to obtain, then one cannot hope to successfully determine that solution.

1.1 Common features shared by reduced-order methods

All reduced bases *require the solution of high-fidelity and therefore very expensive discrete state and/or sensitivity equations and/or adjoint equations.* The idea is that these expensive calculations can be done offline before a state simulation or the optimization of the design parameters or feedback control is attempted. Moreover, one hopes that a single reduced basis can be used for several state simulations or in several design or control settings.

All reduced-basis sets are *global in nature*, i.e., the support of the basis functions is global. Therefore, solving the state or sensitivity or adjoint equations with respect to any of the reduced bases requires the solution of dense linear and nonlinear systems. Thus, unless the dimension of a reduced basis is "small," it cannot be used without some further processing. Unfortunately, in order to obtain meaningful approximations, it is often the case that the use of reduced bases requires the use of a relatively large number of basis functions. However, it is often the case that reduced bases contain "redundant" information in the sense that the dynamics of the state should be well approximated by a set of functions of much lower dimension. The question then arises: how can one extract a reduced basis of smaller dimension that contains all the essential information of a reduced basis of larger dimension? This is where POD and CVT come in and, in this sense, they are *reduced-reduced* basis methods.

Unfortunately, there is no adequate theoretical foundation for reduced-order methods, even in state simulation settings. However, it is certain that without an inexpensive method for reducing the cost of state computations, it is unlikely that the solution of

three-dimensional optimization and control problems involving complex systems, e.g., the Navier-Stokes system, will become routine anytime soon. Thus, it is also certainly true that these methods deserve more study from the computational and theoretical points of view.

2 Reduced-basis methods

All reduced-order methods are reduced basis methods. However, there is a class of methods that use Lagrange bases, Hermite bases, Taylor bases, and snapshot bases (or more precisely, snapshot sets) that have come to be know as *reduced-basis methods.*

Lagrange bases consist of state solutions corresponding to several different values of the parameters (Reynolds number, design parameters, etc.) These solutions are obtained by standard (and expensive) techniques such as finite element or finite volume methods. For example, if one has the design parameters $\{\alpha_j\}_{j=1}^J$, one obtains n approximate state solutions for n sets of parameter values to form the n-dimensional Lagrange reduced basis. *Hermite bases* consist of the state variables and the first derivatives of the state variables with respect to parameters (the sensitivities) determined for different values of the parameters. The state and sensitivity approximations are obtained through standard (and expensive) techniques such as finite element or finite volume methods. Thus, again, if one has the design parameters $\{\alpha_j\}_{j=1}^J$, one chooses M sets of parameter values and then one obtains the corresponding M approximate state solutions and the corresponding MJ sensitivity derivative approximations. The $n = M(J + 1)$ state and sensitivity approximations form the Hermite reduced basis of dimension n. *Taylor bases* consist of the state and derivatives of the state with respect to parameters (sensitivities and higher-order sensitivities) determined for a fixed set of design parameters. The state and derivative approximations are obtained through standard (and expensive) techniques such as finite element or finite volume methods. The Taylor basis may be somewhat complicated to program due to the complexity of the partial differential equations that determine the higher-order sensitivities. In addition, the number of higher-order derivatives grows very rapidly with the number of design parameters, e.g., if one has 10 design parameters, there are 55 different second derivative sensitivities. Thus, the dimension of the Taylor reduced basis grows quickly with the number of parameters and the number of derivatives used.

See [13, 20, 21] for more details and for examples of the use of reduced-basis methods for simulation and optimization problems.

2.1 Snapshot sets

The state of a complex system is determined by parameters that appear in the specification of a mathematical model for the system. Of course, the state of a complex system also depends on the independent variables appearing in the model. *Snapshot sets* consist of state solutions corresponding to several parameter values and/or evaluated at several values of one or more of the dependent variables, e.g., steady-state solutions corresponding to several sets of design parameters or a time-dependent state solution for a fixed set of design parameter values evaluated at several time instants during the evolution process or several state solutions corresponding to different sets of parameter values evaluated at

several time instants during the evolution process. Snapshot sets are often determined by solving the full, very large-dimensional discretized system obtained by, e.g., a finite volume or finite element discretization. Experimental data have also been used to determine a snapshot set. Snapshot sets often contain "redundant" information; therefore, snapshot sets must usually be post-processed to remove as much of the redundancy as possible before they can be used for reduced-order modeling. POD and CVT may be viewed as simply different ways to post-process snapshot sets.

Since snapshot sets are the underpinning for POD and CVT, we briefly discuss how they are generated in practice. At this time, the generation of snapshot sets is an art and not a science; in fact, it is a rather primitive art. The generation of snapshot sets is an exercise in the design of experiments, e.g., for stationary systems, how does one choose the sets of parameters at which the state (and sensitivities) are to be calculated (using expensive, high-fidelity computations) in order to generate the snapshot set? Clearly, some a priori knowledge about the types of states to be simulated or optimized using the reduced-order model is very useful in this regard. The large body of statistics literature on the design of experiments has not been used in a systematic manner

For time-dependent systems, many (ad hoc) measures have been invoked in the hope that they will lead to good snapshot sets. Time-dependent parameters (e.g., in boundary conditions) are used to generate states that are "rich" in transients, even if the state of interest depends only on time-independent parameters. In order to generate even "richer" dynamics, impulsive forcing is commonly used, e.g., starting the evolution impulsively with different strength impulses and introducing impulses in the middle of a simulation. In the future, a great deal of effort needs to be directed towards developing and justifying methodologies for generating good snapshot sets. After all, a POD or CVT basis is only as good as the snapshot set used to generate it.

3 Proper orthogonal decompositions (POD)

Given n snapshots $\widetilde{\mathbf{x}}_j \in \mathbb{R}^N$,$j = 1, \ldots, n$, set

$$\mathbf{x}_j = \widetilde{\mathbf{x}}_j - \widetilde{\boldsymbol{\mu}}\,, \qquad j = 1 \ldots, n, \qquad \text{where} \qquad \widetilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^{n} \widetilde{\mathbf{x}}_j.$$

The set $\{\mathbf{x}_j\}_{j=1}^n$ are the *modified snapshots*. Let $d \le n$. Then, the *POD basis* $\{\boldsymbol{\phi}_i\}_{i=1}^d$ of cardinality d is found by successively solving, for $i = 1, \ldots, d$, the problem

$$\lambda_i = \max_{|\boldsymbol{\phi}_i|=1} \frac{1}{n}\sum_{j=1}^{n} |\boldsymbol{\phi}_i^T \mathbf{x}_j|^2 \qquad \text{and} \qquad \boldsymbol{\phi}_i^T \boldsymbol{\phi}_\ell = 0 \quad \text{for } \ell \le i-1.$$

If $n \ge N$, this decomposition is known as the *direct* method. If $n < N$, it is known as the *snapshot method*; we will only consider the latter case.

Let A denote the $N \times n$ snapshot matrix whose columns are the modified snapshots $\mathbf{x}_j$, i.e.,

$$A = \big(\mathbf{x}_1\ ,\ \mathbf{x}_2\ ,\ \ldots\ ,\ \mathbf{x}_n\big) = \big(\widetilde{\mathbf{x}}_1 - \widetilde{\boldsymbol{\mu}}\ ,\ \widetilde{\mathbf{x}}_2 - \widetilde{\boldsymbol{\mu}}\ ,\ \ldots\ ,\ \widetilde{\mathbf{x}}_n - \widetilde{\boldsymbol{\mu}}\big).$$

Let K denote the the $n \times n$ (normalized) correlation matrix for the modified snapshots, i.e.,

$$K_{j\ell} = \frac{1}{n}\mathbf{x}_j^T \mathbf{x}_\ell \qquad \text{or} \qquad K = \frac{1}{n}A^T A.$$

Let $\boldsymbol{\chi}_i$ with $|\boldsymbol{\chi}_i| = 1$ denote the eigenvector corresponding to the i-th largest eigenvalue λ_i of K. Then, the POD basis is given by $\boldsymbol{\phi}_i = \frac{1}{\sqrt{n\lambda_i}} A\boldsymbol{\chi}_i$. The POD basis is orthonormal, i.e., $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j = 0 \quad \text{for } i \neq j$ and $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1$. POD is closely related to the statistical methods known as Karhunen-Loève analysis or the method of empirical orthogonal eigenfunctions or principal component analysis. POD is also closely related to the singular value decomposition (SVD) of the modified snapshot matrix A. Let $A = U\Sigma V^T$ denote the SVD of A; then, $\sigma_i^2 = n\lambda_i$ for $i = 1, \ldots, n$, where $\sigma_i =$ the i-th singular value of A and $\lambda_i =$ the i-th largest eigenvalue of$K = \frac{1}{n}A^T A$. The POD basis vectors are the first n left singular vectors of the snapshot matrix A, i.e., $\boldsymbol{\phi}_i = \mathbf{u}_i$ for $i = 1, \ldots, n$.

The POD basis is optimal in the following sense. Let $\{\boldsymbol{\psi}_i\}_{i=1}^n$ denote an arbitrary orthonormal basis for the span of the modified snapshot set $\{\mathbf{x}_j\}_{j=1}^n$. Let $P_{\psi,d}\mathbf{x}_j$ denote the projection of the modified snapshot $\mathbf{x}_j$ onto the d-dimensional subspace spanned by $\{\boldsymbol{\psi}_i\}_{i=1}^d$. Clearly we have, for each $j = 1, \ldots, n$,

$$P_{\psi,d}\mathbf{x}_j = \sum_{i=1}^{d} c_{ji}\boldsymbol{\psi}_i \qquad \text{where} \qquad c_{ji} = \boldsymbol{\psi}_i^T \mathbf{x}_j \ \text{ for } i = 1, \ldots, d.$$

Let the error be defined by

$$\mathcal{E} = \sum_{j=1}^{n} |\mathbf{x}_j - P_{\psi,d}\mathbf{x}_j|^2.$$

Then, the minimum error is obtained when $\boldsymbol{\psi}_i = \boldsymbol{\phi}_i$ for $i = 1, \ldots, d$, i.e., when the $\boldsymbol{\psi}_i$'s are the POD basis vectors. The connection between POD and SVD makes it easy to show that the error of the d-dimensional POD subspace is given by

$$\mathcal{E}_{\text{pod}} = \sum_{j=d+1}^{n} \sigma_j^2 = n \sum_{j=d+1}^{n} \lambda_j$$

$n =$ number of snapshots
$d =$ dimension of the POD subspace.

If one wishes for the relative error to be less than a prescribed tolerance δ, i.e., if one wants $\mathcal{E}_{\text{pod}} \leq \delta \sum_{j=1}^n |\mathbf{x}_j|^2$, one should

choose d to be the smallest integer such that $\displaystyle\sum_{j=1}^{d} \sigma_j^2 \Big/ \sum_{j=1}^{n} \sigma_j^2 = \sum_{j=1}^{d} \lambda_j \Big/ \sum_{j=1}^{n} \lambda_j \ \geq \ \gamma = 1 - \delta.$

There have also been several variations introduced in attempts to "improve" POD. For details on POD and its variants and its use in flow simulation and control problems, see, e.g., [1, 2, 3, 5, 6, 10, 11, 12, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28].

4 Centroidal Voronoi tessellations (CVT)

Given a *discrete* set of modified snapshots $W = \{\mathbf{x}_j\}_{j=1}^n$ belonging to $\mathbb{R}^N$, a set $\{V_i\}_{i=1}^k$ is a *tessellation* of W if $\{V_i\}_{i=1}^k$ is a subdivision of W into disjoint, covering subsets, i.e.,

$V_i \subset W$ for $i = 1, \ldots, k$, $V_i \cap V_j = \emptyset$ for $i \neq j$, and $\cup_{i=1}^{k} V_i = W$. Given a set of points $\{\mathbf{z}_i\}_{i=1}^{k}$belonging to $\mathbb{R}^N$ (but not necessarily to W) the *Voronoi region* corresponding to the point $\mathbf{z}_i$ is defined by

$$\widehat{V}_i = \{\, \mathbf{x} \in W \quad : \quad |\mathbf{x} - \mathbf{z}_i| \leq |\mathbf{x} - \mathbf{z}_j| \quad \text{for } j = 1, \ldots, k, \ j \neq i, \},$$

where equality holds only for $i < j$. The set $\{\widehat{V}_i\}_{i=1}^{k}$ is called a *Voronoi tessellation* or *Voronoi diagram* of W corresponding to the set of points $\{\mathbf{z}_i\}_{i=1}^{k}$. The points in the set $\{\mathbf{z}_i\}_{i=1}^{k}$ are called the *generators* of the Voronoi diagram $\{\widehat{V}_i\}_{i=1}^{k}$ of$W$. Given a density function $\rho(\mathbf{y}) \geq 0$, defined for $\mathbf{y} \in W$, the *mass centroid* $\mathbf{z}^*$ of any subset $V \subset W$ is defined by

$$\sum_{\mathbf{y} \in V} \rho(\mathbf{y}) |\mathbf{y} - \mathbf{z}^*|^2 = \inf_{\mathbf{z} \in V^*} \sum_{\mathbf{y} \in V} \rho(\mathbf{y}) |\mathbf{y} - \mathbf{z}|^2,$$

where the sums extend over the points belonging to V. The set V^* can be taken to be V or it can be an even larger set such as all of $\mathbb{R}^N$. In case $V^* = \mathbb{R}^N$, $\mathbf{z}^*$ is the ordinary mean

$$\mathbf{z}^* = \sum_{\mathbf{y} \in V} \rho(\mathbf{y}) \mathbf{y} \Big/ \sum_{\mathbf{y} \in V} \rho(\mathbf{y}).$$

In this case, $\mathbf{z}^* \notin W$ in general.
If $\mathbf{z}_i = \mathbf{z}_i^*$ for $i = 1, \ldots, k$, where

$\{\mathbf{z}_i\}_{i=1}^{k}$ is the set of generating points of the Voronoi tessellation $\{\widehat{V}_i\}_{i=1}^{k}$

$\{\mathbf{z}_i^*\}_{i=1}^{k}$ is the set of mass centroids of the Voronoi regions $\{\widehat{V}_i\}_{i=1}^{k}$,

we refer to the Voronoi tessellation as being a *Centroidal Voronoi tessellation* (CVT). The concept of CVT's can be extended to more general sets, including regions in Euclidean space, and to more general metrics. CVT's are useful in a variety of applications, including optimal quadrature rules, covolume and finite difference methods for PDE's, optimal representation, quantization, and clustering, cell division, data compression, optimal distribution of resources, territorial behavior of animals, optimal placement of sensors and actuators, grid generation in 2D, 3D, and on surfaces, mesh free methods, clustering of gene expression data, image segmentation. See, e.g., [7] for details.
CVT's are optimal in the following sense. Given the discrete set of points $W = \{\mathbf{x}_j\}_{j=1}^{n}$ belonging to $\mathbb{R}^N$, we define the error of a tessellation $\{V_i\}_{i=1}^{k}$ of W and a set of points$\{\mathbf{z}_i\}_{i=1}^{k}$ belonging to $\mathbb{R}^N$ by

$$\mathcal{F}\big((\mathbf{z}_i, V_i), i = 1, \ldots, k\big) = \sum_{i=1}^{k} \sum_{\mathbf{y} \in V_i} \rho(\mathbf{y}) |\mathbf{y} - \mathbf{z}_i|^2.$$

Then, it can be shown that a necessary condition for the error $\mathcal{F}$to be minimized is that the pair $\{\mathbf{z}_i, V_i\}_{i=1}^{k}$ form a CVT of W.
CVT's of discrete sets are closely related to optimal k-means clusters so that Voronoi regions and centroids can be referred to as clusters and cluster centers, respectively. The error $\mathcal{F}$ is also often referred to as the variance, cost, distortion error, or mean square error.

There are several algorithms known for constructing Centroidal Voronoi tessellations of a given set. Lloyd's method is a deterministic algorithm which is the obvious iteration between computing Voronoi diagrams and mass centroids, i.e., a given set of generators is replaced in an iterative process by the mass centroids of the Voronoi regions corresponding to those generators. MacQueen's method is a very elegant probabilistic algorithm which divides sampling points into k sets or clusters by taking means of clusters. We have developed a new probabilistic method which may be viewed as a generalization of both the MacQueen and Lloyd methods and is amenable to efficient parallelization. See [14] for details.

4.1 CVT's and model reduction

CVT's have been successfully used in data compression; one particular application was to image reconstruction. Therefore, it is natural to examine CVT's in another data compression setting, namely reduced-order modeling. The idea, just as it is in the Pod setting, is to extract, from a given set of (modified) snapshots $\{\mathbf{x}_j\}_{j=1}^n$ of vectors in $\mathbb{R}^N$, a smaller set of vectors also belonging to $\mathbb{R}^N$. In the POD setting, the reduced set of vectors was the d-dimensional set of POD vectors $\{\boldsymbol{\phi}_j\}_{i=1}^d$. In the CVT setting, the reduced set of vectors is the k-dimensional set of vectors $\{\mathbf{z}_k\}_{i=1}^k$ that are the generators of a centroidal Voronoi tessellation of the set of modified snapshots. See [4, 8, 9] for details. Just as POD produced an optimal reduced basis in the sense that the error $\mathcal{E}$ is minimized, CVT produces an optimal reduced basis in the sense that the error $\mathcal{F}$ is minimized. One can, in principle, determine the dimension d of an effective POD basis, e.g., using the eigenvalues of the correlation matrix. Similarly, one can, in principle, determine the dimension k of an effective CVT basis by examining the (computable) error $\mathcal{F}(\cdot)$.

A natural question is: why should one use CVT instead of POD? Although justifications have to be substantiated through analyses and extensive numerical experiments, heuristically, one can make some arguments. CVT naturally introduces the concept of clustering into the construction of the reduced basis. CVT is "cheaper" than POD; POD involves the solution of an $n \times n$ eigenproblem, where n is the number of snapshots; CVT requires no eigenproblem solution. CVT can handle many more snapshots. Adaptively changing the reduced basis is much less expensive with CVT.

Actually, one does not have to choose between POD and CVT. They may in fact be combined to define hybrid methods which take advantage of the best features of both methods. See [9].

4.2 Computational experiments using CVTs for model reduction

Several computational examples of the use of CVTs for model reduction in flow simulation problems are given in [4]. Here, we provide some of the results given in that paper. The setting is the Navier-Stokes system in a box with inflow at the lower left-hand corner and outflow at upper right-hand corner. Snapshots (500 of them) were generated by sampling a finite element approximation at 500 evenly-spaced times in the interval $[0, 5]$. The inflow velocity was impulsively changed at $t = 0$ and $t = 2.5$. CVT reduced bases were determined from the 500 (modified) snapshots. See [4] for details.

We tested the accuracy resulting from the use of CVT reduced basis for flow simulations with the two inflow conditions depicted in Figure 1. Figure 2 provides plots of the L^2 error vs. time for the CVT reduced-basis solutions for three different cardinalities for the bases. The results of Figure 2 indicate that, at least for this example, CVT-based model reduction is effective. Of course, much more computational experimentation, including comparison with POD-based results, is necessary in order to document the efficacy of CVT-based reduced-order modeling. These experiments are the focus of our current efforts in CVT-based model reduction.

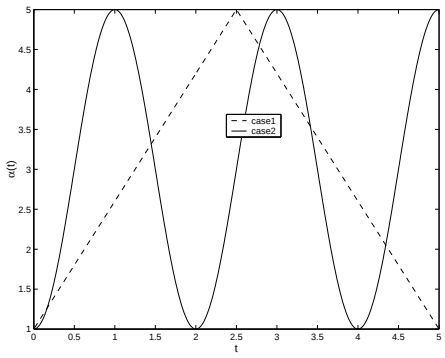


Figure 1: *Two velocity boundary conditions used in testing CVT bases.*

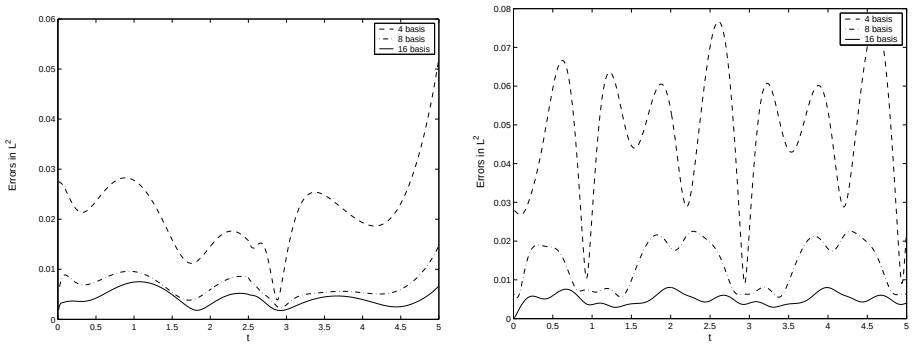


Figure 2: L^2 *errors vs. time in CVT reduced-basis solution vs. time; left: case 1; right: case 2.*

References

[1] N. Aubry, W. Lian, and E. Titi, Preserving symmetries in the proper orthogonal decomposition, *SIAM J. Sci. Comp.* **14**, 1993.

[2] G. Berkooz, P. Holmes, and J. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, *Ann. Rev. Fluid. Mech.* **25**, 1993.

[3] G. Berkooz and E. Titi, Galerkin projections and the proper orthogonal decomposition for equivariant equations, *Phys. Let. A* **174** 1993.

[4] J. Burkardt, M. Gunzburger, and H.-C. Lee, Centroidal Voronoi tessellation-based reduced-order modeling of complex systems, to appear.

[5] E. Christensen, M. Brons, and J. Sorensen, Evaluation of proper orthogonal decomposition-based decomposition techniques applied to parameter-dependent non turbulent flows, *SIAM J. Sci. Comp.* **21**, 2000.

[6] A. Deane, I. Kevrekidis, G. Karniadakis, and S. Orszag, Low-dimensional models for complex geometry flows: applications to grooved channels and circular cylinders, *Phys. Fluids A* **3**, 1991.

[7] Q. Du, V. Faber and M. Gunzburger, Centroidal Voronoi tessellations: applications and algorithms, *SIAM Rev.* **41**, 1999.

[8] Q. Du and M. Gunzburger, Model reduction by proper orthogonal decomposition coupled with centroidal Voronoi tessellation, *Proc. Fluids Engineering Division Summer Meeting*, FEDS2002-31051, ASME, 2002.

[9] Q. Du and M. Gunzburger, Centroidal Voronoi tessellation based proper orthogonal decomposition analysis, *Proc. 8th Conference on Control of Distributed Parameter Systems*, Birkhauser, Basel, 2002.

[10] M. Graham and I. Kevrekidis, Pattern analysis and model reduction: Some alternative approaches to the Karhunen-Lóeve decomposition, *Comp. Chem. Engng.* **20**, 1996.

[11] P. Holmes, J. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, 1996.

[12] P. Holmes, J. Lumley, G. Berkooz, J. Mattingly, and R. Wittenberg, Low-dimensional models of coherent structures in turbulence, *Phys. Rep.* **287**, 1997.

[13] K. Ito and S. Ravindran, A reduced order method for simulation and control of fluid flows, *J. Comput. Phys.* **143** 1998.

[14] L. Ju, Q. Du and M. Gunzburger, Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations, *J. Parallel Comput.* **28**, 2002.

[15] K. Kunisch and S. Volkwein, Control of Burger's equation by a reduced order approach using proper orthogonal decomposition, *JOTA* **102**, 1999.

[16] K. Kunisch and S. Volkwein, Galerkin proper orthogonal decomposition methods for parabolic problems, *Spezialforschungsbereich F003 Optimierung und Kontrolle,* Projektbereich Kontinuierliche Optimierung und Kontrolle, Bericht Nr. 153, Graz, 1999.

[17] D. Lucia, P. King, and P. Beran, Domain decomposition for reduced order modeling of a flow with moving shocks, *AIAA J.*, to appear.

[18] J. Lumley, *Stochastic Tools in Turbulence*, Academic, New York, 1971.

[19] H. Park and M. Lee, Reduction in modes in viscous laminar flows, *Comput. & Fluids*, to appear.

[20] A. Noor, Recent advances in reduction methods for nonlinear problems, *Comput. & Struc.* **13**, 1981, 31-44.

[21] J. Peterson, The reduced basis method for incompressible flow calculations, *SIAM J. Sci. Stat. Comput.* **10**, 1989.

[22] S. Ravindran, A reduced order approach for optimal control of fluids using proper orthogonal decomposition, *Int. J. Numer. Meth. Fluids* **34** 2000.

[23] S. Ravindran, Reduced-order adaptive controllers for fluid flows using POD,*J. Sci. Comput.* **15** 2000.

[24] J. Rodríguez and L. Sirovich, Low-dimensional dynamics for the complex Ginzburg-Landau equations, *Physica D* **43**, 1990.

[25] L. Sirovich, Turbulence and the dynamics of coherent structures, I-III, *Quart. Appl. Math.* **45**, 1987.

[26] N. Smaoui and D. Armbruster, Symmetry and the Karhunen-Loève analysis, *SIAM J. Sci. Comput.* **18**, 1997.

[27] S. Volkwein, Optimal control of a phase field model using the proper orthogonaldecomposition, *ZAMM* **81**, 2001.

[28] S. Volkwein, Proper orthogonal decompositionand singular value decomposition, *Spezialforschungsbereich F003 Optimierung und Kontrolle*, ProjektbereichKontinuierliche Optimierung und Kontrolle, Bericht Nr. 153, Graz, 1999.

Numerical Issues and Influences in the Design of Algebraic Modeling Languages for Optimization

Robert Fourer & David M. Gay
Northwestern University,
Industrial Engineering and Management Science,
2145 Sheridan Road,
Evanston IL 60208 USA
4er@iems.nwu.edu
&
AMPL Optimization LLC

The idea of a modeling language is to describe mathematical problems symbolically in a way that is familiar to people but that allows for processing by computer systems. In particular the concept of an algebraic modeling language, based on objective and constraint expressions in terms of decision variables, has proved to be valuable for a broad range of optimization and related problems.

One modeling language can work with numerous solvers, each of which implements one or more optimization algorithms. The separation of model specification from solver execution is thus a key tenet of modeling language design. Nevertheless, several issues in numerical analysis that are critical to solvers are also important in implementations of modeling languages. So-called presolve procedures, which tighten bounds with the aim of eliminating some variables and constraints, are numerical algorithms that require carefully chosen tolerances and can benefit from directed roundings. Correctly rounded binary-decimal conversion is valuable in portably conveying problem instances and in debugging. Further rounding options offer tradeoffs between accuracy, convenience, and readability in displaying numerical results.

Modeling languages can also strongly influence the development of solvers. Most notably, for smooth nonlinear optimization, the ability to provide numerically computed, exact first and second derivatives has made modeling languages a valuable tool in solver development. The generality of modeling languages has also encouraged the development of more general solvers, such as for optimization problems with equilibrium constraints.

This paper draws from our experience in developing the AMPL modeling language [14, 15] to provide examples in all of the above areas. An associated set of slides [13] shows more completely the contexts from which our examples are taken.

1. Rounding and conversion

AMPL incorporates an interactive `display` command that is designed to produce usefully formatted results with a minimum of effort. Since numbers are often the results of interest, a variety of numerical issues arise.

When linear programs are solved by the simplex method, degenerate basic variables that would take values of zero in exact computation may come out instead having slightly nonzero values, as seen in this AMPL display of a variable named `Make`:

```
ampl: display Make;

:          bands       coils       plate     :=
CLEV      1.91561e-14   1950     3.40429e-14
GARY   1125             1750   300
PITT    775              500   500
```

To suppress these distracting and meaningless entries, AMPL offers the option of specifying that all values less than some "display epsilon" be represented as zeros:

```
ampl: option display_eps 1e-10;

ampl: display Make;

:     bands  coils plate :=
CLEV      0   1950     0
GARY   1125   1750   300
PITT    775    500   500
```

Since tiny magnitudes might have significance in some applications, however, the default display epsilon is left at zero, so that the substitution of zeros never occurs without specific instructions from the modeler.

AMPL shows by default 6 digits of precision, with trailing zeroes suppressed:

```
ampl: display Level;

P1a   450.7
 P3   190.124
P3c  1789.33
```

More or fewer significant digits can be requested:

```
ampl: option display_precision 4;
ampl: display Level;

P1a   450.7
 P3   190.1
P3c  1789

ampl: option display_precision 9;
ampl: display Level;

P1a   450.700489
 P3   190.123755
P3c  1789.33403
```

In some cases, such as where results represent amounts of money, it makes sense to round not to a given precision but to a given decimal place:

```
ampl: display Price;

AA1  16.7051
AC1   5.44585
BC1  48.909
BC2   8.90899
```

```
ampl: option display_round 2;
ampl: display Price;

AA1  16.71
AC1   5.45
BC1  48.91
BC2   8.91
```

These displays rely on numerical routines that provide correctly rounded conversions. *Correctly rounded decimal-to-binary* conversions produce the binary value "closest" to a given decimal number, for a given binary representation and rounding sense (up or down). Clinger [5] showed how to compute these conversions in IEEE double-extended arithmetic, and Gay [16] adapted them to require only double-precision arithmetic. *Correctly rounded binary-to-decimal* conversions produce a decimal representation that, among all those having a specified number of significant digits or digits after the decimal point, comes closest to the original binary value when correctly rounded in a specified rounding sense. Methods for this purpose were proposed by Steele and White [27] and implemented more efficiently by Gay [16]. Gay's work in both cases was motivated in part by AMPL's need for these conversions.

AMPL also incorporates a special option for conversion to "maximum precision" decimal representations:

```
ampl: option display_precision 0;
ampl: display Level;

P1a   450.70048877657206
 P3   190.12375501709528
P3c  1789.334027055151
```

This conversion's result is the *shortest* decimal representation that will yield the original binary representation when correctly rounded back to binary. This option has several uses: to ensure that different numbers in index sets always display differently, to provide for equivalent text and binary forms of communications to solvers, and to assist in diagnosing unexpected behavior of rounding operations.

Experiments reported in [16] showed that the computational times for typical correctly-rounded binary-to-decimal conversions are competitive with the times required by the then available standard C library routines — which did not produce correctly-rounded results. Even conversion to maximum precision is not prohibitively slow, requiring usually between 3 and 10 times the effort of correct binary-to-decimal conversion to 6 places.

As an example of the diagnostic function of maximum precision, consider these results, from a scheduling model, that show how many people are to work each of several numbered schedules:

```
ampl: option display_eps .000001;

ampl: display Work;

 10 28.8    73 28
 18  7.6    87 14.4
 24  6.8   106 23.2
 30 14.4   109 14.4
 35  6.8   113 14.4
 66 35.6   123 35.6
 71 35.6
```

To get implementable results, we might round each fractional number of people up to the next highest integer. Then the total required workforce should be given by

```
ampl: display sum { j in SCHEDS} ceil(Work[j]);
sum{ j in SCHEDS} ceil(Work[j]) = 273
```

When we compute the same total explicitly, however, it comes out with two people fewer:

```
ampl: display 29+8+7+15+7+36+36+28+15+24+15+15+36;
29 + 8 + 7 + 15 + 7 + 36 + 36 + 28 + 15 + 24 + 15 + 15 + 36 = 271
```

The anomaly can be resolved by displaying all nonzero values at maximum precision:

```
ampl: option display_eps 0, display_precision 0;

ampl: display Work;

 10 28.799999999999997        73 28.000000000000018
 18  7.599999999999998        87 14.399999999999995
 24  6.79999999999999         95 -5.876671973951407e-15
 30 14.40000000000001        106 23.200000000000006
 35  6.799999999999995       108  4.685288280240683e-16
 55 -4.939614313857677e-15   109 14.4
 66 35.6                     113 14.4
 71 35.599999999999994       123 35.59999999999999
```

We can see here that `Work[73]` and `Work[108]`, which appeared to have values 28 and 0 when rounded to AMPL's default 6 digits of precision, are in fact slightly greater than those integer values and so are being unexpectedly rounded up.

For situations of this kind, AMPL provides a function `round` that rounds its first argument to the number of places after the decimal point specified by the second argument. Using this function in our example, the correct sum can be obtained by

```
ampl: display sum { j in SCHEDS} ceil(round(Work[j],6));
sum{ j in SCHEDS} ceil(round(Work[j], 6)) = 271
```

Alternatively, the modeler may set an option `solution_round` to specify that *all* values returned by the solver be automatically rounded to a certain number of decimal places, when AMPL retrieves them from the solver. Setting this option to 6, for example, would cause our original expression `sum {j in SCHEDS} ceil(Work[j])` to come out to 271 as expected.

2. Presolve

A presolve phase attempts to reduce the numbers of variables and constraints in a problem instance before sending it to a solver. The underlying motivation for including a presolve phase in AMPL — rather than leaving this operation to each solver — is to provide consistent handling of simple bounds, which most solvers can treat specially to advantage. Explicit simple bounding constraints are removed and the bounds are folded into the definitions of the variables, so that it makes no difference whether a model declares

```
var Sell { PROD,1..T} >= 0;
subject to MLim { p in PROD, t in 1..T}: Sell[p,t] <= market[p,t];
```

or

```
var Sell { p in PROD, t in 1..T} >= 0, <= market[p,t];
```

As a result a modeler need not appreciate the importance of simple bounds to gain the computational advantages of them.

AMPL also incorporates a more powerful presolve, due to Brearley, Mitra and Williams [3], that uses known bounds together with linear constraints in order to deduce tighter bounds. The idea is simple. If we know for example that $l_j \le x_j \le u_j$ for $j = 1, \ldots, n$, and if we have a constraint $\sum_{j=1}^{n} a_{rj} x_j \le b_r$ with $a_{rs} > 0$, then we can deduce

$$x_s \le \frac{1}{a_{rs}} \left(b_r - \sum_{\substack{j \in \mathcal{P} \\ j \ne s}} a_{rj} l_j - \sum_{\substack{j \in \mathcal{N} \\ j \ne s}} a_{rj} u_j \right),$$

where

$$\mathcal{P} \equiv \{j = 1, \ldots, n : \ a_{rj} > 0\}, \quad \mathcal{N} \equiv \{j = 1, \ldots, n : \ a_{rj} < 0\}.$$

Other cases are handled similarly, as described in [12]. Whenever any such inferred bound is tighter than a known bound, it replaces the known bound. Thus progressively tighter bounds are sometimes achieved by iterating several times through the nonzero coefficients a_{ij} of the relevant constraints.

This presolving procedure can reduce problem size in several ways. If the inferred bounds $l_s = u_s$, then x_s can be fixed and eliminated from the problem; if $l_s > u_s$ then no feasible solution is possible. If

$$\sum_{j \in \mathcal{P}} a_{rj} u_j + \sum_{j \in \mathcal{N}} a_{rj} l_j \le b_r,$$

then the rth constraint is redundant and can be dropped; if

$$\sum_{j \in \mathcal{P}} a_{rj} l_j + \sum_{j \in \mathcal{N}} a_{rj} u_j > b_r,$$

then no feasible solution is possible, and the same with $\ge$ implies that the constraint can be "fixed" to $\sum_{j=1}^{n} a_{rj} x_j = b_r$. (Again there are other cases handled similarly, and described in [12].) These are numerical tests, so in practice they may have to be

carried out with respect to tolerances. In fact a variety of tolerance options have proved necessary to get AMPL's presolve to act the way modelers want and expect in a range of circumstances.

As a simple example, consider an AMPL model that specifies the data values as

```
set PROD;  # products

param avail >= 0;          # production hours available in a week
param commit { PROD} >= 0;  # lower limit on tons sold in a week
param market { PROD} <= 0;  # upper limit on tons sold in a week
```

and for which the variables `Make[p]` and single constraint `Time` are defined as follows:

```
var Make { p in PROD} >= commit[p], <= market[p];

subject to Time: sum { p in PROD} (1/rate[p]) * Make[p] <= avail;
```

If we set hours available to 13, then AMPL's presolve phase reports that no feasible solution is possible:

```
ampl: let avail := 13;
ampl: solve;

presolve: constraint Time cannot hold:
    body <= 13 cannot be >= 13.2589; difference = -0.258929
```

The reason for this infeasibility is not hard to see. If in the constraint we use the lower bounds `commit[p]` as the values of the variables, we see that hours available must be at least 13.2589 (to six digits) to allow the minimum possible production:

```
ampl: display sum { p in PROD} (1/rate[p]) * commit[p];
sum{ p in PROD} 1/rate[p]*commit[p] = 13.2589
```

AMPL's previous message that "`body <= 13 cannot be >= 13.2589`" is reporting this same observation.

Here are AMPL's responses to three more values of `avail`:

```
ampl: let avail := 13.2589;
ampl: solve;

presolve: constraint Time cannot hold:
    body <= 13.2589 cannot be >= 13.2589; difference = -2.85714e-05

ampl: let avail := 13.25895;
ampl: solve;

MINOS 5.5: optimal solution found.
0 iterations, objective 61750.10714

ampl: let avail := 13.258925;
ampl: solve;

presolve: constraint Time cannot hold:
    body <= 13.2589 cannot be >= 13.2589; difference = -3.57143e-06
Setting $presolve_eps >= 4.29e-06 might help.
```

In the first case we see that 13.2589 is not quite enough; evidently it is rounded down

when it is converted to six significant decimal digits. In the second case, a slightly higher value, 13.25895, proves sufficient to allow a feasible solution, and hence an optimal value can be reported. For the third case, the value has been taken between the previous two, at 13.258925. Again presolve reports no feasible solution, but it suggests a new tolerance setting that might make a difference.

There are actually two tolerances at work in this example. First, infeasibility is detected in the rth constraint if (continuing our example)

$$b_r - \sum_{j \in \mathcal{P}} a_{rj} l_j - \sum_{j \in \mathcal{N}} a_{rj} u_j$$

is less than the negative of the value in the AMPL option `presolve_eps`. Once infeasibility has been detected, presolve also calculates a (necessarily larger) hypothetical value of `presolve_eps` that might lead to a determination of feasibility instead. This value is reported if it is considered small enough to be meaningful — specifically, if it is no larger than the value in AMPL option `presolve_epsmax`. Presolve reports that this increase "might" make a difference because actions of other presolve steps may have additional effects that cannot be quickly foreseen.

The tolerances `presolve_eps` and `presolve_epsmax` relate to determining infeasibility through inconsistent inferred bounds on variables or constraints. Analogous tolerances `presolve_fixeps` and `presolve_fixepsmax` play the same roles in the determination of whether bounds are close enough that a variable should be fixed or an inequality constraint should be made an equality. And also the rth constraint is dropped as redundant when

$$b_r - \sum_{j \in \mathcal{P}} a_{rj} u_j - \sum_{j \in \mathcal{N}} a_{rj} l_j$$

is at least as large as a tolerance given by `constraint_drop_tol`.

All three of `presolve_eps`, `presolve_fixeps`, and `constraint_drop_tol` are set by default to zero, allowing in effect no tolerance for computational imprecision. Experience suggests that, when the presolve routines can be written to use the *directed rounding* — lower bounds toward $-\infty$ and upper bounds toward $+\infty$ — that is available with IEEE arithmetic [24], then decisions on infeasibility, equality, and redundancy are made correctly without any assistance from explicit tolerances (as long as the problem involves exactly known data). The tolerance options are retained, however, for use on platforms that do not offer directed rounding and with problems having data of limited precision.

A final set of tolerances relate to variables that are allowed to take only integer values. When a new lower bound is inferred for an integer variable, that bound may be immediately rounded up to the next highest integer. Similarly, an inferred upper bound may be immediately rounded down. In these situations the AMPL presolver must again take care not to round up a bound that is only slightly greater than an integer, or to round down a bound that is only slightly less than an integer. Presolve thus only rounds a lower bound l_j up or an upper bound u_j down if $l_j - \text{floor}(l_j)$ or $\text{ceil}(u_j) - u_j$ is greater than a given tolerance. The tolerance value is taken from the value of the AMPL option `presolve_inteps`, which is set to `1e-6` by default. Messages suggesting that a slightly higher value of the tolerance might make a difference are governed by an

option `presolve_intepsmax` that works with `presolve_inteps` in the same way that `presolve_epsmax` works with `presolve_eps`.

Presolve operations can be performed very efficiently. For a run that reported

```
Presolve eliminates 1769 constraints and 8747 variables.

19369 variables, all linear
3511 constraints, all linear; 150362 nonzeros
1 linear objective; 19369 nonzeros
```

AMPL's processing required only about 2 seconds in total on a fast PC, of which about $\frac{1}{2}$ second was in presolve. For a larger run that reported

```
Presolve eliminates 32989 constraints and 54819 variables.

327710 variables, all linear
105024 constraints, all linear
1 linear objective; 317339 nonzeros
```

AMPL's processing time was 23 seconds, of which only 4 were in presolve.

3. Modeling language influences on solvers

In addition to linear and mixed-integer programs, algebraic modeling languages can express a broad variety of nonlinear programs. The expressive abilities of modeling languages have in fact at times gone beyond the computational abilities of available solvers. Extensions to solver technology have come about as a result.

We summarize here two cases in which the generality of AMPL has encouraged new solver developments. One involves higher-order derivatives, and the other an extension for specifying equilibrium conditions.

Second derivatives. Expressing a nonlinear program in an algebraic modeling language involves nothing more than using the available operators to form nonlinear expressions, possibly in conjunction with elementary nonlinear functions. The following is for example a nonlinear objective function in AMPL:

```
minimize Energy:
   sum { (j,k) in PAIRS} x[j,k] *
      (c[j,k] + log (x[j,k] / sum { m in J: (m,k) in PAIRS} x[m,k]));
```

Nonlinear constraints are written using the same sorts of expressions along with relational operators.

Given a nonlinear model and a data instance, AMPL generates an expression graph for the nonlinear part of each objective and constraint. The AMPL version of a nonlinear solver incorporates a *driver* that sets up a data structure to represent the expression graphs. The solver later calls AMPL library routines that use the graph data structures to evaluate the nonlinear functions; specifically, the solver passes a current iterate to the AMPL routines, and gets back the values of the nonlinear function at that iterate. A detailed explanation of this process is given by Gay [20]. (Solvers that work directly with expression graphs can also be accommodated by AMPL, but we are concerned here only with conventional solvers that address nonlinearities through function evaluations.)

Many solvers require gradients as well as function values. AMPL's evaluation routines apply *automatic differentiation* [21] to compute gradient values at the same time as function values. As described by Gay [17], AMPL employs in particular an approach known as "backward" AD, which makes two passes through the graph for each nonlinear expression $f(x)$:

- ▷ a forward sweep computes $f(x)$ and saves information on $\partial o/\partial a_i$ for each argument a_i to each operation o;
- ▷ a backward sweep recursively computes $\partial f/\partial o$ for each operation o, and hence ultimately $\nabla f(x)$.

The work of computing the gradient vector can be shown to be bounded by a small multiple of the work of computing the function value alone, though possibly at a large cost in additional space in the expression-graph data structure. Backward AD is more accurate and efficient than finite differencing, and has substantially better worst-case efficiency than straightforward symbolic differentiation.

The concepts of automatic differentiation can be applied as well to computing second-derivative information. Hessian-vector products $\nabla^2 f(x)v$ can be determined by applying backward AD to compute $v^T \nabla f(x)$, or equivalently to compute $\nabla_x (df(x+\tau v)/d\tau|_{\tau=0})$. The entire Hessian can thus be determined through the Hessian-vector products $\nabla^2 f(x) e_j$ with the n unit vectors. Alternatively, if the function of interest can be determined to have a group partially separable structure [6, 22, 23],

$$f(x) \;=\; \sum_{i=1}^{q} \theta_i \left(\sum_{j=1}^{r_i} \phi_{ij}(U_{ij} x) \right) \;=\; \sum_{i=1}^{q} \theta_i \left(\Phi_i(x) \right),$$

where U_{ij} is $m_{ij} \times n$ with $m_{ij} \ll n$, $\phi_{ij} \colon \Re^{m_{ij}} \mapsto \Re$, and $\theta_i \colon \Re \mapsto \Re$, then the gradient and Hessian also have simplified structures,

$$\nabla f(x) = \sum_{i=1}^{q} \theta_i'(\Phi_i(x)) \nabla \Phi_i(x),$$

$$\nabla^2 f(x) = \sum_{i=1}^{q} \left(\theta_i'(\Phi_i(x)) \nabla^2 \Phi_i(x) + \theta_i''(\Phi_i(x)) \nabla \Phi_i(x) \nabla \Phi_i(x)^T \right)$$

where

$$\nabla \Phi_i(x) = \sum_{j=1}^{r_i} U_{ij}^T \nabla \phi_{ij}(U_{ij} x), \quad \nabla^2 \Phi_i(x) = \sum_{j=1}^{r_i} U_{ij}^T \, \nabla^2 \phi_{ij}(U_{ij} x) \, U_{ij}.$$

Thus the Hessian can be computed from sums of outer products involving the generally much smaller Hessians of the component functions ϕ_{ij}.

The AMPL function-evaluation routines have been extended as described in Gay [18, 19] to provide Hessians in dense or sparse form, as a full (symmetric) matrix or just the lower triangle. Hessian-vector products are also available, to be used for example in iteratively solving the linear equations that define the steps taken by some algorithms. Given Lagrange multipliers as well as the current iterate, the routines return the Hessian

of the Lagrangian or its product with a vector. AMPL's library routines also perform an initial walk of the expression graphs to detect group partially separable structure, using a hashing scheme to spot common subexpressions; the modeler need not identify this structure.

How has this feature been critical to nonlinear solver development? Its addition to AMPL encouraged researchers' interest in extending interior-point methods to nonlinear optimization problems. These methods can be viewed as solving a nonlinear problem such as

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{Subject to} & h_i(x) \geq 0, \quad i = 1, \ldots, m \end{array}$$

by applying a modified Newton's method to the "centered" Karush-Kuhn-Tucker optimality conditions,

$$\nabla f(x) = \nabla h(x)^T y, \;\; h(x) = w, \;\; WYe = \mu e,$$

where $W = \mathrm{diag}(w)$, $Y = \mathrm{diag}(y)$. The result is a Newton linear system of the form

$$\begin{bmatrix} -\nabla^2 \left(f(x) - h(x)^T y \right) & \nabla h(x)^T \\ \nabla h(x) & WY^{-1} \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \cdots$$

that incorporates the Hessian of the Lagrangian in the upper left-hand block.

Given this extension to AMPL, a solver developer need only add an AMPL driver to have full access to the necessary second-derivative information. Hundreds of standard test problems — as well as any of the developer's own benchmarks — are immediately available, without any need to write and debug new code for Hessian computations. This facility has contributed to rapid development of new interior-point codes such as LOQO [28], KNITRO [4], and MOSEK [1]. The success of these codes has in turn stimulated developers of other kinds of codes, such as CONOPT [2] and PATH [9], to pursue possibilities for taking advantage of second-derivative information.

Complementarity problems. A classical complementarity condition says that two inequalities must hold, at least one with equality. Collections of complementarity conditions, known as complementarity problems, arise as equilibrium conditions for problems in economics and in engineering, and represent optimality conditions for nonlinear programs, bi-level programs, and bi-matrix games.

To support solvers designed specifically for complementarity problems, the operator `complements` has been added to the AMPL language [10]. Thus it is possible to state a collection of conditions such as

```
subject to Lev_Compl { j in ACT}:
   Level[j] >= 0 complements
      sum { i in PROD} Price[i] * io[i,j] <= cost[j];
```

AMPL also provides support for "mixed" complementarity conditions, which hold between a double inequality and an expression:

```
subject to Lev_Compl { j in ACT}:
   level_min[j] <= Level[j] <= level_max[j] complements
      cost[j] - sum { i in PROD} Price[i] * io[i,j];
```

This condition is satisfied if the double-inequality holds with equality at its lower limit and the expression is ≥ 0, or if the double-inequality holds with equality at its upper limit and the expression is ≤ 0, or otherwise if the expression $= 0$.

The initial motivation for complementarity conditions in AMPL was to support codes such as PATH [7] that solve "square" systems of such conditions — where the number of variables equals the number of complementarity conditions plus the number of equality constraints. Once a convenient notation for complementarity conditions was available, however, there was nothing to stop modelers from writing non-square systems and adding objective functions. Indeed, uses for these "mathematical programs with equilibrium constraints" — or MPECs — are of growing interest.

Thus the availability of complementarity conditions in AMPL encouraged development of algorithmic approaches to solving general MPECs, especially through the application of nonlinear optimization methods to modifications of the MPEC conditions [8, 11, 25, 26]. The necessary modifications can be carried out by a solver driver, so that AMPL's convenient representation of the complementarity conditions is maintained at the modeler's level.

References

[1] E.D. Andersen and K.D. Andersen, "The MOSEK Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm." In *High Performance Optimization,* H.Frenk, K.Roos, T.Terlaky, and S.Zhang, eds., Kluwer Academic Publishers (2000) pp. 197–232.

[2] ARKI Consulting & Development A/S, "The CONOPT Algorithm." At `www.conopt.com/Algorithm.htm`.

[3] A.L. Brearley, G. Mitra and H.P. Williams, "Analysis of Mathematical Programming Problems Prior to Applying the Simplex Method." *Mathematical Programming* **8** (1975) 54–83.

[4] R. Byrd, M.E. Hribar, and J. Nocedal, "An Interior Point Method for Large Scale Nonlinear Programming." *SIAM Journal on Optimization* **9** (1999) 877–900.

[5] W.D. Clinger, "How to Read Floating Point Numbers Accurately." In *Proceedings of the ACM SIGPLAN'90 Conference on Programming Language Design and Implementation,* White Plains, NY, June 20-22, 1990, pp. 92–101.

[6] A.R. Conn, N.I.M. Gould, and Ph.L. Toint, *LANCELOT, a Fortran Package for Large-Scale Nonlinear Optimization (Release A).* Springer Series in Computational Mathematics 17, Springer-Verlag (1992).

[7] S.P. Dirkse and M.C. Ferris, "The PATH Solver: A Non-Monotone Stabilization Scheme for Mixed Complementarity Problems." *Optimization Methods and Software* **5** (1995) 123–156.

[8] M.C. Ferris, S.P. Dirkse and A. Meeraus, "Mathematical Programs with Equilibrium Constraints: Automatic Reformulation and Solution via Constrained Optimization." Numerical Analysis Group Research Report NA-02/11, Oxford University Computing Laboratory, Oxford University (2002), `web.comlab.ox.ac.uk/oucl/publications/natr/na-02-11.html`.

[9] M.C. Ferris and K. Sinapiromsaran, "Formulating and Solving Nonlinear Programs as Mixed Complementarity Problems." In *Optimization,* V.H. Nguyen, J.J. Strodiot and P. Tossings, eds., volume 481 of *Lecture Notes in Economics and Mathematical Systems,* Springer-Verlag (2000) pp. 132–148.

[10] M.C. Ferris, R. Fourer and D.M. Gay, "Expressing Complementarity Problems in an Algebraic Modeling Language and Communicating Them to Solvers." *SIAM Journal on Optimization* **9** (1999) 991–1009.

[11] R. Fletcher and S. Leyffer, "Numerical Experience with Solving MPECs as NLPs." University of Dundee Report NA/210 (August 2002). At `www.mcs.anl.gov/~leyffer/MPEC-num-2.ps.Z`.

[12] R. Fourer and D.M. Gay, "Experience with a Primal Presolve Algorithm." In*Large Scale Optimization: State of the Art,* W.W. Hager, D.W. Hearn and P.M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, 1994, pp. 135–154.

[13] R. Fourer and D.M. Gay, "Numerical Issues and Influences in the Design of Algebraic Modeling Languages for Optimization." Slides for presentation at the 20th Biennial Conference on Numerical Analysis, University of Dundee, Scotland, 24-27 June 2003. At `www.ampl.com/REFS`.

[14] R. Fourer, D.M. Gay and B.W. Kernighan, "A Modeling Language for Mathematical Programming." *Management Science* **36** (1990) 519–554.

[15] R. Fourer, D.M. Gay and B.W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming,* 2nd edition. Duxbury Press, Belmont, CA (2003).

[16] D.M. Gay, "Correctly Rounded Binary-Decimal and Decimal-Binary Conversions." Numerical Analysis Manuscript 90-10, AT&T Bell Laboratories, Murray Hill, NJ (1990). At `www.ampl.com/REFS/rounding.pdf`.

[17] D.M. Gay, "Automatic Differentiation of Nonlinear AMPL Models." In *Automatic Differentiation of Algorithms: Theory, Implementation, and Application,* A. Griewank and G. Corliss, eds., SIAM, Philadelphia, PA (1991) pp. 61–73.

[18] D.M. Gay, "More AD of Nonlinear AMPL Models: Computing Hessian Information and Exploiting Partial Separability." In *Computational Differentiation: Techniques, Applications, and Tools,* M. Berz, C. Bischof, G. Corliss and A. Griewank, eds., SIAM, Philadelphia, PA (1996) pp. 173–184.

[19] D.M. Gay, "Automatically Finding and Exploiting Partially Separable Structure in Nonlinear Programming Problems." Technical report, Bell Laboratories, Murray Hill, NJ (1996). At `www.ampl.com/REFS/psstruc.pdf`.

[20] D.M. Gay, "Hooking Your Solver to AMPL." Technical report, Bell Laboratories, Murray Hill, NJ (1993; revised 1994, 1997). At `www.ampl.com/REFS/hooking2.pdf` or `www.ampl.com/REFS/HOOKING`.

[21] A. Griewank, "On Automatic Differentiation." In *Mathematical Programming: Recent Developments and Applications,* M. Iri and K. Tanabe, eds., Kluwer Academic Publishers, Dordrecht (1989) pp. 83–107.

[22] A. Griewank and Ph.L. Toint, "On the Unconstrained Optimization of Partially Separable Functions." In *Nonlinear Optimization 1981,* M.J.D. Powell, ed., Academic Press (1982) pp. 301–312

[23] A. Griewank and Ph.L. Toint, "Partitioned Variable Metric Updates for Large Structured Optimization Problems." *Numerische Mathematik* **39** (1982) 119–137.

[24] *IEEE Standard for Binary Floating-Point Arithmetic,* Institute of Electrical and Electronics Engineers, New York, NY (1985). ANSI/IEEE Standard 754-1985.

[25] S. Leyffer, "Mathematical Programs with Complementarity Constraints." Argonne National Laboratory Preprint ANL/MCS-P1026-0203 (February 2003). At `www.mcs.anl.gov/~leyffer/mpec-survey.ps.gz`.

[26] S. Leyffer, "Complementarity Constraints as Nonlinear Equations: Theory and Numerical Experiences." Argonne National Laboratory Preprint ANL/MCS-P1054-0603 (June 2003). At `www.mcs.anl.gov/~leyffer/MPEC-NCP-02.ps.gz`.

[27] G.L. Steele and J.L. White, "How to Print Floating-Point Numbers Accurately." In *Proceedings of the ACM SIGPLAN'90 Conference on Programming Language Design and Implementation,* White Plains, NY, June 20-22, 1990, pp. 112–126.

[28] R.J. Vanderbei, "LOQO: an Interior Point Code for Quadratic Programming." *Optimization Methods and Software* **11-12** (1999) 451–484.

Solution of non-symmetric, real positive linear systems

G. H. Golub

Department of Computer Science,
Stanford University,
Serra Street, Stanford 94305, USA

golub@stanford.edu

We discuss methods which use a Hermitian/Skew-Hermitian splitting (HSS) iteration and an inexact variant for solving non-symmetric linear equations which are real positive. Theoretical analyses show that the HSS method converges unconditionally to the unique solution of the system of linear equations. Moreover, we derive an upper bound of the contraction factor of the HSS iteration which is dependent solely on the spectrum of the Hermitian part. Numerical examples are presented to illustrate the effectiveness of both HSS and IHSS iterations. In addition, several important generalizations are presented.

The talk is based on the technical reports listed below, the most pertinent being [7]. These reports may be downloaded from

http://sccm.stanford.edu/nf-publications-tech.html#start-2001

References

[1] Analysis of a Preconditioned Iterative Method for the Convection-Diffusion Equation, Daniele Bertaccini, Gene H. Golub, Stefano Serra-Capizzano. Stanford University Report SCCM-03-13.

[2] Block Triangular and Skew-Hermitian Splitting Methods for Positive Definite Linear Systems, Zhong-Zhi Bai, Gene H. Golub, Lin-Zhang Lu, Jun-Feng Yin. Stanford University Report SCCM-03-09.

[3] An Iterative Method for Generalized Saddle Point Problems, Michele Benzi and Gene H. Golub. Stanford University Report SCCM-02-14.

[4] Optimization of the Hermitian and Skew-Hermitian Splitting Iteration for Saddle-Point Problems, Michele Benzi, Martin J. Gander, and Gene H. Golub. Stanford University Report SCCM-02-13.

[5] Preconditioned Hermitian and Skew-Hermitian Splitting Methods for non-Hermitian Positive Semidefinite Linear Systems, Zhong-Zhi Bai, Gene H. Golub and Jian-Yu Pan. Stanford University Report SCCM-02-12.

[6] Preconditioned HSS methods for the solution of non-Hermitian positive definite linear systems, Daniele Bertaccini, Gene H. Golub, Stefano S. Capizzano and Cristina Tablino Possio. Stanford University Report SCCM-02-11.

[7] Hermitian and Skew-Hermitian Splitting Methods For Non-Hermitian Positive Definite Linear Systems, Zhong-Zhi Bai, Gene H. Golub, Michael K. Ng. Stanford University Report SCCM-01-06.

School of Mathematics

Linear Algebra Techniques in Interior Point Methods for Optimization

Jacek Gondzio
Email: J.Gondzio@ed.ac.uk
URL: http://www.maths.ed.ac.uk/~gondzio

NA Conference, Dundee, June 2003

Outline: Direct Methods

- Linear Algebra in IPMs
- LP, QP, NLP: Linear Algebra is the same
- Symmetric Systems:
  - Positive Definite vs Indefinite Systems
  - Quasi-definite Systems
  - Primal and Dual Regularization
- Unavoidable Ill-conditioning
  - IPM Scaling Matrices
  - Dikin's Bound
- Primal-Dual Regularized Factorization
- Exploiting Structure in IPMs

Linear Algebra of IPM for LP

First order optimality conditions

$$\begin{aligned} Ax &= b, \\ A^T y + s &= c, \\ XSe &= \mu e. \end{aligned}$$

Newton's direction

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^T & I \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} \xi_p \\ \xi_d \\ \xi_\mu \end{bmatrix},$$

where

$$\begin{bmatrix} \xi_p \\ \xi_d \\ \xi_\mu \end{bmatrix} = \begin{bmatrix} b - Ax \\ c - A^T y - s \\ \mu e - XSe \end{bmatrix}.$$

Use the third equation to eliminate

$$\begin{aligned} \Delta s &= X^{-1}(\xi_\mu - S\Delta x) \\ &= -X^{-1}S\Delta x + X^{-1}\xi_\mu, \end{aligned}$$

from the second equation and get

$$\begin{bmatrix} -\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} \xi_d - X^{-1}\xi_\mu \\ \xi_p \end{bmatrix}.$$

where $\Theta = XS^{-1}$ is a diagonal scaling matrix.

IPMs: LP, QP & NLP

Augmented system in **LP**

$$\begin{bmatrix} -\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} r \\ h \end{bmatrix}.$$

Eliminate Δx from the first equation and get normal equations

$$(A\Theta A^T)\Delta y = g.$$

Augmented system in **QP**

$$\begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} r \\ h \end{bmatrix}.$$

Eliminate Δx from the first equation and get normal equations

$$(A(Q + \Theta^{-1})^{-1}A^T)\Delta y = g.$$

Augmented system in **NLP**

$$\begin{bmatrix} Q(x,y) & A(x)^T \\ A(x) & -ZY^{-1} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} r \\ h \end{bmatrix}.$$

Eliminate Δx from the first equation and get normal equations

$$(AQ^{-1}A^T + \Theta^{-1})\Delta y = g.$$

Factorization of Indefinite Matrix

Two step solution method:

- factorization to LDL^T form,
- backsolve to compute direction Δy.

Two options are possible:

1. Replace diagonal matrix D with a block-diagonal one and allow 2×2 (indefinite) pivots

$$\begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & a \\ a & d \end{bmatrix}.$$

Hence obtain a decomposition $H = LDL^T$ with **block-diagonal** D.

2. Regularize indefinite matrix to produce a **quasidefinite** matrix

$$K = \begin{bmatrix} -E & A^T \\ A & F \end{bmatrix},$$

where
$E \in \mathcal{R}^{n\times n}$ is positive definite,
$F \in \mathcal{R}^{m\times m}$ is positive definite, and
$A \in \mathcal{R}^{m\times n}$ has full row rank.

5

Quasidefinite Matrices

A symmetric matrix is called **quasidefinite** if

$$K = \begin{bmatrix} -E & A^T \\ A & F \end{bmatrix},$$

where $E \in \mathcal{R}^{n\times n}$ and
$F \in \mathcal{R}^{m\times m}$ are positive definite, and
$A \in \mathcal{R}^{m\times n}$ has full row rank.

Symmetric nonsingular matrix K is *factorizable* if there exists a diagonal matrix D and a unit lower triangular matrix L such that $K = LDL^T$.

The symmetric matrix K is *strongly factorizable* if for any permutation matrix P a factorization $PKP^T = LDL^T$ exists.

Vanderbei (1995) proved that
Symmetric QDFM's are strongly factorizable.
SIOPT 5 (1995) 100-113.

For any quasidefinite matrix
there exists a **Cholesky-like** factorization

$$\bar{H} = LDL^T,$$

where
D is **diagonal** but **not positive definite**:
has n negative pivots;
and m positive pivots.

6

From Indefinite to Quasidefinite Matrix

Indefinite matrix

$$H = \begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix}.$$

Vanderbei *SIOPT* 5 (1995) 100-113.
Replace $Ax = b$ with $Ax + s = b$

$$H_V = \begin{bmatrix} -\Theta_s^{-1} & 0 & I \\ 0 & -Q - \Theta^{-1} & A^T \\ I & A & 0 \end{bmatrix}$$

and eliminate Θ_s^{-1}

$$K = \begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & \Theta_s \end{bmatrix}.$$

Saunders (1996) SIAM Adams & Nazareth (eds)

$$H_S = \begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix} + \begin{bmatrix} -\gamma^2 I_n & 0 \\ 0 & \delta^2 I_m \end{bmatrix},$$

where $\gamma\delta \geq \sqrt{\varepsilon} = 10^{-8}$.

Altman & Gondzio *OMS* 11-12 (99) 275-302.
Use **dynamic regularization**

$$\bar{H} = \begin{bmatrix} -\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} + \begin{bmatrix} -R_p & 0 \\ 0 & R_d \end{bmatrix},$$

$R_p \in \mathcal{R}^{n\times n}$ is a *primal* regularization
$R_d \in \mathcal{R}^{m\times m}$ is a *dual* regularization.

7

Primal Regularization

Primal barrier problem

$$\begin{array}{ll} \min & z_P = c^T x + \frac{1}{2} x^T Q x - \mu \sum_{j=1}^n (\ln x_j + \ln s_j) \\ \text{s. to} & Ax = b, \\ & x + s = u, \\ & x, s > 0 \end{array}$$

$$\begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} f \\ h \end{bmatrix}.$$

Primal **regularized** barrier problem

$$\begin{array}{ll} \min & z_P + \dfrac{1}{2}(x - x_0)^T R_p (x - x_0) \\ \text{s. to} & Ax = b, \\ & x + s = u, \\ & x, s > 0 \end{array}$$

$$\begin{bmatrix} -Q - \Theta^{-1} - R_p & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} f' \\ h \end{bmatrix},$$

where

$$f' = f - R_p(x - x_0).$$

8

Dual Regularization

Dual barrier problem

$$\begin{array}{ll} \max & z_D = b^T y - u^T w - \frac{1}{2}x^T Q x + \mu \sum_{j=1}^{n} (\ln z_j + \ln w_j) \\ \text{s. to} & A^T y + z - w - Qx = c, \\ & x \geq 0, z, w > 0 \end{array}$$

$$\begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} f \\ h \end{bmatrix}.$$

Dual **regularized** barrier problem

$$\begin{array}{ll} \max & z_D - \frac{1}{2}(y - y_0)^T R_d (y - y_0) \\ \text{s. to} & A^T y + z - w - Qx = c, \\ & x \geq 0, z, w > 0 \end{array}$$

$$\begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & R_d \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} f \\ h' \end{bmatrix},$$

where

$$h' = h - R_d(y - y_0).$$

9

HOPDM Solver for LP, QP and NLP
Higher Order Primal Dual Method

Problem	Dimensions				LOQO	HOPDM
	m	n	nz(A)	nz(Q)	nz(L)	nz(L)
nug12	3192	8856	44244	0	3091223	1969957
nug15	6330	22275	110700	0	-	7374972
cvxqp1_m	500	1000	1498	2984	71487	75973
cvxqp1_l	5000	10000	14998	29984	4056820	3725045
cvxqp2_m	250	1000	749	2984	52917	51923
cvxqp2_l	2500	10000	7499	29984	2923584	2754141
cvxqp3_m	750	1000	2247	2984	79957	90433
cvxqp3_l	7500	10000	22497	29984	4411197	4291057

200 MHz Pentium II PC, Linux.

Problem	LOQO		HOPDM	
	iters	time	iters	time
nug12	24	4417.7	13	1140.3
nug15	-	-	15	10276.6
cvxqp1_m	32	13.78	9	6.63
cvxqp1_l	72	18361.1	11	2874.4
cvxqp2_m	16	4.06	9	4.02
cvxqp2_l	25	3849.4	8	1353.7
cvxqp3_m	49	25.45	9	9.11
cvxqp3_l	100	27447.6	8	2461.2

10

Ill-conditioning

Assume **Normal Equations** are used in LP and the feasible IPM is used ($\xi_p = 0$ and $\xi_d = 0$)

$$(A\Theta A^T)\Delta y \;=\; A\Theta r,$$

where $\Theta = XS^{-1}$ and $r = -X^{-1}\xi_\mu$.

Optimal Partition:

Basic variables	$x_B \to x_B^* > 0$	$s_B \to s_B^* = 0$
Non-basic variables	$x_N \to x_N^* = 0$	$s_N \to s_N^* > 0$

For **basic** variables: $\Theta_j = x_j/s_j \to \infty$;
For **non-basic** variables: $\Theta_j = x_j/s_j \to 0$.

Hence

$$A\Theta A^T = \sum_{j\in\mathcal{B}} \theta_j a_{.j} a_{.j}^T + \sum_{j\in\mathcal{N}} \theta_j a_{.j} a_{.j}^T \to \sum_{j\in\mathcal{B}} \theta_j a_{.j} a_{.j}^T.$$

The matrix $H = A\Theta A^T$ usually has a huge condition number $\kappa(H)$. Although $\kappa(H) \gg 1/\epsilon$, where ϵ is the relative precision of the computer (e.g. $\epsilon = 10^{-16}$), IPMs do converge.

11

Dikin's Bound

Theorem: (Dikin, 1974)
Upravlaemye Sistemy 12 (1974) pp 54-60.

Let $A \in \mathcal{R}^{m\times n}$ be a full row rank matrix;
g be a vector of dimension n; and
D_+ be the set of $n \times n$ diagonal positive definite matrices.
Then

$$\sup_{D\in D_+} \|(ADA^T)^{-1}ADg\| \;=\; \max_{\mathcal{J}\in\mathcal{J}(A)} \|A_{\mathcal{J}}^{-T} g_{\mathcal{J}}\|$$

$$\sup_{D\in D_+} \|(ADA^T)^{-1}AD\| \;=\; \max_{\mathcal{J}\in\mathcal{J}(A)} \|A_{\mathcal{J}}^{-T}\|$$

where $\mathcal{J}(A)$ is the set of column indices associated with nonsingular $m \times m$ submatrices of A.

Corollary:
The linear system arising in IPMs for LP

$$(A\Theta A^T)\Delta y \;=\; A\Theta r,$$

produces more accurate solutions than those one could have expected from a "classical" worst-case analysis.

12

Forsgren and Sporre (2001) generalized Dikin's result for a subclass of positive definite weight matrices W. *SIMAX* 22 (2001) 42-56.

Lemma:
Let $A \in \mathcal{R}^{m \times n}$ be a full row rank matrix;
g be a vector of dimension n; and
W_+ be the set of $n \times n$ matrices defined as

$$W = \sum_{i=1}^{k} \alpha_i W_i,$$

where $\alpha_i > 0$ and $W_i = U_i D_i U_i^T$ with U_i bounded and D_i diagonal positive definite $\forall i = 1, ..., k$.
Then

$$\sup_{W \in W_+} \|(AWA^T)^{-1} AWg\|$$
$$\sup_{W \in W_+} \|(AWA^T)^{-1} AW\|$$

are bounded.

This Lemma extends Dikin's result to quadratic and nonlinear optimization.

The Lemma does not hold for arbitrary positive definite matrix W.

Interior Point Methods:

- are well-suited to large-scale optimization
- can take advantage of the parallelism

Large problems are "structured":

- partial separability
- spatial distribution
- dynamics
- uncertainty
- etc.

Object-Oriented Parallel Solver (OOPS)

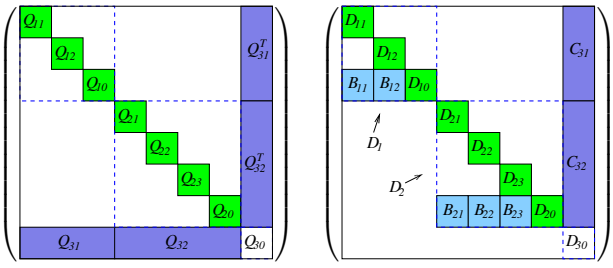
- Exploits structure
- Runs in parallel
- Solves problems with millions of variables

Andreas Grothey will talk about OOPS.

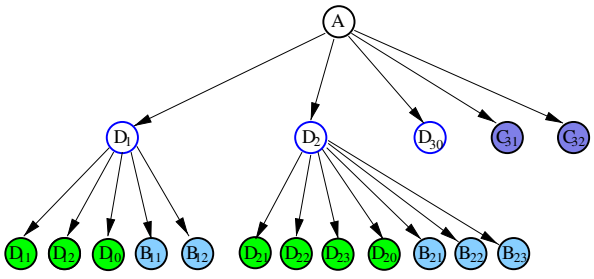
Gondzio & Sarkissian:
Math Prog 96 (2003) 561-584.
Gondzio & Grothey:
SIOPT 13 (2003) 842-864.

Tree Description of Block-Structures

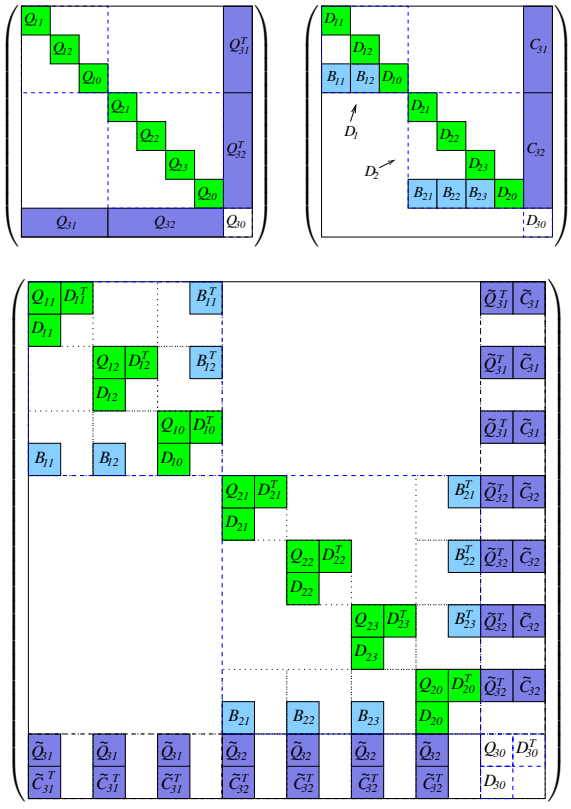
Structured Matrix:



Associated Tree:



Reordered Augmented Matrix



Cholesky factors sometimes get hopelessly dense.

QAP (Quadratic Assignment Problems) and NUG problems (dual QAPs)

Prob	Dimensions		
	rows	columns	nonzeros
qap12	3192	8856	38304
qap15	6330	22275	94950
nug12	3192	8856	38304
nug15	6330	22275	94950

Normal Equations:

Prob	nz(AAt)	nz(LLt)	Flops
qap12	74592	2135388	2.378e+9
qap15	186075	8191638	1.792e+10
nug12	74592	2789960	4.014e+9
nug15	186075	11047639	3.240e+10

Augmented System:

Prob	nz(A)	nz(LLt)	Flops
qap12	38304	1969957	2.046e+9
qap15	94950	7374972	1.522e+10
nug12	38304	1969957	2.046e+9
nug15	94950	7374972	1.522e+10

- Unavoidable Ill-conditioning:
  - benign in direct approach;
  - challenge for iterative approach.
- Positive Definite vs Indefinite Systems
- Preconditioners for Structured Matrices
- Preconditioners for Indefinite System
  - Motivation
    - Sparsity Issues
    - Numerical Properties
  - Spectral Analysis
  - Influence of Regularizations
- Conclusions
- What's to Come in IPMs

Iterative Methods

Normal Equations or **Augmented System**:

- NE is positive definite:
  can use conjugate gradients;
- AS is indefinite:
  can use BiCGSTAB, GMRES, QMR;

AS is generally more flexible.
Oliveira (1997) PhD Thesis, Rice Univ.
Oliveira & Sorensen (1997) TR, Rice Univ.
→ It is better to precondition AS.

O, OS show that all preconditioners for the NE have an equivalent for the AS while the opposite is not true.

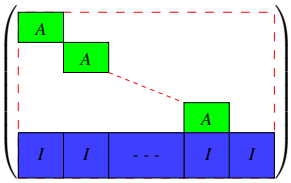
After all, NE is equivalent to a restricted order of pivoting in AS.

Iterative Methods

Many attempts (LP, QP, NLP and PDE):

- Gill, Murray, Ponceleon, Saunders
  SIMAX 13 (1992) 292-311.
- Lukšan & Vlček
  NLAA 5 (1998) 219-247.
- Golub & Wathen
  SISC 19 (1998) 530-539.
- Murphy, Golub & Wathen
  SISC 21 (2000) 1969-1972.
- Keller, Gould & Wathen
  SIMAX 21 (2000) 1300-1317.
- Perugia & Simoncini
  NLAA 7 (2000) 585-616.
- Castro
  SIOPT 10 (2000) 852-877.
- Gould, Hribal & Nocedal
  SISC 23 (2001) 1376-1395.
- Durazzi & Ruggiero
  NLAA (to appear).
- Rozložník & Simoncini
  SIMAX 24 (2002) 368-391.

Castro *SIOPT* 10 (2000) 852-877.



Normal-equations matrix

$$\begin{bmatrix} A_1A_1^T & & & A_1B_1^T \\ & A_2A_2^T & & A_2B_2^T \\ & & \ddots & \vdots \\ & & A_nA_n^T & A_nB_n^T \\ B_1A_1^T & B_2A_2^T & \cdots & B_nA_n^T \quad \sum_{i=1}^{n+1} B_iB_i^T \end{bmatrix} = \begin{bmatrix} E & B^T \\ B & F \end{bmatrix},$$

where E and F are positive definite.

E is easily invertible (block-diagonal).
The inverse of Schur complement matrix $F-BE^{-1}B^T$ can be written as the power series:

$$(F - BE^{-1}B^T)^{-1} = \sum_{i=0}^{\infty} (F^{-1}BE^{-1}B^T)^i F^{-1}.$$

Finite approximation of the series:
→ Very efficient preconditioner.

Murphy, Golub & Wathen
SISC 21 (2000) 1969-1972.

Consider a matrix

$$H = \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix},$$

where
$Q \in \mathcal{R}^{n\times n}$ is positive definite, and
$A \in \mathcal{R}^{m\times n}$ has full row rank.

Consider the preconditioner which incorporates an exact Schur complement $AQ^{-1}A^T$.
For example:

$$P_1 = \begin{bmatrix} Q & 0 \\ 0 & AQ^{-1}A^T \end{bmatrix} \quad \text{or} \quad P_2 = \begin{bmatrix} Q & A^T \\ 0 & AQ^{-1}A^T \end{bmatrix}.$$

The preconditioned matrices $P^{-1}H$ have only two or three distinct eigenvalues.

MGW conclude:
"The approximations of the Schur complement lead to preconditioners which can be very effective even though they are in no sense approximate inverses".

CG with Indefinite Preconditioner

Consider the indefinite matrix

$$H = \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix},$$

where
$Q \in \mathcal{R}^{n\times n}$ is positive definite, and
$A \in \mathcal{R}^{m\times n}$ has full row rank.
The CG method may fail when applied to an indefinite system.

Rozložník & Simoncini
SIMAX 24 (2002) 368-391.

RS consider the preconditioner P which guarantees that all eigenvalues of the preconditioned matrix $P^{-1}H$ are positive and bounded away from zero.

Although $P^{-1}H$ is indefinite

- the CG can be applied to this problem,
- the asymptotic rate of convergence of CG is approximately the same as that obtained for a positive definite matrix with the same eigenvalues as the original system.

Indefinite Block Preconditioner

Consider again the matrix

$$H = \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix},$$

where
$Q \in \mathcal{R}^{n\times n}$ is positive definite, and
$A \in \mathcal{R}^{m\times n}$ has full row rank.

Consider a preconditioner of the form:

$$P = \begin{bmatrix} D & A^T \\ A & 0 \end{bmatrix},$$

where $D \in \mathcal{R}^{n\times n}$ is positive definite.

Keller, Gould & Wathen
SIMAX 21 (2000) 1300-1317.

Theorem. *Assume that A has rank m ($m < n$). Then, $P^{-1}H$ has at least $2m$ unit eigenvalues, and the other eigenvalues are positive and satisfy*

$$\lambda_{min}(D^{-1}Q) \;\le\; \lambda \;\le\; \lambda_{max}(D^{-1}Q).$$

Proof: The preconditioned matrix (left) reads

$$P^{-1}H = \begin{bmatrix} D & A^T \\ A & 0 \end{bmatrix}^{-1} \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} =$$
$$= \begin{bmatrix} D^{-1} - D^{-1}A^TM^{-1}AD^{-1} & D^{-1}A^TM^{-1} \\ M^{-1}AD^{-1} & -M^{-1} \end{bmatrix} \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix}$$
$$= \begin{bmatrix} D^{-1}Q - D^{-1}A^TM^{-1}AU & 0 \\ M^{-1}AU & I_m \end{bmatrix} = \begin{bmatrix} X & 0 \\ Y & I_m \end{bmatrix},$$

where $M = AD^{-1}A^T$, $U = D^{-1}Q - I$.

$P^{-1}H$ has m linearly independent eigenvectors associated with the eigenvalue $\lambda = 1$ since for $w_i \in \mathcal{R}^m$

$$P^{-1}H \begin{bmatrix} 0 \\ w_i \end{bmatrix} = \begin{bmatrix} 0 \\ w_i \end{bmatrix}.$$

The remaining n eigenvectors are the same as those of the matrix $X = D^{-1}Q - D^{-1}A^TM^{-1}AU$.

Matrix X has at least m other unit eigenvalues. Indeed, for any $x \in \mathcal{R}^m$ we write

$$X^TA^Tx = (I + U^T(I - A^TM^{-1}AD^{-1}))A^Tx =$$
$$= A^Tx + U^T(A^Tx - A^Tx) = A^Tx.$$

The remaining n–m eigenvalues and eigenvectors of $P^{-1}H$ have to satisfy

$$\begin{array}{lcl} Qx + A^Ty = \lambda Dx + \lambda A^Ty \\ Ax = \lambda Ax. \end{array}$$

If $\lambda \neq 1$ the second equation yields $Ax = 0$. Let us multiply the first equation by x^T. Recalling that $x^TA^T = 0$ we obtain

$$x^TQx = \lambda x^TDx, \quad \Rightarrow \quad \lambda = \frac{x^TQx}{x^TDx} = q(D^{-1}Q).$$

The last expression is the Rayleigh quotient of the generalized eigenproblem $Dv = \mu Qv$. Since both D and Q are positive definite we have for every $x \in \mathcal{R}^n$

$$0 < \lambda_{\min}(D^{-1}Q) \leq \frac{x^TQx}{x^TDx} \leq \lambda_{\max}(D^{-1}Q)$$

and finally

$$\lambda_{\min}(D^{-1}Q) \leq \lambda \leq \lambda_{\max}(D^{-1}Q).$$

Conclusion:
The preconditioner satisfies the requirements of **Rozložník & Simoncini**.

How to choose D?

Bergamaschi, Gondzio & Zilli,
Preconditioning indefinite systems in interior point methods for optimization,
Tech. Rep. MS-02-02.

Augmented system in **QP, NLP**

$$H = \begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix}.$$

Drop off-diagonal elements from Q:
Replace

$$-Q - \Theta^{-1}$$

with

$$D = -diag(Q) - \Theta^{-1}.$$

Preconditioners: Motivation

Sparsity issues: irreducible blocks in QP.

Consider the matrices

$$Q = \begin{bmatrix} \mathbf{x} & \mathbf{x} & & & \\ \mathbf{x} & \mathbf{x} & & & \\ & & x & & \\ & & & x & \\ & & & & x \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} x & & & & x \\ x & & x & & \\ & x & x & & x \\ & x & & x & \end{bmatrix},$$

giving

$$H = \begin{bmatrix} \mathbf{x} & \mathbf{x} & & & & x & x & & \\ \mathbf{x} & \mathbf{x} & & & & & & x & x \\ & & x & & & x & x & & \\ & & & x & & & & & x \\ & & & & x & x & & x & \\ x & & & & x & & & & \\ x & & x & & & & & & \\ & x & x & & x & & & & \\ & x & & x & & & & & \end{bmatrix}.$$

If the elimination starts from h_{11} or h_{22}, then

$$H = \begin{bmatrix} \mathbf{x} & \mathbf{x} & & & & x & x & \mathbf{f} & \mathbf{f} \\ \mathbf{x} & \mathbf{x} & & & & \mathbf{f} & \mathbf{f} & x & x \\ & & x & & & & x & x & \\ & & & x & & & & & x \\ & & & & x & x & & x & \\ x & \mathbf{f} & & & x & & & & \\ x & \mathbf{f} & x & & & & & & \\ \mathbf{f} & x & x & & x & & & & \\ \mathbf{f} & x & & x & & & & & \end{bmatrix}.$$

Conclusion:
Drop off-diagonal elements form Q.

D is a diagonal matrix
→ Free choice between NE and AS.

Preconditioner 1
Compute the Cholesky-like factorization.

$$P_1 = \begin{bmatrix} D & A^T \\ A & 0 \end{bmatrix} = L\bar{D}L^T.$$

Preconditioner 2
Reduce the system to Normal Equations $AD^{-1}A^T$, compute the Cholesky factorization

$$AD^{-1}A^T = L_0 D_0 L_0^T,$$

and use:

$$P_2 = \begin{bmatrix} D & A^T \\ A & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ AD^{-1} & L_0 \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & -D_0 \end{bmatrix} \begin{bmatrix} I & D^{-1}A^T \\ 0 & L_0^T \end{bmatrix}.$$

The **regularization**

$$\bar{H}_R = \begin{bmatrix} -Q & A^T \\ A & 0 \end{bmatrix} + \begin{bmatrix} -R_p & 0 \\ 0 & R_d \end{bmatrix},$$

changes the **eigenvalues** of the preconditioned matrix:

without the regularization:

$$\lambda(P^{-1}H) = \frac{x^T Q x}{x^T D x}$$

with the regularization:

$$\lambda(P_R^{-1}H_R) = \frac{-x^T Q x + \delta}{-x^T D x + \delta},$$

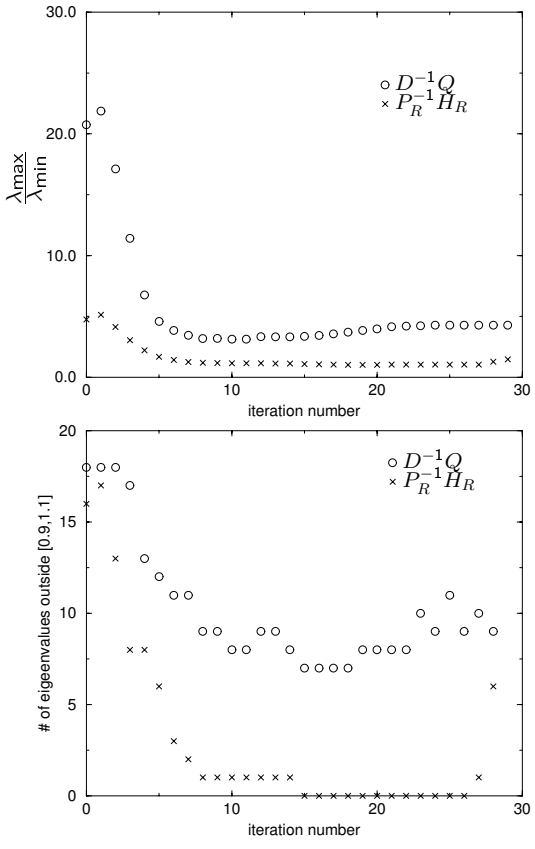
where $\delta = x^T R_p\, x + y^T R_d\, y > 0$.

For any $\alpha, \beta, t > 0$, the function $h(t) = \frac{\alpha+t}{\beta+t}$ is *increasing* if $\frac{\alpha}{\beta} \leq 1$, and *decreasing* if $\frac{\alpha}{\beta} > 1$.

Hence:
if $\lambda(P^{-1}H) < 1$, then $\lambda(P_R^{-1}H_R) > \lambda(P^{-1}H)$.
if $\lambda(P^{-1}H) > 1$, then $\lambda(P_R^{-1}H_R) < \lambda(P^{-1}H)$.

The use of regularization improves the **clustering** of eigenvalues.

Influence of Regularization: `q25fv47`



HOPDM: Direct vs Iterative Methods

Problem	Dimensions		nonzeros(L)		
	nz(A)	nz(Q)	Direct	AS-Prec	NE-Prec
cvxqp1_m	1498	2984	75973	4739	4768
cvxqp1_l	14998	29984	3725045	71833	89241
cvxqp2_m	749	2984	51923	1031	315
cvxqp2_l	7499	29984	2754141	10579	3379
cvxqp3_m	2247	2984	90433	9527	14018
cvxqp3_l	22497	29984	4291057	149488	271780

QMR: Freund & Nachtigal (1991,1994).
QMR is asked for 10^{-3} accuracy.
500 MHz Pentium III PC, Linux, 256 MB.

Problem	Direct		AS-Prec		NE-Prec	
	its	time	its	time	its	time
cvxqp1_m	9	2.35	11	1.59	11	1.64
cvxqp1_l	11	1267.53	13	32.51	13	38.50
cvxqp2_m	9	1.27	10	1.01	10	1.06
cvxqp2_l	8	547.91	10	17.87	10	18.10
cvxqp3_m	9	3.40	11	1.94	11	2.37
cvxqp3_l	8	958.59	10	42.03	10	57.12

QMR: Freund & Nachtigal (1991,1994).
QMR is asked for 10^{-3} accuracy.
500 MHz Pentium III PC, Linux, 256 MB.

Problem	AS-Prec				NE-Prec			
	IPM	ItS|	Max	Avr	IPM	ItS|	Max	Avr
cvxqp1_m	11	338	20	14	11	338	20	14
cvxqp1_l	13	481	20	17	13	487	20	17
cvxqp2_m	10	307	20	13	10	307	20	13
cvxqp2_l	10	389	20	18	10	389	20	18
cvxqp3_m	11	303	20	13	11	296	20	12
cvxqp3_l	10	415	20	19	10	374	20	17

NL iterations (QMR) in the last IPM iteration:

Problem	AS-Prec		NE-Prec	
	Predictor	Corrector	Predictor	Corrector
cvxqp1_m	11	11	11	11
cvxqp1_l	16	12	14	12
cvxqp2_m	9	1	9	1
cvxqp2_l	18	16	18	16
cvxqp3_m	10	7	9	7
cvxqp3_l	20	14	15	13

GMRES: Saad & Schultz (1986).
BiCGSTAB: Van der Vorst (1992).
QMR: Freund & Nachtigal (1991,1994).

All approaches iterate until 10^{-3} accuracy is reached but perform no more than 20 iterations.
All approaches use the AS preconditioner.

500 MHz Pentium III PC, Linux, 256 MB.

Problem	GMRES			BiCGSTAB			QMR		
	IP	ItS|	time	IP	ItS|	time	IP	ItS|	time
cvxqp1_s	12	177	0.1	12	137	0.1	9	189	0.1
cvxqp1_m	11	307	1.3	11	233	1.3	11	338	1.6
cvxqp1_l	13	503	24.9	13	357	28.9	13	481	32.5
cvxqp2_s	27	217	0.1	16	153	0.1	10	235	0.1
cvxqp2_m	16	270	0.9	21	221	1.1	10	307	1.0
cvxqp2_l	19	404	16.1	10	243	14.9	10	389	18.1
cvxqp3_s	11	162	0.1	11	114	0.1	11	181	0.1
cvxqp3_m	11	306	1.6	11	226	1.7	11	303	1.9
cvxqp3_l	10	375	28.8	10	272	32.5	10	415	42.0

Conclusions

Direct Methods are reliable but occasionally excessively expensive.

Iterative Methods are promising but:

- are sometimes unpredictable;
- need tuning;
- depend upon preconditioners.

An **Augmented System** offers more freedom

- when used in the direct approach,
- when used to compute the preconditioners for the iterative approach.

Regularization is helpful.

What's to come in IPMs?

Direct Methods:

- small improvements:
  - reordering strategies
  - implementation (cache, supernodes)
- exploiting structure in huge problems (implicit inverse representations)

Iterative Methods:

- new preconditioners

Challenge:

Find an inverse representation with the number of nonzeros comparable to that of $\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix}$.

Spectral Methods for Discontinuous Problems

David Gottlieb[1] & Sigal Gottlieb[2]
Division of Applied Mathematics,
Brown University,
Providence RI, USA
dig@mail.cfm.brown.edu
&
University of Massachusetts
Dartmouth,
North Dartmouth MA. USA

1 Introduction

Spectral methods have emerged as powerful computational techniques for simulation of complex, smooth physical phenomena. Among other applications they have contributed to our understanding of turbulence by successfully simulating incompressible turbulent flows, have been extensively used in meteorology and geophysics, and have been recently applied to time domain electromagnetics. Several issues arise when applying spectral methods to problems which feature sharp gradients and discontinuities. In the presence of such phenomena the accuracy of high order methods deteriorates. This is due to the well known Gibbs phenomenon that states that the pointwise convergence of global approximations of discontinuous functions is at most first order. In the presence of a shock wave global approximations are oscillatory and converge nonuniformly. Recent advances in the theory and application of spectral methods indicate that high order information is retained in stable spectral simulations of discontinuous phenomena and can be recovered by suitable postprocessing techniques.

2 The Gibbs Phenomenon

The partial Fourier sum

$$\sum_{k=-N}^{N} \hat{f}_k e^{\pi i k x}$$

based on the first $2N+1$ Fourier coefficients of a nonsmooth function $f(x)$, converges slowly away from the discontinuity and features non-decaying oscillations. This behavior of all global approximations of nonsmooth functions is known as the classical Gibbs phenomenon (see [20]).

Euler (in 1755) was probably the first to witness the phenomenon in the Fourier expansion of the function $f(x) = x$. Wilbraham (in 1848) analyzed this series and discussed the oscillatory behavior of the approximation. However, the first to address the issue of reconstructing a discontinuous function from its partial series was the eminent physicist

[1]Work supported by AFOSR grant F49620-02-1-0113 and DOE grant DE-FG02-98ER25346
[2]Work supported by NSF grant no. DMS 0106743

Albert Michelson. In 1898 Michelson, together with Stratton, built a harmonic analyzer, a mechanical device which stored the Fourier coefficients of a given curve. A detailed description of this device is given in the 1910 edition of the Britannica. In their Philosophical magazine paper (1898) they presented graphs of functions reconstructed from their Fourier coefficients. One of those curves was a square wave that displayed the Gibbs oscillations. Shortly after the appearance of this paper, Michelson, in a letter to the British journal NATURE (October 6 1898), pointed out the difficulty in reconstructing the function $f(x) = x$ from its Fourier coefficients. In particular Michelson claimed that looking at the point $x_0 = 1 - \frac{1}{N}$ no convergence is observed. Michelson states:

> The idea that a real discontinuity can replace a sum of continuous curves is so utterly at variance with the physicists' notion of quantity that it seems to me to be worth while giving a very elementary statement of the problem in such simple form that the mathematicians can at once point to the inconsistency if any there be.

This appeal to mathematicians was answered by the mathematician A.E.H. Love in the next issue of NATURE (December 1898), though probably not in the spirit that Michelson intended. Love claimed that the partial sum does converges pointwise and for this type of convergence, the point has to be unchanged in the limit process. Insultingly, Love recommended to Michelson and his physicist friends some elementary books in mathematics. Love, of course, overlooked the concept of uniform convergence which is at the heart of this issue. This seems to be the gist of Michelson's terse response in the December 29, 1898 issue of Nature that restated his original example. The same issue contains a letter from Gibbs beginning with the observation that in Love's response "the point of view of Professor Michelson is hardly considered" and attempting to analyze the phenomenon. However Gibbs seems to argue that there was indeed uniform convergence. The same issue contains also the (now more respectful) response of Love to Gibbs' letter which now admits the importance of the issue of uniform convergence.

It was not until the April 27, 1899 issue of Nature that Gibbs presented the correct analysis of this phenomenon. It was Poincare's eminence that ultimately decided this issue in a letter (forwarded by Michelson to Nature and published in May 1899) stating the correct behavior of the partial Fourier sums.

Sufficient conditions for the removal of the Gibbs phenomenon were given in [21]. Consider a function $f(x) \in L^2[-1,1]$ and assume that there is a subinterval $[a,b] \subset [-1,1]$ in which $f(x)$ is analytic. (For convenience we define the local variable, $\xi = -1 + 2\frac{x-a}{b-a}$ such that if $a \le x \le b$ then $-1 \le \xi \le 1$.) Let the family $\{\Psi_k(x)\}$, be orthonormal under a scalar product $(\,\cdot\,,\,\cdot\,)$, and denote the finite expansion of $f(x)$ in this basis by $f_N(x)$,

$$f_N(x) \quad = \quad \sum_{k=0}^{N} (f, \Psi_k)\Psi_k(x).$$

Let the family $\{\Phi_k^\lambda(\xi)\}$ be Gibbs complementary to the family $\{\Psi_k(x)\}$ (see [21] for its exact definition), then the postprocessed reconstruction given by

$$g_N(x) = \sum_{l=0}^{\lambda} \langle f_N, \Phi_l^\lambda \rangle_\lambda \Phi_l^\lambda(\xi(x))$$

converges exponentially to $f(x)$, i.e.

$$\max_{a\le x\le b}|f(x)-g_N(x)| \;\le\; e^{-qN}, \qquad q>0.$$

In a series of papers [15]–[19] we showed that the Gegenbauer polynomials

$$\Phi_k^\lambda(\xi) \;=\; \frac{1}{\sqrt{h_k^\lambda}}C_k^\lambda(\xi)$$

which are orthonormal under the inner product $\langle\cdot,\cdot\rangle_\lambda$ defined by

$$\langle f,g\rangle_\lambda \;=\; \int_{-1}^{1}(1-\xi^2)^{\lambda-\frac{1}{2}}f(\xi)g(\xi)\,d\xi$$

are Gibbs complementary to all commonly used spectral approximations.
The theory presented above does not prescribe an optimal way of constructing a Gibbs complementary basis. This is still an open question. The Gegenbauer method is not robust, it is sensitive to roundoff errors and to the choice of the parameters λ and m. A different implementation of the Gegenbauer postprocessing method has been suggested recently by Jung and Shizgal [25]. To explain the differences between the direct Gegenbauer method and the inverse Gegenbauer method suggested in [25], consider the case of the the Fourier expansion of a nonperiodic problem. The Fourier approximation $f_N(x)$ of $f(x)$ is

$$f_N(x) \;=\; \sum_{k=-N}^{N}\hat{f}_k e^{ik\pi x},$$

where $\hat{f}_k=\big(f(x),e^{ik\pi x}\big)$, and we construct

$$f_N^m(x) \;=\; \sum_{l=0}^{m}\hat{g}_l C_l^\lambda(x),$$

where $\hat{g}_l=\langle f_N,C_l^\lambda(x)\rangle_\lambda$. In the Inverse method we use the relationship

$$\hat{f}_k=(f_N^m(x),e^{ik\pi x})$$

and invert to find $\hat{g}_l$.
Thus if we define the matrix $W_{kl}=(C_l^\lambda(x),e^{ik\pi x})_F=\int_{-1}^{1}(1-x^2)^{\lambda-\frac{1}{2}}C_l^\lambda(x)e^{ik\pi x}dx$, and $\hat{f}_k=(f,e^{ik\pi x})$

$$\sum_{l=0}^{m}W_{kl}g_l^\lambda=\hat{f}_k.$$

The method seems to be less sensitive to roundoff errors or to the choice of parameters. In particular if the original function is a polynomial, the inverse method is exact.

3 Recovering Order of Accuracy for Solutions of PDEs

Just as spectral accuracy can be recovered in spectral approximations of nonsmooth functions, it can also be recovered in the case of discontinuous solutions of linear hyperbolic equations.

Consider the hyperbolic system of the form

$$\frac{\partial U}{\partial t} = \mathcal{L}U; \;\; U(t=0) = U_0.$$

Let u be the spectral approximation to U. For smooth solutions we have the classical error estimate:

$$||U - u|| \leq K||U_0||_s \frac{1}{N^{s-1}}.$$

This estimate, obviously, requires the initial condition to be smooth everywhere and does not apply in the case of piecewise smooth initial conditions. However it has been proven in [1] that :

$$|(U(T) - u(T), \phi)| \leq K||\phi||_s \frac{1}{N^s}$$

for any smooth function ϕ provided that the numerical initial conditions are preprocessed. Equation (3) implies that the Fourier coefficients of u approximate those of U with spectral accuracy. It is therefore possible to postprocess to get spectral accuracy for the point values in any interval where the solution U is smooth. In the case of the nonlinear hyperbolic system

$$\frac{\partial U}{\partial t} + \frac{\partial f(U)}{\partial x} = 0,$$

Lax ([23]) argued that high order information is contained in a convergent high resolution scheme.

To stabilize the spectral schemes we use either the Spectral Viscosity (SV) Method or the Super Spectral Viscosity (SSV) Method (see [27],[28],[24]). These methods amount to the addition of viscosity terms for the high modes of the solution. In [10] we demonstrate that this is equivalent to filtering, and in fact, filtering can be seen as an efficient way of applying these methods.

The theory developed by Tadmor and Tadmor and Maday demonstrates that both the SV and SSV methods converge to the correct entropy solution for Fourier and Legendre approximations to scalar nonlinear hyperbolic equations. Carpenter, Gottlieb and Shu [2] proved that even for systems, if the solution converges it converges to the correct entropy solution.

4 Applications

There is extensive literature reporting results of the application of spectral methods to shock wave problems [12],[13],[27],[28][24].

In [4] the authors compared ENO and spectral methods for the numerical simulations of shock- cylinder interactions in the case of reactive flows. The authors demonstrated

that spectral methods required fewer resources than the ENO schemes for comparable accuracy. In [3], the author considered interactions of shock waves and entropy waves as well as interactions of shock waves and vortices. The calculations involved solutions of the two dimensional Euler equations and the results compared well with ENO methods. In [10] [9] we presented spectral simulations of the Richtmyer-Meshkov instabilities and showed comparisons with WENO schemes. It has been shown that with increased order of accuracy the WENO results converge to the spectral ones.

A more extensive study of spectral simulations of compressible reactive high Mach number flows has been reported in [6]. In this work the interaction of shock waves and hydrogen jets were studied. This involves the solution of the Navier Stokes equations with chemical interactions. The work gives a clear demonstration of the fact that spectral methods are very suitable for studies of complicated flows that involve shock waves.

Several model problems confirm Lax's argument concerning the information contained in high resolution schemes. Shu and Wang [26] recovered spectral accuracy for the nonlinear Burgers equation where discontinuity develops and moves around the domain. It is interesting to note that the Gegenbauer postprocessing technique recovers the design order accuracy even for finite difference schemes. In [22] the postprocessing recovered design accuracy, in the maximum norm, for the WENO steady state solution of a converging–diverging nozzle.

References

[1] S. Abarbanel, D. Gottlieb and E. Tadmor, *Spectral methods for discontinuous problems*, in "Numerical Methods for Fluid Dynamics II", ed. by K.W. Morton and M.J. Baines, Oxford University Press, (1986), pp.128-153.

[2] M. Carpenter, D. Gottlieb and C.W. Shu, *On the Conservation and Convergence to Weak Solutions of Global Schemes*, to appear in JSC.

[3] W. S. Don, *Numerical Study of Pseudospectral Methods in Shock Wave Applications*, Journal of Computational Physics, **110**, pp. 103-111, 1994.

[4] W. S. Don, C. Quillen, *Numerical simulation of Reactive Flow, Part I : Resolution*, Journal of Computational Physics **122**, pp. 244-265, 1995.

[5] W. S. Don & D. Gottlieb, *High Order Methods for Complicated Flows Interacting with Shock Waves*, AIAA 97-0538.

[6] W. S. Don & D. Gottlieb, *Spectral Simulations of Supersonic Reactive Flows*, SIAM J. Numer. Anal., **35**, No. 6, pp. 2370-2384, Dec. 1998.

[7] W. S. Don, D. Gottlieb & J. H. Jung, *A multi-domain spectral method for supersonic reactive flows*, Journal of Computational Physics, submitted for publication.

[8] W. S. Don, D. Gottlieb & J. H. Jung, *Multi-domain spectral method approach to supersonic combustion of recessed cavity flame-holders*, JANNAF 38th Combustion, 26th Airbreathing Propulsion, 20th Propulsion Systems Hazards, and 2nd Modeling and Simulation Subcommittees Joint Meeting, held in Destin, FL 8-12 April 2002.

[9] W. S. Don, D. Gottlieb, & C. W. Shu, *High Order Numerical Methods for the Two Dimensional Richtmyer-Meshkov Instability, Part I.*, Conference proceeding for the International

Workshop for the Physics of Compressible Turbulence Mixing, Laser and Particle Beams, to appear.

[10] S. Gottlieb and D. Gottlieb *High Order Methods for Reactive Compressible Flows* to appear in the proceeding of Euromech 446.

[11] D. Gottlieb, M. Y. Hussaini and S. A. Orszag, *Introduction: Theory and Applications of Spectral Methods*, in Spectral Methods for Partial Differential Equations, R. Voigt, D. Gottlieb and M.Y. Hussaini, ed. SIAM, Philadelphia, 1984. pp. 1-54.

[12] D. Gottlieb, L. Lustman and S.A. Orszag,*Spectral calculations of one dimensional inviscid flows with shocks*, SIAM J. Sci. Stat. Comput.,**2**, 296–310 (1981).

[13] D. Gottlieb L. Lustman and C.L. Street, *Spectral Methods for Two Dimensional Flows*, in Spectral Methods for PDEs, (SIAM, Philadelphia 1984).

[14] D. Gottlieb, S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS Conference Series in Applied Mathematics **26**, SIAM, (1977).

[15] D. Gottlieb, C. W. Shu, A. Solomonoff, H. Vandeven, *On the Gibbs phenomenon I: recovering exponential accuracy from the Fourier partial sum of a non-periodic analytic function*, J. Comp. and Appl. Math. **43**, pp. 81–98 (1992).

[16] D. Gottlieb, C. W. Shu, *Resolution properties of the Fourier method for discontinuous waves*, Computer Meth. in Appl. Mech. and Eng. **116**, pp. 27–37 (1994).

[17] D. Gottlieb, C. W. Shu, *On the Gibbs phenomenon III: recovering exponential accuracy in a sub-interval from a spectral partial sum of a piecewise in a sub-interval from a spectral partial sum of a piecewise analytic function*, SIAM. J.Numer. Anal. **33**, pp. 280-290 (1996).

[18] D. Gottlieb, C. W. Shu, *On the Gibbs phenomenon IV: recovering exponential accuracy in a sub-interval from a Gegenbauer partial sum of a piecewise analytic function*, Math. Comp. **64**, pp. 1081-1095 (1995).

[19] D. Gottlieb, C. W. Shu, *On the Gibbs phenomenon V: recovering exponential accuracy from collocation point values of a piecewise analytic function*, Numerische Mathematik **71**, pp. 511–526 (1995).

[20] D. Gottlieb, C. W. Shu, *The Gibbs phenomenon and its resolution*, SIAM Review, 39, pp.644–668 (1997).

[21] D. Gottlieb and C.W. Shu, *General theory for the resolution of the Gibbs phenomenon*, Accademia Nazionale Dei Lincey, ATTI Dei Convegni Lincey 147, pp. 39–48 (1998).

[22] S. Gottlieb, D. Gottlieb and C.W. Shu, to appear.

[23] P. D. Lax, *Accuracy and resolution in the computation of solutions of linear and nonlinear equations*, in *Recent advances in Numerical Analysis, Proc. Symp.*, Mathematical Research Center, University of Wisconsin, Academic Press, pp. 107-117 (1978).

[24] H. Ma, *Chebyshev–Legendre Super Spectral Viscosity Method for Nonlinear Conservation Laws*, submitted to SIAM J. Numer. Anal.

[25] B.D. Shizgal and J.H. Jung, *On the resolution of the Gibbs phenomenon*, to appear.

[26] C.-W. Shu and P. Wong, *A note on the accuracy of spectral methods applied to nonlinear conservation laws*, Journal of Scientific Computing, v10 (1995), pp.357-369.

[27] E. Tadmor, *Convergence of spectral methods for nonlinear conservation laws*, SIAM J. Numer. Anal., **26**, pp. 30–44 (1989).

[28] E. Tadmor, *Shock capturing by the spectral viscosity method*, Proceedings of ICOSAHOM 89, Elsevier Science Publishers B. V., North Holland, IMACS (1989).

The filter-idea and its application to the nonlinear feasibility problem

Nick Gould & Philippe L. Toint
Computational Science and Engineering Department
Rutherford Appleton Laboratory,
Chilton, Oxfordshire, England
gould@rl.ac.uk
&
Department of Mathematics,
University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium,
philippe.toint@fundp.ac.be

1 Introduction

We consider the solution of the general smooth feasibility problem, that is the problem of finding a vector $x \in \mathbb{R}^n$ such that

$$c_{\mathcal{E}}(x) = 0, \tag{1.1}$$

and

$$c_{\mathcal{I}}(x) \geq 0, \tag{1.2}$$

where $c_{\mathcal{E}}(x)$ and $c_{\mathcal{I}}(x)$ are smooth functions from $\mathbb{R}^n$ into $\mathbb{R}^m$ and $\mathbb{R}^q$, respectively. If such a point cannot be found, it is desirable to find a local minimizer of the constraint violations. We choose here to consider the Eucliden norm of these violations, that is to find a local minimizer of the function

$$\min_x \tfrac{1}{2}\|\theta(x)\|^2, \tag{1.3}$$

where we define

$$\theta(x) \stackrel{\rm def}{=} \begin{pmatrix} c_{\mathcal{E}}(x) \\ [c_{\mathcal{I}}(x)]_- \end{pmatrix} \in \mathbb{R}^p, \tag{1.4}$$

with $\|\cdot\|$ denoting the Euclidean norm, with $p = m + q$ and $[c_{\mathcal{I}}(x)]_- = \min[0, c_{\mathcal{I}}(x)]$, the minimum being taken componentwise. An important special case of this problem is when $q = 0, m = n$ and thus $\mathcal{E} = \{1, \ldots, n\}$, which gives systems of smooth nonlinear equations. The problem under consideration is therefore not only fairly general, but also practically important because a large number of applications can be cast in this form. Moreover, solving the feasibility problem may also occur as a subproblem in practically more complicated contexts, such as the the "restoration" phase in the solution of the nonlinear programming problem using filter methods (see [6], [5], [8], [9] or [7], amongst others).

The method of choice for solving (1.1)–(1.2) or (1.3) is Newton's method, because of its fast convergence properties. However, as is well-known, Newton's method must be

safeguarded to ensure that it converges to a solution even from starting points that are far from the solution, a feature that is not automatic otherwise. Various safeguarding techniques are known, including the use of linesearches (see [16], [3], [17], [18], ...) or trust regions (see [14], [15], or Chapter 16 of [2]). More recently, [11] and [10] have proposed a method that combines the basic trust-region mechanism with filter techniques: not only did they prove global convergence for the algorithm, but they also reported very encouraging numerical experience. The objective of this note is to outline this method.

2 The filter algorithm and its algorithmic options

2.1 The objective function, its models and the step

We consider an algorithm which aims at minimizing

$$f(x) = \tfrac{1}{2}\|\theta(x)\|^2.$$

For simplicity of exposition, we assume that the problem only contains nonlinear equations ($q = 0$). In this case, we may build two distinct local quadratic models of $f(x)$ in the neighbourhood of a given iterate x_k. The first is the Gauss-Newton model, and is given by

$$m_k^{\rm GN}(x_k + s) = \tfrac{1}{2}\|c_{\mathcal{E}}(x_k) + J_{\mathcal{E}}(x_k)s\|^2, \tag{2.1}$$

where $J_{\mathcal{E}}(x_k)$ is the Jacobian of $c_{\mathcal{E}}(x)$ at x_k. The second is the full second-order Newton model

$$m_k^{\rm N}(x_k + s) = m_k^{\rm GN}(x_k + s) + \tfrac{1}{2}\sum_{j\in\mathcal{E}} c_j(x_k)\langle s, \nabla^2 c_j(x_k)s\rangle, \tag{2.2}$$

which includes an additional term involving the curvature of the equality constraints.

In our method, we have chosen to compute the step s_k by minimizing one of these models in some region surrounding the current iterate x_k, defined by the constraint

$$\|s_k\|_k \leq \tau_k\Delta_k, \tag{2.3}$$

where Δ_k is a trust-region radius which is updated in the usual trust-region manner (see Chapters 6 and 17 of [2], for instance). The parameter $\tau_k \geq 1$ allows for steps that potentially extend much beyond the limit of the trust region itself, in the case where convergence seems satisfactory. The precise mechanism for determining τ_k will be discussed in more detail below. The $\|\cdot\|_k$ norm appearing in (2.3) is a preconditioned Euclidean norm, that is $\|s\|_k^2 = \langle s, P_k^{-1}s\rangle$, where P_k is a symmetric positive-definite preconditioning matrix that is used at the k-th iteration. The solution of the subproblem of minimizing $m_k^{\rm GN}(x_k+s)$ or $m_k^{\rm N}(x_k+s)$ subject to (2.3) is computed approximately using the Generalized Lanczos Trust-Region (GLTR) method of [12] as implemented in the GLTR module of GALAHAD (see [13]). Besides using $P_k = I$ (i.e. no preconditioning at all), our package, named FILTRANE, can also be instructed to use a diagonal preconditioning which is obtained by extracting the diagonal of the matrix $H_k \stackrel{\rm def}{=} J_{\mathcal{E}}(x_k)J_{\mathcal{E}}(x_k)^T$, or a banded preconditioning matrix of semi-bandwidth 5 obtained by extracting the corresponding part of H_k and modifying it if necessary to ensure its positive definiteness (see [1] for details of that procedure). It also allows a user-defined preconditioning via its reverse communication interface.

2.2 The filter-trust-region mechanism

Once the step s_k has been computed, we define the *trial point* $x_k^+ = x_k + s_k$ and consider the question of deciding whether or not it is acceptable as our next iterate x_{k+1}. We use a so-called filter to answer this question.

In order to define our filter, we first say that a point x_1 *dominates* a point x_2 whenever

$$|\theta_i(x_1)| \leq |\theta_i(x_2)| \text{ for all } i \in \mathcal{E}.$$

Thus, if iterate x_{k_1} dominates iterate x_{k_2}, the latter is of no real interest to us since x_{k_1} is at least as good as x_{k_2} for each i. All we need to do now is to remember iterates that are not dominated by other iterates using a structure called a filter. A *filter* is a list $\mathcal{F}$ of m-tuples of the form $(|\theta_{1,k}|, \ldots, |\theta_{m,k}|)$ such that, for $k \neq \ell$,

$$|\theta_{i,k}| < |\theta_{i,\ell}| \text{ for at least one } i \in \mathcal{E}.$$

Filter methods then accept the new trial iterate x_k^+ if it is not dominated by any other iterate in the filter. While the idea of not accepting dominated trial points is simple and elegant, it needs to be refined a little in order to provide an efficient algorithmic tool. In particular, we do not wish to accept a new point x_k^+ if $\theta_k^+ \stackrel{\text{def}}{=} \theta(x_k^+)$ is too close to being dominated by another point already in the filter. To avoid this situation, we slightly strengthen our acceptability condition. More formally, we say that a new trial point x_k^+ is *acceptable for the filter* $\mathcal{F}$ if and only if

$$\forall \theta_\ell \in \mathcal{F} \quad \exists\, i \in \mathcal{E} \quad |\theta_i(x_k^+)| < \Big[|\theta_{i,\ell}| - \gamma_\theta \|\theta_\ell\| \Big]_+ \tag{2.4}$$

where $\gamma_\theta \in (0, 1/\sqrt{m})$ is a small positive constant and $[w]_+ = \max[0, w]$.

In order to avoid cycling, and assuming the trial point is acceptable in the sense of (2.4), we may wish to add it to the filter, so as to avoid other iterates that are worse, that is we perform the simple operation $\mathcal{F} \leftarrow \mathcal{F} \cup \{\theta_k\}$. This may however cause an existing filter value θ_ℓ to be *strongly dominated* in the sense that

$$\exists\, \theta_q \in \mathcal{F} \quad \forall j \in \{1, \ldots, p\} \quad |\theta_{j,\ell}| \geq |\theta_{j,q}| - \gamma_\theta \|\theta_\ell\|. \tag{2.5}$$

If this happens, we simplify later comparisons by removing θ_ℓ from the filter.

If the trial point is not acceptable for the filter, it may nevertheless be acceptable for the usual trust-region mechanism. This requires that $\|s_k\| \leq \Delta_k$ and that

$$\rho_k = \frac{f(x_k) - f(x_k^+)}{m_k(x_k) - m_k(x_k^+)} \tag{2.6}$$

is sufficiently positive. Our algorithm therefore combines the filter and trust-region acceptability criteria to allow a potentially larger set of trial points to be accepted.

Inequality constraints are treated in a way entirely similar to that used for equalities: as already mentioned in (1.4) we define θ to measure the violation of the inequality constraints. Although the ℓ_2-penalty function (1.3) has discontinuous second derivatives on the boundary of the set of vectors satisfying the inequality constraints, this does not seem to create numerical difficulties when using the Gauss-Newton model.

2.3 An outline of the algorithm and a glimpse of numerical performance

We now outline the FILTRANE algorithm using the ideas developed above. This outline is presented as Algorithm 2.1, page 76. Clearly, this description leaves a number of points unspecified and not fully explained. A more detailed discussion is beyond the scope of this note, and we refer the reader to [10] for an in-depth analysis.

Algorithm 2.1: Outline of the Filter-Trust-Region Algorithm

Step 0: Initialization.
An initial point x_0 and an initial trust-region radius $\Delta_0 > 0$ are given, as well as constants $0 < \gamma_0 \leq \gamma_1 < 1 \leq \gamma_2$, $\gamma_\theta \in (0, 1/\sqrt{p})$, $0 < \eta_1 < \eta_2 < 1$. Compute $c_0 = c(x_0)$ and θ_0. Set $k = 0$, $\mathcal{F} = \emptyset$, and select $\tau_0 \geq 1$.

Step 1: Test for termination.
If either θ_k or $\|\nabla f(x_k)\|$ is sufficiently small, stop.

Step 2: Choose a model and a norm.
Choose a norm $\|\cdot\|_k$ for (2.3). Set m_k to be either $m_k^{\rm GN}$ or $m_k^{\rm N}$.

Step 3: Determine a trial step.
Compute a step s_k using the GLTR algorithm. If the model is found to be nonconvex and $\tau_k > 1$, reenter the GLTR algorithm with $\tau_k = 1$. Compute the trial point $x_k^+ = x_k + s_k$.

Step 4: Evaluate the residual at the trial step.
Compute $c(x_k^+)$ and $\theta_k^+ = \theta(x_k^+)$. Define ρ_k according to (2.6).

Step 5: Test to accept the trial step.

- If x_k^+ is acceptable for the current filter:
  Set $x_{k+1} = x_k^+$, select $\tau_{k+1} \geq 1$ and add θ_k^+ to $\mathcal{F}$ if either $\rho_k < \eta_1$ or $\|s_k\| > \Delta_k$.
- If x_k^+ is not acceptable for the current filter:
  If $\|s_k\| \leq \Delta_k$ and $\rho_k \geq \eta_1$, set $x_{k+1} = x_k^+$ and select $\tau_{k+1} \geq 1$. Else, set $x_{k+1} = x_k$ and $\tau_{k+1} = 1$.

Step 6: Update the trust-region radius.
If $\|s_k\| \leq \Delta_k$, update the trust-region radius by choosing

$$\Delta_{k+1} \in \begin{cases} [\gamma_0\Delta_k, \gamma_1\Delta_k] & \text{if } \rho_k < \eta_1, \\ [\gamma_1\Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2) \\ [\Delta_k, \gamma_2\Delta_k] & \text{if } \rho_k \geq \eta_2; \end{cases}$$

otherwise, set $\Delta_{k+1} = \Delta_k$. Increment k by one and go to Step 1.

We simply attempt to motivate the reader to investigate further by showing performance profiles comparing the classical trust-region framework to that including the filter technique (using the Gauss-Newton model (2.1)) introduced above on a set of 122 problems

from the CUTEr collection (see [13]). Given a set of test problems and a set of competing algorithms, the i-th performance profiles $p_i(\alpha)$ indicates the fraction of problems for which the i-th algorithm is within a factor α of the best for a given metric (see [4] for a formal definition of performance profiles and a discusion of their properties) Comparisons of both iteration counts and CPU times indicate that the filter technique results in a considerably improved performance compared to the more classical technique, and this is true both in reliability (on the right of the profiles) and in efficiency (on their left).

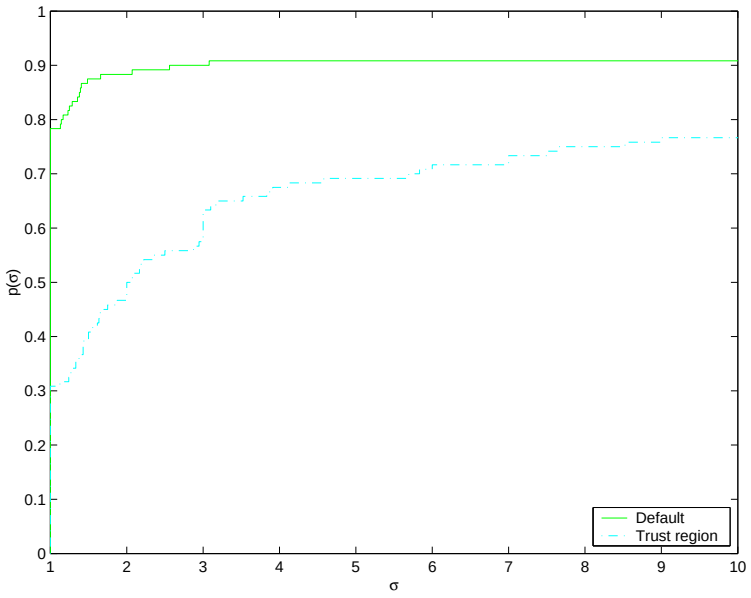


Figure 2.3: Iteration performance profiles $p(\alpha)$ for the default FILTRANE variant (including filter) and the pure trust-region variant (no filter)

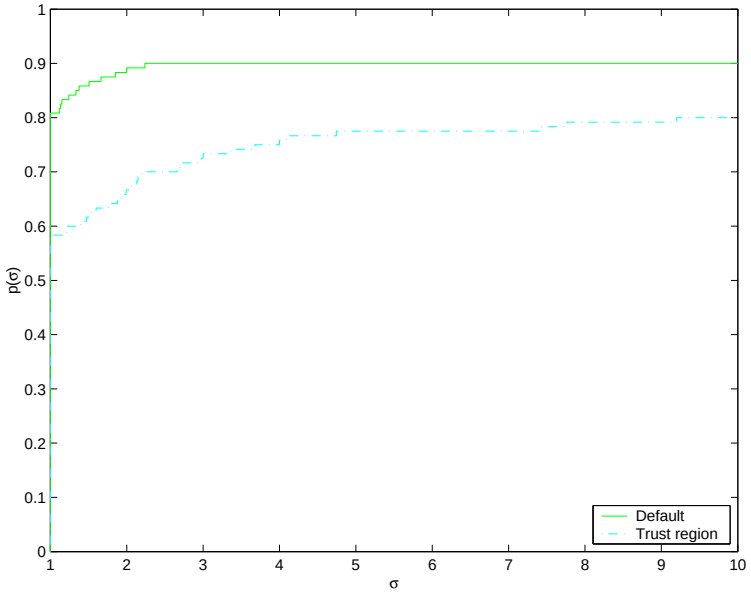


Figure 2.4: CPU time performance profiles $p(\alpha)$ for the default FILTRANE variant (including filter) and the pure trust-region variant (no filter)

3 Conclusion

We have briefly outlined the main ideas and the algorithm that are behing the FILTRANE package, which is available in the GALAHAD library of optimization programs (see [13] and `http://galahad.rl.ac.uk/galahad-www/`). We have also indicated that FILTRANE appears to be remarkably robust and efficient. We hope the supplied pointers will encourage the reader to pursue this subject of research, maybe with the help of the two more detailed papers on the subject, namely [11] and [10].

Acknowledgements

The second author is indebted to the Belgian National Fund for Scientific Research for its support during his 2002-2003 sabbatical leave. Some of this research was sponsored by EPSRC grants GR/R46641 and GR/S02969/01.

Current reports available from

`http://www.numerical.rl.ac.uk/reports/reports.html`

or by anonymous ftp from the directory
"`pub/reports`" on `thales.math.fundp.ac.be`

References

[1] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. LANCELOT*: a Fortran package for large-scale nonlinear optimization (Release A)*. Number 17 *in* 'Springer Series in Computational Mathematics'. Springer Verlag, Heidelberg, Berlin, New York, 1992.

[2] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 *in* 'MPS-SIAM Series on Optimization'. SIAM, Philadelphia, USA, 2000.

[3] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1983. Reprinted as *Classics in Applied Mathematics 16*, SIAM, Philadelphia, USA, 1996.

[4] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, **91**(2), 201–213, 2002.

[5] R. Fletcher and S. Leyffer. User manual for filterSQP. Numerical Analysis Report NA/181, Department of Mathematics, University of Dundee, Dundee, Scotland, 1998.

[6] R. Fletcher and S. Leyffer. Nonlinear programming without a penalty function. *Mathematical Programming*, **91**(2), 239–269, 2002.

[7] R. Fletcher, N. I. M. Gould, S. Leyffer, Ph. L. Toint, and A. Wächter. Global convergence of trust-region SQP-filter algorithms for nonlinear programming. *SIAM Journal on Optimization*, **13**(3), 635–659, 2002*a*.

[8] R. Fletcher, S. Leyffer, and Ph. L. Toint. On the global convergence of a filter-SQP algorithm. *SIAM Journal on Optimization*, **13**(1), 44–59, 2002*b*.

[9] C. C. Gonzaga, E. Karas, and M. Vanti. A globally convergent filter method for nonlinear programming. Technical report, Department of Mathematics, Federal University of Santa Catarina, Florianopolis, Brasil, 2002.

[10] N. I. M. Gould and Ph. L. Toint. FILTRANE, a Fortran 95 filter-trust-region package for solving systems of nonlinear equalities, nonlinear inequalities and nonlinear least-squares problems. Technical Report 03/15, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, 2003.

[11] N. I. M. Gould, S. Leyffer, and Ph. L. Toint. A multidimensional filter algorithm for nonlinear equations and nonlinear least-squares. Technical Report TR-2003-004, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, 2003*a*.

[12] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, **9**(2), 504–525, 1999.

[13] N. I. M. Gould, D. Orban, and Ph. L. Toint. GALAHAD—a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. *Transactions of the ACM on Mathematical Software*, **29**(4), (to appear), 2003*b*.

[14] J. J. Moré and D. C. Sorensen. Newton's method. *in* G. H. Golub, ed., 'Studies in Numerical Analysis', number 24 *in* 'MAA Studies in Mathematics', pp. 29–82, Providence, Rhode-Island, USA, 1984. American Mathematical Society.

[15] J. Nocedal. Trust region algorithms for solving large systems of nonlinear equations. *in* W. Liu, T. Belytschko and K. C. Park, eds, 'Innovative Methods for Nonlinear Problems', pp. 93–102. Pineridge Press, 1984.

[16] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, London, 1970.

[17] Ph. L. Toint. Numerical solution of large sets of algebraic nonlinear equations. *Mathematics of Computation*, **46**(173), 175–189, 1986.

[18] Ph. L. Toint. On large scale nonlinear least squares calculations. *SIAM Journal on Scientific and Statistical Computing*, **8**(3), 416–435, 1987.

Finite Differences in a Small World

Desmond J. Higham[(1)]
Department of Mathematics,
University of Strathclyde,
Glasgow G1 1XH, Scotland
djh@maths.strath.ac.uk

1 Introduction

Many complex networks in nature exhibit two properties that are seemingly at odds. They are *clustered*—neighbors of neighbors are very likely to be neighbors—and they are *small worlds*—any two nodes can typically be connected by a relatively short path. Watts and Strogatz [17] referred to this as the *small world phenomenon* and proposed a network model that was shown through simulation to capture the two properties. The model incorporates a parameter that interpolates between fully local and fully global regimes. As the parameter is varied the small world property is roused before the clustering property is lost. These computations were later backed up by a semi-heuristic mean-field analysis in [12]. Motivated by [17] and the book [16], many authors have shown experimentally that the small world phenomenon can be found across a range of areas in science, medicine and technology [3, 8, 11, 10, 14, 15]. In this work we show that the small world phenomenon also emerges through a local-to-global cutoff in the context of Markov chains. Our model is based on a simple random walk and hence is relevant to many physical, sociological, epidemiological and computational applications [1, 2, 9, 13]. In particular, it may be regarded as a *teleporting random walk*, of the type used by the search engine Google to locate pages on the web [6, 7], restricted to a simple ring network.

An extremely attractive feature of our model is that it can be analyzed rigorously by using tools from numerical analysis.

Full details of the material summarized here can be found in [5]. Recent work that deals with the question of navigation in a small world setting appears in [4].

2 Model

Our model is a perturbed version of a period random walk on a ring. Imagine the numbers $1, 2, 3, \ldots, N$ arranged like the hours on a clock. Take a periodic random walk around this clock—at each time level flip a fair coin; if it lands heads, move clockwise, and if it lands tails, move anticlockwise. Now alter the process so that before taking each step you toss another, biased, coin that lands heads with small probability δ. In the event that the biased coins lands heads, then pick a number between 1 and N uniformly at random and move there. Otherwise, follow the fair coin procedure above. In this model, most of the time we follow a traditional periodic random walk, but occasionally (when the biased coin lands heads) we take a *teleporting jump* across the network. This teleporting idea is used by search engines such as Google to locate pages on the web. (Traverse the web by following links out of pages, but, to avoid reaching a cycle or a dead end, take

[(1)]Supported by a Research Fellowship from the Leverhulme Trust

the occasional jump). Teleporting is very similar in spirit to the rewiring/shortcutting process that has been observed to produce small world networks of the type described in section 1, and we will show that an analogous cutoff effect arises.

3 Result

The small world phenomenon for networks involves the pathlength between nodes. Analogously, we may study the small world phenomenon for our random walk model by measuring the *mean hitting time*, that is, the average number of steps that it takes to reach B starting from A, where $A, B \in \{1, 2, \ldots, N\}$ are chosen uniformly at random.

The problem of analysing the mean hitting time can be set up as a linear algebraic system of dimension $N-1$. We are interested in behaviour for large systems, that is, the $N \to \infty$ limit.

The key to the analysis is to identify the mean hitting time system as a finite difference approximation to an underlying continuum limit. The continuum limit itself depends upon N (i.e. the grid spacing), but careful analysis shows that convergence takes place in an appropriate sense, so that limiting behaviour may be captured.

The analysis shows that it is natural to let the teleporting jump probability, δ, scale with N according to

$$\delta = \frac{K}{N^2}, \qquad \text{for fixed } K.$$

In this regime, we find that, up to $O(N^{-1})$ terms, the mean hitting time, normalized by the mean hitting time when $\delta = 0$, has the form

$$y = \frac{6}{K}x, \tag{3.7}$$

where

$$x = \frac{\sqrt{2K}}{2\tanh\frac{\sqrt{2K}}{2}} - 1 \tag{3.8}$$

is an expression for the expected number of teleporting jumps taken per random walk.

We may now examine the small world property of the model by plotting y against x as K is varied. This shows how the introduction of shortcuts (teleporting jumps) makes it easier to get around the network. Figure 3.5 gives the picture. We see that there is an abrupt drop in the hitting time as x increases beyond $x \approx 10^{-1}$. This "order 1 shortcut cutoff effect" agrees with the fundamental observation of Watts and Strogatz [17] for their small world network model.

References

[1] H. C. Berg, *Random Walks in Biology*, Princeton University Press, 1983.

[2] H. Caswell, *Matrix Population Models*, Sinauer Associates, Inc., Sunderland, MA, second ed., 2001.

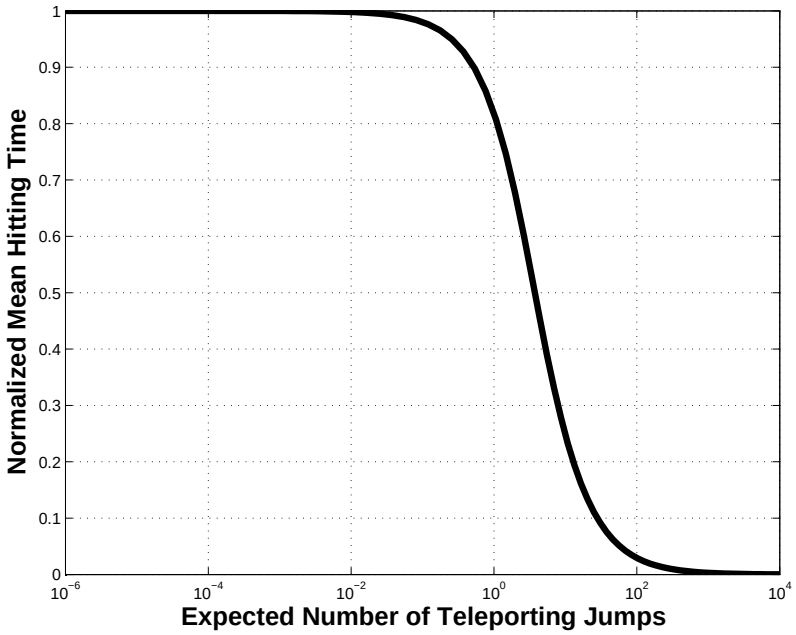


Figure 3.5: x-axis is the average number of teleporting jumps per excursion, y-axis is the mean hitting time (normalized by dividing by the mean hitting time with $\delta = 0$). The picture was produced from (3.7)–(3.8) by varying K.

[3] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell, *Relevant cycles in chemical reaction networks*, Advances in Complex Systems, 4 (2001), pp. 207–226.

[4] D. J. Higham, *Greedy pathlengths and small world graphs*, Tech. Rep. 8, University of Strathclyde, Glasgow, UK, 2002.

[5] ——, *A matrix perturbation view of the small world phenomenon*, SIAM J. Matrix Anal. Appl., (to appear (2003)).

[6] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, *Extrapolation methods for accelerating pagerank computations*, Proceedings of the Twelfth International World Wide Web Conference, (2003).

[7] C. Moler, *The world's largest matrix computation*, MATLAB News and Notes, October (2002).

[8] J. M. Montoya and R. V. Solé, *Small world patterns in food webs*, Journal of Theoretical Biology, 214 (2002), pp. 405–412.

[9] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[10] M. E. J. Newman, *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci., 98 (2001), pp. 404–409.

[11] ——, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.

[12] M. E. J. Newman, C. Moore, and D. J. Watts, *Mean-field solution of the small-world network model*, Physical Review Letters, 84 (2000), pp. 3201–3204.

[13] J. R. Norris, *Markov Chains*, Cambridge University Press, 1997.

[14] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[15] A. Wagner and D. A. Fell, *The small world inside large metabolic networks*, Proc. Roy. Soc. London, B., 268 (2001), pp. 1803–1810.

[16] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, 1999.

[17] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

Monotonicity for Time Discretizations

Willem Hundsdorfer & Steven J. Ruuth
CWI, Amsterdam,
The Netherlands
willem.hundsdorfer@cwi.nl
&
Department of Mathematics,
Simon Fraser University,
Burnaby, Canada
sruuth@sfu.ca

This review is largely based on material from [7, 8], where additional results and a more precise presentation can be found.

1 Introduction

The preservation of monotonicity properties – like maximum principles and positivity – is often essential for numerical schemes to approximate non-smooth solutions in a qualitatively correct manner. In these notes a review is given of monotonicity results for time stepping with Runge-Kutta and linear multistep methods.

We consider an ODE system in $\mathbb{R}^m$

$$w'(t) = F(w(t)), \quad w(0) = w_0 .$$

In the applications this will originate from a PDE after suitable spatial discretization. Then approximations $w_n \approx w(t_n)$, $t_n = n\Delta t$, are obtained by a time stepping method with step size Δt.

In these notes we shall deal with the property

$$\|w_n\| \leq \|w_0\| \qquad \text{for } \ n \geq 1 , \ w_0 \in \mathbb{R}^m, \tag{1.9}$$

where $\|\cdot\|$ is a given semi-norm. This may be, for example, the maximum-norm or the total variation over the components. In the latter case (1.9) is called the TVD property (total variation diminishing). It is assumed that

$$\|v + \Delta t\, F(v)\| \leq \|v\| \qquad \text{for } \ 0 < \Delta t \leq \Delta t_{FE} , \ v \in \mathbb{R}^m, \tag{1.10}$$

where Δt_{FE} can be viewed as the maximal step size for the forward Euler method. Condition (1.10) is often easy to verify [5, 8]. The goal is to specify for higher-order methods the *monotonicity threshold* $C > 0$ such that (1.9) holds whenever $\Delta t \leq C\, \Delta t_{FE}$.

Related monotonicity properties are *positivity* ($w_n \geq 0$ whenever $w_0 \geq 0$) and the *comparison principle* ($w_n \geq v_n$ whenever $w_0 \geq v_0$), with inequalities for vectors component-wise. Properties like these and (1.9) are often called *monotonicity*. The main motivation for wanting such properties is to avoid oscillations in the numerical solutions and to prevent over- and undershoots.

Example 1 Before considering general results for time stepping methods, in the Sections 2, 3 and 4, let us first consider some simple examples.

The implicit Euler method

$$w_n = w_{n-1} + \Delta t\, F(w_n) \tag{1.11}$$

is unconditionally monotone, $C_{BE} = \infty$. This is easily seen from

$$\Big(1 + \frac{\Delta t}{\Delta t_{FE}}\Big) w_n \;=\; w_{n-1} + \frac{\Delta t}{\Delta t_{FE}} \Big(w_n + \Delta t_{FE}\, F(w_n) \Big)$$

with application of the triangle inequality to the right-hand side.
The implicit trapezoidal rule

$$w_n = w_{n-1} + \frac{1}{2}\,\Delta t\, F(w_{n-1}) + \frac{1}{2}\,\Delta t\, F(w_n) \tag{1.12}$$

has threshold factor $C_{ITR} = 2$; the method consists of a forward Euler half-step followed by a backward Euler half-step (with $\frac{1}{2}\Delta t$).
The explicit trapezoidal rule (modified Euler)

$$\overline{w}_n = w_{n-1} + \Delta t\, F(w_{n-1})\,, \qquad w_n = w_{n-1} + \frac{1}{2}\,\Delta t\, F(w_{n-1}) + \frac{1}{2}\,\Delta t\, F(\overline{w}_n) \tag{1.13}$$

has threshold factor $C_{ETR} = 1$. This becomes more apparent by writing the second stage as

$$w_n = \frac{1}{2}\, w_{n-1} + \frac{1}{2}\,\Big(\overline{w}_n + \Delta t\, F(\overline{w}_n) \Big)\,.$$

From these simple examples we see that methods have to be rewritten sometimes in a form that is more convenient to make the monotonicity apparent. More importantly, we see that there is no direct relation with the usual stability properties of the methods. After all, the implicit trapezoidal rule is A-stable, whereas its explicit counterpart (1.13) is only conditionally stable. In fact, the backward Euler method will turn out to be the only well-known method with threshold value $C = \infty$.

Example 2 To illustrate the relevance of monotonicity, we consider the Buckley-Leverett equation

$$u_t + f(u)_x = 0\,, \qquad f(u) = \frac{3u^2}{3u^2 + (1-u)^2}\,,$$

for $t \geq 0$, $0 \leq x \leq 1$, with inflow condition $u(0,t) = 1$ and an initial block-function: $u(x,0)$ is zero on $(0, \frac{1}{2}]$, one on $(\frac{1}{2}, 1]$. We use a fixed grid with mesh width $\Delta x = 5 \cdot 10^{-3}$ and a flux-limited spatial discretization (van Leer type; see [8] for the semi-discrete form). This then defines our ODE system. The PDE solution consists of two shocks followed by rarefaction waves; in the semi-discrete solution the shocks are slightly diffused, over a few grid cells.
The implicit BDF2 scheme

$$w_n = \frac{4}{3} w_{n-1} - \frac{1}{3} w_{n-2} + \frac{2}{3} \Delta t\, F(w_n) \tag{1.14}$$

has order 2 and it is A-stable, that is, unconditionally stable in a von Neumann analysis. Its explicit counterpart, the extrapolated BDF2 scheme

$$w_n = \frac{4}{3} w_{n-1} - \frac{1}{3} w_{n-2} + \frac{4}{3} \Delta t\, F(w_{n-1}) - \frac{2}{3} \Delta t\, F(w_{n-2}) \tag{1.15}$$

also has order 2. It is stable with the present spatial discretization for Courant numbers up to 0.5, approximately. As we shall see below, the two schemes have approximately the same monotonicity threshold.

Numerical solutions are given in the Figures 1.6 and 1.7. Let us first consider the results with $\Delta t = \frac{1}{800}$. Then both methods give results close to the exact semi-discrete solution. Next we take $\Delta t = \frac{1}{400}$. Then the explicit scheme is becoming unstable. However, also the (unconditionally stable) implicit scheme gives bad results; we now have a wrong location and height of the shocks. This is due to *loss of monotonicity*, giving over- and undershoots after the shocks. (Global overshoots would occur with different initial and boundary conditions, e.g., $u(x,0) = 0$, $u(0,t) = \frac{1}{2}$.)

Example 3 The standard examples for which monotonicity is relevant arise in hyperbolic conservation laws. To a lesser extend monotonicity is also important for certain parabolic examples. As an illustration consider the Fisher equation

$$u_t = \epsilon u_{xx} + \gamma u(1-u^2)$$

with traveling wave solution $u(x,t) = (1 + e^{\lambda(x-1-\alpha t)})^{-1}$, where $\lambda = \frac{1}{2}\sqrt{2\gamma/\epsilon}$ and $\alpha = \frac{3}{2}\sqrt{2\gamma\epsilon}$. The parameters are taken as $\gamma = \epsilon^{-1} = 100$ and $0 < x < L = 6$, $0 < t \leq 1$, with homogeneous Neumann conditions at the boundaries. The mesh width is $\Delta x = 10^{-2}$, and standard second-order differences are used in space.

Then with the explicit Euler method we need approximately $\Delta t \lesssim 1/300$ for monotonicity. Here, we consider the implicit 2-stage Gauss Runge-Kutta method of order 4, and $\Delta t = \frac{1}{100}, \frac{1}{50}$. The results are shown in Figure 1.8. This method is A-stable but its monotonicity properties are poor ($C = 0$, see [17]). The large oscillations are initiated by small negative solution values that are amplified towards -1 by the reaction term. Because the solution is smooth on fine grids, the threshold $C = 0$ is pessimistic for this example. Nonetheless, it is obvious that the method cannot take large steps without loosing the correct qualitative behaviour.

2 Runge-Kutta methods

The first results on monotonicity were derived for *linear systems* by Bolley & Crouzeix [1] and Spijker [15]. This gives upper bounds for nonlinear systems. For Runge-Kutta methods with s stages and order p, the maximal threshold value $C = C_{RK}$ of explicit methods with $p = s$ is $C_{RK} = 1$. Moreover, among the well-known implicit methods we have unconditional monotonicity, $C_{RK} = \infty$, only for the backward Euler method [1].

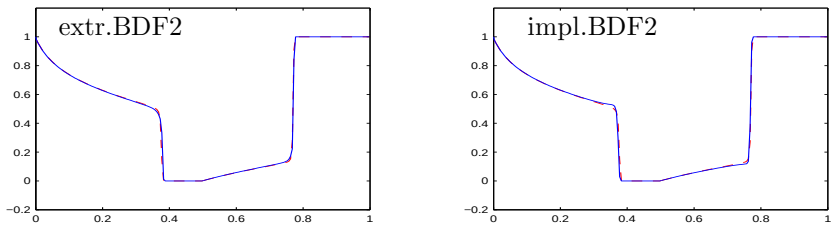


Figure 1.6: BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $\Delta t = 1/800$. The dashed line is a time-accurate semi-discrete solution ($\Delta x = 5 \cdot 10^{-3}$).

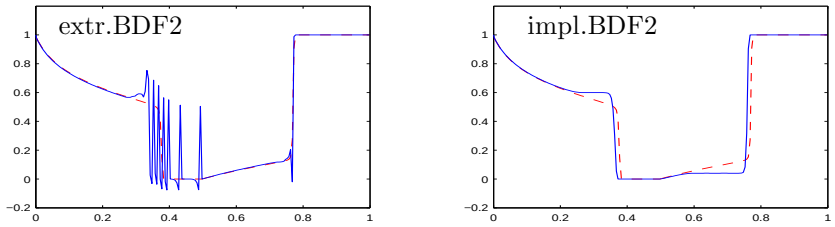


Figure 1.7: BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $\Delta t = 1/400$. The dashed line is a time-accurate semi-discrete solution ($\Delta x = 5 \cdot 10^{-3}$).

Results for *nonlinear systems/* were obtained first by Shu & Osher [14], using forms with combinations of Euler steps. For example, a diagonally implicit method can be written as

$$\left.\begin{aligned} v_0 &= w_{n-1}, \\ v_i &= \sum_{j=0}^{i-1} (p_{ij} v_j + q_{ij} \Delta t F(v_j)) + q_{ii} \Delta t F(v_i), \qquad i = 1, \ldots, s, \\ w_n &= v_s. \end{aligned}\right\} \tag{2.16}$$

If all $p_{ij}, q_{ij} \geq 0$, then

$$\|w_n\| \ \leq \ \|w_{n-1}\|$$

holds under step size restriction

$$\Delta t \ \leq \ C_{RK} \, \Delta t_{FE} \,, \qquad C_{RK} = \min_{1 \leq j \leq i-1} \left(\frac{p_{ij}}{q_{ij}} \right) . \tag{2.17}$$

Necessary and sufficient conditions were obtained by Kraaijevanger [10] on nonlinear contractivity. A recent result of Ferracina & Spijker [2] shows that with the Shu-Osher form (2.16) an optimal choice of p_{ij}, q_{ij} leads as well to necessary conditions, also with fully implicit methods. (Note that for a given method in the form of a Butcher tableau, there is some freedom in the choice of the p_{ij}, q_{ij}).

For further results on classes of Runge-Kutta methods with optimal threshold $C_{RK} > 0$, for given p and s, see for instance [2, 4, 5, 8, 10, 16].

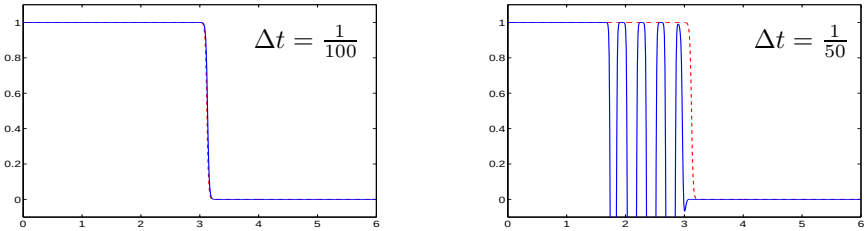


Figure 1.8: Results for the Fisher equation with the 2-stage Gauss RK method, $\Delta t = 1/100, 1/50$, at $t = 1$. The dashed line is the exact PDE solution.

Example 4 The optimal explicit second-order methods are given by

$$\left.\begin{aligned} v_0 &= w_{n-1}\,, \\ v_i &= v_{i-1} + \frac{1}{s-1}\,\Delta t\, F(v_{i-1})\,, \qquad i = 1, \ldots, s-1\,, \\ w_n &= \frac{1}{s}\, w_{n-1} + \frac{s-1}{s}\left(v_{s-1} + \frac{1}{s-1}\,\Delta t\, F(v_{s-1})\right). \end{aligned}\right\} \tag{2.18}$$

This class of methods was derived by Kraaijevanger [9, 10]. The monotonicity threshold factor for these methods is given by

$$C_{RK} = s - 1\,.$$

Kraaijevanger's results were formulated in terms of contractivity. The same methods were derived by Gerisch & Weiner [3] and Spiteri & Ruuth [16], studying related monotonicity properties.

These methods have nicely shaped stability regions (where stability is valid for $z = \Delta t\,\lambda \in \mathcal{S}$ with the scalar test equation $w' = \lambda w$), see Figure 2.9.

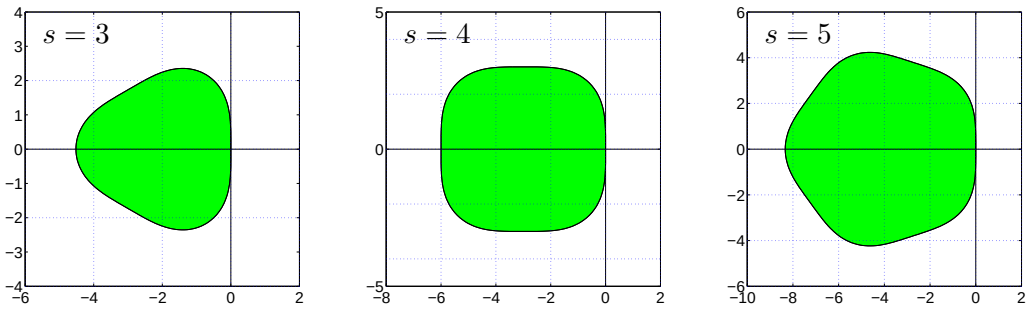


Figure 2.9: Stability regions $\mathcal{S}$ for the optimal 2nd-order explicit Runge-Kutta methods, $s = 3, 4, 5$.

3 Linear multistep methods with arbitrary starting values

Consider a linear multistep method

$$w_n \;=\; \sum_{j=1}^{k} \Big(a_j w_{n-j} + b_j \Delta t\, F(w_{n-j})\Big) \;+\; b_0 \Delta t\, F(w_n)\,. \tag{3.19}$$

If all $a_j, b_j \geq 0$, then

$$\|w_n\| \;\leq\; \max_{1\leq j\leq k} \|w_{n-j}\|$$

will hold under the step size restriction

$$\Delta t \;\leq\; C_{LM}\, \Delta t_{FE}\,, \qquad C_{LM} = \min_{1\leq j\leq k} \left(\frac{a_j}{b_j}\right). \tag{3.20}$$

This result is due to Shu [13] (with $b_0 = 0$), originally in terms of total variations. Related results for linear systems were given in [1] (on positivity), and in [15, 11] (on contractivity).

Using the order conditions, it was shown by Lenferink [11] that the maximal size of the threshold factor C_{LM} for explicit k-step methods of order p is bounded by

$$\begin{cases} C_{LM} \leq 1 & \text{if } p = 1\,, \\ C_{LM} \leq \dfrac{k-p}{k-1} & \text{if } p \geq 2\,. \end{cases} \tag{3.21}$$

The bound for $k = 1$ is attained by the forward Euler method. Optimal higher-order multistep methods have been constructed by Shu [13], Lenferink [11] and Gottlieb et al. [5].

Example 5 The explicit 3-step method

$$w_n = \frac{3}{4}\, w_{n-1} + \frac{1}{4}\, w_{n-3} + \frac{3}{2}\, \Delta t\, F(w_{n-1}) \tag{3.22}$$

has order $p = 2$ and $C_{LM} = \frac{1}{2}$. The explicit 4-step method

$$w_n = \frac{8}{9}\, w_{n-1} + \frac{1}{9}\, w_{n-4} + \frac{4}{3}\, \Delta t\, F(w_{n-1}) \tag{3.23}$$

has $p = 2$ and threshold factor $C_{LM} = \frac{2}{3}$. The stability regions of these methods is given in Figure 3.10.

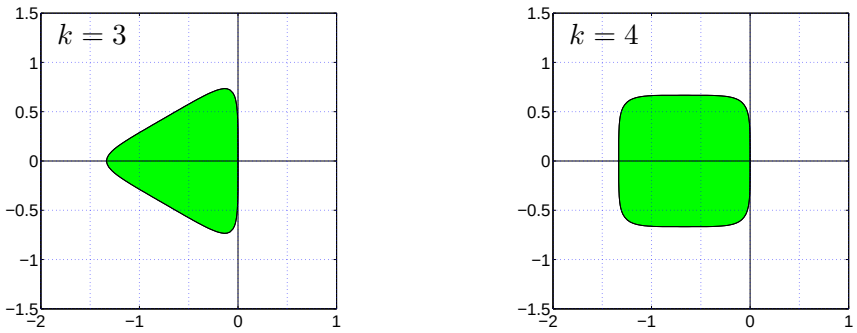


Figure 3.10: Stability regions $\mathcal{S}$ for the optimal second-order explicit k-step methods, $k = 3, 4$.

4 Linear multistep methods with starting procedures

The above results with arbitrary starting values exclude many schemes that are useful in practice and also may give unnecessary step size restrictions. For example, for explicit methods with $p = k = 2$ we cannot have $C_{LM} > 0$ in view of (3.21). Also, Adams and BDF schemes are excluded. This is due to insistence on *arbitrary* initial vectors, which may give too strong restrictions.

For instance, application of the BDF2 method to the trivial problem $w'(t) = 0$ yields

$$w_2 = \frac{4}{3}\, w_1 - \frac{1}{3}\, w_0\,.$$

Obviously, we do not have $\|w_2\| \le \max\big(\|w_0\|, \|w_1\|\big)$ for arbitrary w_0, w_1; but it is also obvious that only $w_1 = w_0$ makes sense here. Therefore we will look at the methods in combination with *starting procedures*. In the following presentation, based on [7], we restrict ourselves to 2-step methods.

The standard form of a 2-step method reads

$$w_n - b_0\Delta t\, F_n = a_1 w_{n-1} + a_2 w_{n-2} + b_1\Delta t\, F_{n-1} + b_2\Delta t\, F_{n-2} \tag{4.24}$$

for $n \ge 2$, with $F_n = F(w_n)$. Let $\theta \ge 0$. By subtracting recursively terms $\theta^j w_{n-j}$ and using (4.24), the multistep scheme can be written out fully up to the starting values,

$$\begin{aligned} w_n - b_0\Delta t\, F_n &= (a_1 - \theta)w_{n-1} + (b_1 + \theta b_0)\Delta t\, F_{n-1} \\ &+ \sum_{j=2}^{n-2} \theta^{j-2}\Big((a_2 + \theta a_1 - \theta^2)w_{n-j} + (b_2 + \theta b_1 + \theta^2 b_0)\Delta t\, F_{n-j}\Big) \\ &+ \theta^{n-3}\Big((a_2 + \theta a_1)w_1 + (b_2 + \theta b_1)\Delta t\, F_1 + \theta a_2 w_0 + \theta b_2 \Delta t\, F_0\Big)\,. \end{aligned}$$

Define

$$C^*_{LM} = \max_{\theta}\ \min\left(\frac{a_1 - \theta}{b_1 + \theta b_0}\ ,\ \frac{a_2 + \theta a_1 - \theta^2}{b_2 + \theta b_1 + \theta^2 b_0}\right)\ , \tag{4.25}$$

where θ is restricted to get nonnegative coefficients. For the starting procedure that yields w_1 we assume that

$$\begin{aligned} &\|w_1\| \le M\|w_0\|\,, \quad \|w_2\| \le M\|w_0\|\,, \\ &\|(a_2 + \theta a_1)w_1 + (b_2 + \theta b_1)\Delta t\, F_1 + \theta a_2 w_0 + \theta b_2 \Delta t\, F_0\| \ \le\ (a_2 + \theta)M\,\|w_0\|\,, \end{aligned} \tag{4.26}$$

with constant $M \ge 1$. Then we have the following result [7].

Theorem 1 *Suppose (4.26) and $\Delta t \le C^*_{LM}\,\Delta t_{FE}$. Then $\|w_n\| \le M\|w_0\|$ for all $n \ge 1$.*

We note that for any starting procedure, there will be an $M \ge 1$ such that (4.26) is valid with $\Delta t \le C^*_{LM}\,\Delta t_{FE}$. To get genuine monotonicity, that is, $M = 1$, conditions on the starting procedure and perhaps an additional step size restriction will be needed.

Example 6 : Explicit second-order 2-step methods. The explicit methods with $k = p = 2$ form a one-parameter family,

$$w_n \ =\ (2 - \xi)\, w_{n-1} + (\xi - 1)\, w_{n-2} + \ (1 + \frac{1}{2}\xi)\,\Delta t\, F_{n-1} + (\frac{1}{2}\xi - 1)\,\Delta t\, F_{n-2}\,. \tag{4.27}$$

The parameter ξ should be in the interval $(0, 2]$ for zero-stability. Interesting examples are $\xi = 1$ (Adams-Bashforth) and $\xi = \frac{2}{3}$ (extrapolated BDF2). Here

$$C^*_{LM} \ =\ \frac{2(1 + \xi)(2 - \xi)}{(2 + \xi)^2}\ .$$

To have $\|w_n\| \le \|w_0\|$ for all n, there can be an additional restriction $\Delta t \le C^0_{LM}\Delta t_{FE}$ for the starting procedure.
A natural starting procedure for (4.27) is given by the forward Euler method,

$$w_1 = w_0 + \Delta t\, F_0\,.$$

We then have the following result (with $M = M(\xi) \ge 1$):

$$\begin{aligned} &\Delta t \le C^*_{LM}\,\Delta t_{FE}\,, \quad 0 < \xi \le 2 \quad \Longrightarrow \quad \|w_n\| \le M\,\|w_0\|\,, \\ &\Delta t \le \frac{2-\xi}{2+\xi}\Delta t_{FE}\,, \quad \frac{2}{3} \le \xi \le 2 \quad \Longrightarrow \quad \|w_n\| \le \|w_0\|\,. \end{aligned}$$

Here it should be noted that the restriction for genuine monotonicity ($M = 1$) can be relaxed with more general starting procedures. In numerical experiments (linear advection, Burgers' equation) the restriction $\Delta t \le C^*_{LM}\,\Delta t_{FE},\ 0 < \xi \le 2$ was found to be practically relevant. For small $\xi > 0$ the methods often gave inaccurate results due to compression; good numerical results were obtained for $\xi = \frac{2}{3}$ (extrapolated BDF2). See [7] for details.

Example 7 : Implicit second-order 2-step methods. The implicit methods with $k = p = 2$ form a two-parameter family,

$$\begin{aligned} w_n - \eta\Delta t\, F_n \;&=\; (2-\xi)w_{n-1} + (\xi-1)w_{n-2} \\ &+ (1 + \frac{1}{2}\xi - 2\eta)\Delta t\, F_{n-1} + (\eta + \frac{1}{2}\xi - 1)\Delta t\, F_{n-2} \end{aligned} \tag{4.28}$$

with $0 < \xi \le 2$, $\eta \ge 0$; for A-stability we need $\eta \ge \frac{1}{2}$. Interesting classes are $\xi = \frac{2}{3}$ (BDF2-type) and $\xi = 1$ (2-step Adams type).
Determination of the optimal factors C^*_{LM} is easy numerically. Plots for $\xi = \frac{2}{3}, 1$ as function of η are given in Figure 4.11. For the familiar implicit BDF2 method ($\xi = \eta = \frac{2}{3}$) we have $C^*_{LM} = \frac{1}{2}$, which is even less than for its explicit counterpart ($\xi = \frac{2}{3}$, $\eta = 0$) where $C^*_{LM} = \frac{5}{9}$. In fact, for these two methods the same thresholds are found if only linear problems are considered [6].

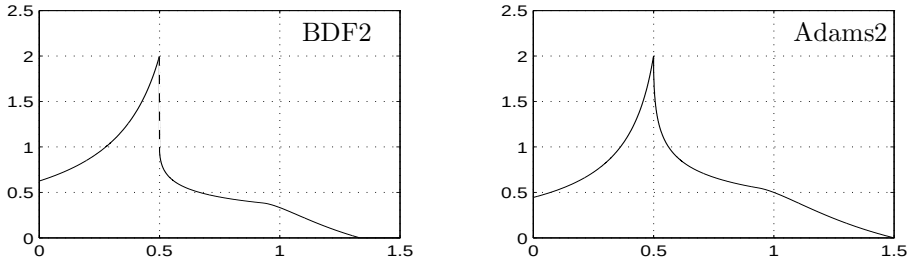


Figure 4.11: Threshold values C^*_{LM} versus $\eta \in [0, 1.5]$, with $\xi = \frac{2}{3}$ (left) and $\xi = 1$ (right).

Overall, these results are disappointing. The largest numbers $C^*_{LM} = 2$ are found for the values $\eta = \frac{1}{2}$ (similar behaviour with other $\xi \in (0, 2]$). Experimental verification of the curves in Figure 4.11 is given in [7] for the advection model problem $u_t + u_x = 0$ with first-order spatial differences.

5 Summary

For Runge-Kutta methods the basic monotonicity theory is fairly complete. There is a continuing search for 'optimal' explicit methods and theoretical refinements; see [2, 16, 12] for some recent papers. From a practical point of view, it is important to note that for many standard methods – including implicit (A-stable) methods – quite small step sizes may be needed if monotonicity properties are crucial in an application.

For linear multistep methods, on the other hand, the theory is less well developed. Inclusion of starting procedures in the considerations is needed to get reasonable step size restrictions for monotonicity or boundedness. This inclusion allows statements with classes of methods that are important in practice (such as Adams and BDF-type). In the numerical tests in [7] the standard Adams-Bashforth schemes ($k \leq 3$) and the extrapolated BDF schemes ($k \leq 4$) performed better than special constructed methods [5, 11, 13] with positive coefficients.

Apart from results with 2-step methods, some higher-order methods were analyzed in [7], but each class of methods did require a separate analysis. Optimality statements for general higher-order methods are lacking. Also results for predictor-corrector schemes are at present still absent.

References

[1] C. Bolley, M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*. RAIRO Anal. Numer. 12 (1978), 237–245.

[2] L. Ferracina, M.N. Spijker, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*. Report MI 2003-08, Univ. Leiden, 2003.

[3] A. Gerisch, R. Weiner, *On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods*. Comp. and Math. with Appl. 45 (2003), 53–67.

[4] S. Gottlieb, C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*. Math. Comp. 67 (1998), 73–85.

[5] S. Gottlieb, C.-W. Shu, E. Tadmor, *Strong stability preserving high-order time discretization methods*. SIAM Review 42 (2001), 89–112.

[6] W. Hundsdorfer, *Partially implicit BDF2 blends for convection dominated flows*. SIAM J. Numer. Anal. 38 (2001), 1763–1783.

[7] W. Hundsdorfer, S.J. Ruuth, R.J. Spiteri, *Monotonicity-preserving linear multistep methods*. SIAM J. Numer. Anal. 41 (2003), 605–623.

[8] W. Hundsdorfer, J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Series in Computational Mathematics 33, Springer Verlag, 2003.

[9] J.F.B.M. Kraaijevanger, *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*. Numer. Math. 48 (1986), 303–322.

[10] J.F.B.M. Kraaijevanger, *Contractivity of Runge-Kutta methods*. BIT 31 (1991), 482–528.

[11] H.W.J. Lenferink, *Contractivity preserving explicit linear multistep methods.* Numer. Math. 55 (1989), 213–223.

[12] S.J. Ruuth, R.J. Spiteri, *High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations.* Submitted, 2003.

[13] C.-W. Shu, *Total-variation-diminishing time discretizations.* SIAM J. Sci. Stat. Comput. 9 (1988), 1073–1084.

[14] C.-W. Shu, S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes.* J. Comput. Phys. 77 (1988), 439–471.

[15] M.N. Spijker, *Contractivity in the numerical solution of initial value problems.* Numer. Math. 42 (1983), 271–290.

[16] R.J. Spiteri, S.J. Ruuth, *A new class of optimal high-order strong-stability-preserving time-stepping schemes.* SIAM J. Numer. Anal. 40 (2002), 469–491.

[17] J.A. van de Griend, J.F.B.M. Kraaijevanger, *Absolute monotonicity of rational functions occurring in the numerical study of initial value problems.* Numer. Math. 49 (1986), 413–424.

Numerical methods for convection-diffusion problems or The 30 years war

Martin Stynes
Department of Mathematics,
National University of Ireland,
Cork, Ireland
m.stynes@ucc.ie

Abstract

Convection-diffusion problems arise in the modelling of many physical processes. Their typical solutions exhibit boundary and/or interior layers. Despite the linear nature of the differential operator, these problems pose still-unanswered questions to the numerical analyst.

This talk will give a selective overview of numerical methods for the solution of convection-diffusion problems, while placing them in a historical context. It examines the principles that underpin the competing numerical techniques in this area and presents some recent developments.

1 Talk overview

To quote the opening words of Morton's book [17]: "Accurate modelling of the interaction between convective and diffusive processes is the most ubiquitous and challenging task in the numerical approximation of partial differential equations."

I shall describe the nature of (steady-state) convection-diffusion problems, then draw some comparisons between the development of numerical methods for convection-diffusion problems during the last 30 years and that well-known 17th-century conflict known as the 30 Years War, whose history most Europeans learn during their schooldays.

Let's begin by reminding ourselves of its main features.

2 The phases of the 30 Years War

The 30 Years War war began in 1618 as a struggle between certain Catholic and Protestant states, but eventually sucked in all the major European countries. It devastated much of present-day Germany and the Czech Republic; indeed, not until the 2nd World War, 300 years later, was there a comparable amount of destruction in any European war.

As various protagonists entered or left the the conflict, the action passed through several *phases* [10, pp.252–255]: the 1618–23 Bohemian phase, the 1624–29 Danish phase, the 1630–35 Swedish phase, and the 1635–48 Franco-Swedish phase.

To indicate how science was progressing at this time, we note that Kepler's 3rd law (proportionality of the square of the period of revolution of a planet to the cube of the

length of the major axis of its orbit) was published in 1619. He was then living in Linz, in Austria.
The war ended with the Peace of Westphalia (Westfalen in German) in 1648.

3 Convection-diffusion problems

These take the form

$$-\varepsilon\Delta u + \boldsymbol{\beta}\cdot\nabla u = f \text{ on } \Omega, \tag{3.29}$$

with some boundary conditions on $\partial\Omega$, where $-\varepsilon\Delta u$ models diffusion and $\boldsymbol{\beta}\cdot\nabla u$ models convection. The parameter ε is positive but small; think of it as say 10^{-6}. The region Ω is any reasonable domain in n dimensions, where $n \geq 1$. The differential operator is elliptic, so under suitable hypotheses on $\boldsymbol{\beta}$ and the boundary data, (3.29) has a solution in $C^2(\Omega)$. Here we take $\boldsymbol{\beta} \approx O(1)$, i.e., *convection dominates diffusion*:

$$\frac{|\text{coefficient of } \nabla u|}{|\text{coefficient of } \Delta u|} = \frac{|\boldsymbol{\beta}|}{|\varepsilon|} >> 1.$$

For most boundary conditions, this is an example of a singularly perturbed partial differential equation (PDE). Convection-diffusion PDEs arise in many applications [17] such as the linearized Navier-Stokes equations and the drift-diffusion equation of semiconductor device modelling.
To get some feeling for the behaviour of solutions to such problems, let's consider a simple example in one dimension, that is, an ordinary differential equation (ODE):

$$\begin{aligned} -\varepsilon u'' + u' &= 2 \quad \text{on } (0,1), \\ u(0) = u(1) &= 0. \end{aligned}$$

Then

$$\begin{aligned} u(x) &= 2x + \frac{2(e^{-1/\varepsilon} - e^{-(1-x)/\varepsilon})}{1 - e^{-1/\varepsilon}} \\ &= 2x - 2e^{-(1-x)/\varepsilon} + O(e^{-1/\varepsilon}). \end{aligned} \tag{3.30}$$

Here $2x$ is the solution of the first-order problem $u'(x) = 2$, $u(0) = 0$; the rapidly-decaying exponential $e^{-(1-x)/\varepsilon}$ is a boundary layer function—i.e., it is not large but its first-order derivative is large near $x = 1$; and the final term $O(e^{-1/\varepsilon})$ is negligible.
For PDEs also, the solution u of (3.29) has a structure analogous to (3.30): it can be written as the sum of the solution to a 1st-order hyperbolic problem + layer(s) + negligible terms. Let's make this a little more precise. Divide the boundary $\partial\Omega$ into 3 parts:

$$\begin{aligned} \text{inflow boundary } \partial^-\Omega &= \{x \in \partial\Omega : \boldsymbol{\beta}.\boldsymbol{n} < 0\}, \\ \text{outflow boundary } \partial^+\Omega &= \{x \in \partial\Omega : \boldsymbol{\beta}.\boldsymbol{n} > 0\}, \\ \text{tangential flow boundary } \partial^0\Omega &= \{x \in \partial\Omega : \boldsymbol{\beta}.\boldsymbol{n} = 0\}, \end{aligned}$$

where $\boldsymbol{n}$ is the outward-pointing unit normal to $\partial\Omega$. Then the 1st-order hyperbolic problem is $\boldsymbol{\beta} \cdot \nabla u = f$ on Ω, with boundary data specified on $\partial^-\Omega$. Professor Ron Mitchell, one the founders of the Dundee conferences, used to refer to the difficulties one can face in attempting to solve accurately this innocent-looking problem as "the great embarrassment of numerical analysis." Usually (depending on the precise boundary conditions in (3.29)) the solution u has an *exponential boundary layer* along $\partial^+\Omega$ and *parabolic/characteristic boundary layers* along $\partial^0\Omega$. Exponential layers are, at each point in $\partial^+\Omega$, essentially the same as the function $e^{-(1-x)/\varepsilon}$, but characteristic layers have a much more complicated structure and cannot be defined by an ODE. They nevertheless have the layer quality of fast decay in a narrow region (roughly of width $O(\sqrt{\varepsilon}\,)$ along $\partial^0\Omega$.

There will also be a characteristic layer in the interior of Ω emanating from each point of discontinuity in the boundary conditions on $\partial^-\Omega$ (think how a discontinuity would be propagated across Ω by $\boldsymbol{\beta} \cdot \nabla u = f$; the effect of the diffusion term $-\varepsilon\Delta u$ is to smooth this discontinuity into a continuous but steep layer).

For some illustrations of how solutions to such problems can look, see, e.g., [14, 15].

4 Numerical instability

Standard numerical approximations of differential equations use a central difference approximation of the convective term. That is, for ODEs, one approximates $u'(x_i)$ by $(u_{i+1}^N - u_{i-1}^N)/(2h)$ in the usual notation, where h is the local mesh-width. On quasiuniform meshes this yields oscillatory and inaccurate solutions; see, e.g., [6, Fig. 5].

One can give many indirect explanations of this poor performance. For example, a careful inspection of the analysis of standard numerical methods reveals an assumption that the diffusion coefficient is bounded away from 0; but in (3.29) the parameter ε can be very small, and this means that the standard analysis is no longer valid. An alternative explanation: the matrices generated by the approximations of the differential operator are not M-matrices when ε is small relative to the local mesh-width, so their inverses can be expected to have both positive and negative entries. As a consequence the computed solution will display oscillations.

One means of eliminating oscillations is to approximate the convective derivative by a non-centered approximation called *upwinding*: when solving $-\varepsilon u'' + u' = f$ on a uniform mesh of width h, replace

$$u'(x_i) \mapsto (u_{i+1}^N - u_{i-1}^N)/(2h)$$

by

$$u'(x_i) \mapsto (u_i^N - u_{i-1}^N)/h,$$

while discretizing $-\varepsilon u''(x_i)$ in the usual way. (Here $\{u_i^N\}_{i=0}^N$ is the computed solution.) That is, the approximation of $u'(x_i)$ uses values of u_i^N that are chosen away from the boundary layer at $x = 1$. It is easy to check that this modified discretization of $-\varepsilon u'' + u'$ yields an M-matrix; consequently the computed solution is more stable and no longer has non-physical oscillations.

While the unwanted oscillations have disappeared, this has come at a price: layers in the computed solution are excessively smeared, i.e., are not as steep as they should be. See, e.g., [6, Fig. 5]. To motivate a way of addressing this shortcoming, we observe that on a uniform mesh of width h, upwinding yields

$$\begin{aligned}(-\varepsilon u'' + u')(x_i) &\mapsto \frac{-\varepsilon}{h^2}(u_{i+1}^N - 2u_i^N + u_{i-1}^N) + \frac{1}{h}(u_i^N - u_{i-1}^N) \\ &= -\left(\varepsilon + \frac{h}{2}\right)\frac{1}{h^2}(u_{i+1}^N - 2u_i^N + u_{i-1}^N) + \frac{1}{2h}(u_{i+1}^N - u_{i-1}^N).\end{aligned}$$

That is, upwinding applied to $-\varepsilon u'' + u'$ is the same method as standard central differencing applied to $-(\varepsilon + h/2)u'' + u'$. To put this in words, we can regard upwinding as the standard discretization of a modified differential equation—modified by artificially increasing the diffusion coefficient by $h/2$.

Now we see the possibility of modifying the diffusion coefficient by some other quantity before applying a standard numerical method, with the aim of retaining stability while introducing less smearing of layers in the computed solution. This way of thinking turns out to be quite fruitful; in fact, stable numerical methods on uniform meshes for convection-diffusion ODEs are usually equivalent to modifying the diffusion in the original differential equation then applying a standard method (e.g., central differencing)—but for PDEs, the connection may be less straightforward.

Summary: when a standard numerical method is applied to a convection-diffusion problem, if there is too little diffusion, then the computed solution is oscillatory, while if there is too much diffusion, the computed layers are smeared.

One can add artificial diffusion using finite difference, finite element or finite volume methods. See [19] for many illustrations of how this can be done. In the rest of this talk, I shall discuss a few well-known techniques for the numerical solution of convection-diffusion problems that operate in this way.

5 1969–early 1990s: the international phase

Our history of numerical methods for convection-diffusion problems begins about 30 years ago, in 1969. In this year, two significant Russian papers [3, 7] analysed new numerical methods for convection-diffusion ODEs.

In [3], Bakhvalov considered an upwinded difference scheme on a layer-adapted graded mesh. Such meshes are based on a logarithmic scale (the inverse of the exponential layer function that we met in (3.30)). They are very fine inside the boundary layer and coarse outside. The fineness of the mesh means that the added artificial diffusion is very small inside the layer, and consequently the layer is not smeared excessively.

We shall return later to Bakhvalov's idea, as initially it was less influential than [7], where A.M.Il'in used a uniform mesh but chose the amount of added artificial diffusion in such a way that for constant-coefficient ODEs the computed solution agrees exactly with the true solution at the meshpoints. The amount of artificial diffusion involves exponentials, and schemes of this type are called *exponentially-fitted* difference schemes. See [18] or [19] for details of the scheme. (In fact the same scheme had been used much earlier in [1] but no analysis of its behaviour was given there.)

During the next 20 years, researchers from many countries developed Il'in-type schemes for many singularly perturbed ODEs and some PDEs. See [19] for references; here we just mention Griffiths and Mitchell from Dundee.

The original Il'in paper used a complicated technique called the "double-mesh principle" to analyse the difference scheme. This became obsolete overnight when in 1978 Kellogg and Tsan published a revolutionary and famous paper [9] that was gratefully seized on by other researchers in the area. Their paper showed how to design *barrier* or *comparison functions* to convert truncation errors to computed errors, and also gave for the first time sharp a priori estimates for the solution of the convection-diffusion ODE. (Historical note: it was the first of many papers by Bruce Kellogg on convection-diffusion problems, and it was the only mathematical paper that Alice Tsan ever wrote!)

Today exponential fitting is still used for instance in the well-known package PLTMG and in semiconductor device modelling (where it's known as the Scharfetter-Gummel scheme). In [2], Angermann gives an example of an exponentially-fitted scheme that does a remarkable job of capturing an interior layer on a uniform mesh.

A related idea is the residual-free bubbles FEM that has been developed and analysed in recent years by Brezzi, Franca, et al. This doesn't explicitly contain exponentials, but it is based on the idea of solving a local problem exactly [5], as is Il'in's method [18].

6 1979–mid 1990s: the Swedish phase

The work described in the previous section is finite difference in nature. In 1979, Hughes and Brooks [6] introduced the *streamline diffusion finite element method* (SDFEM) for convection-diffusion problems. This kick-started a development of finite element methods for convection-diffusion problems that continues to this day. Many researchers have participated in this effort, but from the point of view of analysis, the dominant character has been Claes Johnson from Sweden. See [19] for an overview of the relevant literature.

The SDFEM (also known as SUPG) works as follows. Consider the PDE (3.29) with homogeneous Dirichlet boundary conditions, and Ω a convex bounded subset of R^2. Assume that $\boldsymbol{\beta}$ is constant with $|\boldsymbol{\beta}| = 1$. Write for convenience $\boldsymbol{\beta} \cdot \nabla u \equiv u_\beta$.

Suppose we have a triangular mesh on Ω. We'll discuss only piecewise linears here, but there is an analogous, slightly more complicated method for piecewise polynomials of higher degree (see [19]). For the piecewise linear trial space $V^N \subset H_0^1(\Omega)$, the SDFEM is: find $u^N \in V^N$ such that

$$\varepsilon(\nabla u^N, \nabla v^N) + ((u^N)_\beta, v^N + \delta v^N_\beta) = (f, v^N + \delta v^N_\beta)$$

for all $v \in V^N$, where δ is a user-chosen locally constant parameter with $\delta \geq 0$; typically $\delta = O(\text{local mesh diameter})$

This is roughly equivalent to altering the PDE from $-\varepsilon\Delta u + u_\beta = f$ to $-\varepsilon\Delta u - \delta u_{\beta\beta} + u_\beta = f$, then applying the standard Galerkin FEM—i.e., *we have added artificial diffusion only in the streamline/flow direction.* This stabilizes the SDFEM and can remove the outflow-layer oscillations one would obtain if the standard Galerkin FEM were applied directly to (3.29); moreover, as diffusion is added only in the direction of flow, one does not smear characteristic layers.

How exactly should one choose δ? No "optimal" formula is known. Different choices introduce different amounts of diffusion. In [14, Figs. 2, 3] one sees the striking effect of different choices of δ on the sharpness of computed outflow boundary layers.

In practice the SDFEM yields accurate solutions away from layers and local error estimates of Johnson, Schatz and Wahlbin, and Niijima, reflect this behaviour. See [19]. On subdomains Ω_0 of Ω that lie "away from" layers,

$$|(u - u^N)(x)| \leq Ch^{11/8} \ln(1/h) \; ||u||_{C^2(\Omega_0)}.$$

This is almost sharp: numerical results of Zhou imply that in general $O(h^{3/2})$ is the best possible bound for piecewise linears.

But the stabilization of the SDFEM has little effect along characteristic layers. Kopteva [11] shows that one obtains only $O(\delta)$ pointwise accuracy inside parabolic boundary and interior layers, and as $\delta = O(h)$ typically, this means one can at best get first-order convergence inside characteristic layers, even on special meshes.

The idea of the SDFEM has generated several related FEMs for convection-diffusion problems: the Galerkin least-squares FEM, negative-norm stabilization of the FEM, and the currently popular discontinuous Galerkin FEM.

7 1990–present: the Russian-Irish phase

Recall from Section 5 that Bakhvalov-type graded meshes can be used to solve convection-diffusion problems. In 1990 the Russian mathematician Grisha Shishkin showed that instead one could use a simpler piecewise uniform mesh. This idea has been enthusiastically propagated throughout the 1990s by a group of Irish mathematicians: Miller, O'Riordan, Hegarty and Farrell. See [4, 19] and their bibliographies.

The Shishkin mesh is chosen a priori. It is very fine near layers but coarse otherwise. For example, if the domain Ω is the unit square and the problem is

$$-\varepsilon \Delta u + b_1 u_x + b_2 u_y = f, \quad \text{with } b_1 > 0, b_2 > 0,$$

so one has boundary layers along $x = 1$ and $y = 1$, then this tensor-product mesh has transition points (where the mesh switches from coarse to fine) at $1 - \lambda_x$ and $1 - \lambda_y$ on the x- and y-axes respectively, where $\lambda_x = (4\varepsilon/b_1) \ln N$ and $\lambda_y = (4\varepsilon/b_2) \ln N$. Here N is the number of mesh points in each coordinate direction. The fine and coarse mesh regions on the coordinate axes each contain $N/2$ mesh intervals. See Figure 1 for the mesh with $N = 8$ (the mesh rectangles have been bisected into triangles to permit the use of a piecewise linear FEM). Shishkin and his coauthors favour the use of upwinding (see Section 4) on this mesh. Since the mesh is fine at the boundary layers, upwinding does not smear these layers. The computed solution has no non-physical oscillations, and one usually obtains almost first-order (i.e., up to a factor $\ln N$) pointwise convergence at the mesh nodes. Unlike the SDFEM, one does not have to manage a free parameter. The computed solutions look satisfactory [4].

One can of course instead use a FEM on a Shishkin mesh. Linß and Stynes [13] give numerical results for linears and bilinears on these meshes, and show that bilinears are more accurate in the layer regions. Further theoretical evidence that bilinears are superior to linears is given in [20].

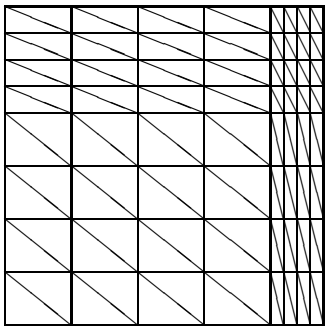


Figure 7.12: Shishkin mesh with $N = 8$

The drawbacks to Shishkin meshes are that one must know the location and nature of the layers a priori, and up to now the method has been implemented only on rectangular domains. Curved interior layers have not been tested numerically using exact Shishkin meshes, but in [15] an interior layer is computed accurately using PLTMG and a simple heuristic approximation of a Shishkin mesh. Furthermore, the analysis of schemes on these meshes requires strong assumptions on the data of the problem to ensure sufficient differentiability of solution u and thereby justify the choice of mesh.

An excellent survey of the published literature for layer-adapted meshes (Shishkin, Bakhvalov, etc.) applied to convection-diffusion problems is given by Linß [12].

8 A historical connection

If you read a little about the 30 Years War, you will almost certainly learn that in 1631 a certain large German city was almost completely destroyed during that conflict. The same German city has played a significant role in the development of numerical methods for convection-diffusion problems. I refer to *Magdeburg*.

Late 20th-century mathematicians from Magdeburg who have worked on numerical methods for convection-diffusion problems include *Goering,* Tobiska, Roos, Lube, Felgenhauer, John, Matthies, Risch, Schieweck, ... Herbert Goering was the father of this school; all other names here were his students or his students's students from Magdeburg.

9 Where is our "Peace of Westphalia"?

Can we find a numerical method that is completely satisfactory for all convection-diffusion problems? The general consensus seems to be that in the future we will use adaptive meshes based on a posteriori error indicators. Unfortunately the development of this theory for convection-diffusion problems is only beginning; John [8] numerically investigates several standard a posteriori error indicators, and concludes that all are unsatisfactory to varying degrees. Consequently I have not discussed this promising line of attack in my talk.

To conclude, I would like to offer to young researchers embarking on a study of convection-diffusion problems some advice drawn from my own experience : always try the easiest

case first—it may be harder than you expect!

References

[1] D.N.de G. Allen and R.V. Southwell, Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. Appl. Math. 8 (1955), 129–145.

[2] L. Angermann, A finite element method for the numerical solution of convection-dominated anisotropic diffusion equations. Numer. Math. 85 (2000), 175–195.

[3] N.S. Bakhvalov, On the optimization of the methods for solving boundary value problems in the presence of a boundary layer. (Russian) Ž. Vyčisl. Mat. i Mat. Fiz. 9 (1969), 841–859.

[4] P. A. Farrell, A. F. Hegarty, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Robust Computational Techniques for Boundary Layers*, volume 16 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall/CRC Press, Boca Raton, U.S.A., 2000.

[5] L.P. Franca, A. Nesliturk and M. Stynes, On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method. Comp. Methods Appl. Mech. Engrg. 166 (1998), 35–49.

[6] T.J.R. Hughes and A. Brooks, A multidimensional upwind scheme with no crosswind diffusion. Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979), AMD, 34, Amer. Soc. Mech. Engrs. (ASME), New York (1979), 19–35.

[7] A.M. Il'in, A difference scheme for a differential equation with a small parameter multiplying the highest derivative. (Russian) Mat. Zametki 6 (1969), 237–248.

[8] V. John, A numerical study of a posteriori error estimators for convection-diffusion equations. Comput. Methods Appl. Mech. Engrg. 190 (2000), 757–781.

[9] R.B. Kellogg and A. Tsan, Analysis of some difference approximations for a singular perturbation problem without turning points. Math. Comp. 32 (1978), 1025–1039.

[10] H. Kinder and W. Hilgemann, *The Penguin Atlas of World History Vol. 1.* Penguin Books, London, 1974.

[11] N. Kopteva, How accurate is the streamline-diffusion FEM inside characteristic (boundary and interior) layers? Preprint No. 3 (2003), School of Mathematics, Applied Mathematics and Statistics, National University of Ireland, Cork, 2003 (submitted for publication).

[12] T. Linß, Layer-adapted meshes for convection-diffusion problems. Comput. Methods Appl. Mech. Engrg. 192 (2003), 1061–1105.

[13] T. Linß and M. Stynes, Numerical methods on Shishkin meshes for linear convection-diffusion problems. Comput. Methods Appl. Mech. Engrg. 190 (2001), 3527–3542.

[14] N. Madden and M. Stynes, Linear enhancements of the streamline diffusion method for convection-diffusion problems. Computers Math. Applic. 32 (1996), 29–42.

[15] N. Madden and M. Stynes, Efficient generation of oriented meshes for solving convection-diffusion problems. Int. J. Numer. Methods Engrg. 40 (1997), 565–576.

[16] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Fitted Numerical Methods For Singular Perturbation Problems – Error Estimates In The Maximum Norm For Linear Problems In One And Two Dimensions*. World Scientific, 1996.

[17] K.W. Morton, *Numerical Solution of Convection-Diffusion Problems*. Chapman & Hall, London, 1996.

[18] H.-G. Roos, Ten ways to generate the Il'in and related schemes. J. Comput. Appl. Math. 53 (1994), 43–59.

[19] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Vol. 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1996.

[20] M. Stynes and L. Tobiska, The SDFEM for a convection-diffusion problem with a boundary layer: optimal error analysis and enhancement of accuracy, SIAM J. Numer. Anal. (to appear).

Contributed Papers

On Wilkinson's Problem
Rafikul Alam (The University of Manchester, Department of Mathematics, Oxford Road, Manchester M13 9PL, UK)
with S Bora
ralam@maths.man.ac.uk

Qualitative behaviour of splitting methods for advection-reaction equations
Adérito Araújo (University of Coimbra, Department of Mathematics, Largo D. Dinis, Coimbra 3000, Portugal)
with J A Ferreira & P Oliveira;
alma@mat.uc.pt

Incomplete orthogonal distance regression
Abdullah Atieg (University of Dundee, Mathematics Division, Dundee DD1 4HN, UK)
with G A Watson
aatieg@maths.dundee.ac.uk

Advection on irregulars grids
Mohammed Babatin (University of Dundee, Mathematics Division, Dundee DD1 4HN, UK)
with D F Griffiths
mbabatin@maths.dundee.ac.uk

Lagrange-remap methods for the Euler Equations
David Bailey (Reading University, Department of Mathematics, PO BOX 220, Reading RG6 6AX, UK)
with P Glaister & P K Sweby
smr00dab@reading.ac.uk

Superconvergence in a fully discrete linear finite element method
Silvia Barbeiro (University of Coimbra, Department of Mathematics, Portugal)
with J A Ferreira & R D Grigorieff
silvia@mat.uc.pt

Convection-diffusion and differential quadrature
Ken Barrett (Coventry University, Department of Mathematics, Priory St, Coventry CV1 5FB, UK)
k.barrett@coventry.ac.uk

Robust implementation of the HZ algorithm for the tridiagonal-diagonal eigenvalue problem
Michael Berhanu (University of Manchester, Department of Mathematics, Oxford Road, Manchester M13 9PL, UK)
with F Tisseur
mberhanu@ma.man.ac.uk

Improving spectral methods with optimized rational interpolation
Jean-Paul Berrut (University of Fribourg, Department of Mathematics, Pérolles, CH-1700 Fribourg, Switzerland)
with H D Mittelmann
jean-paul.berrut@unifr.ch

A global approximation method for eigenvalues on polygons and applications
Timo Betcke (Oxford University, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK)
with L N Trefethen
timo.betcke@comlab.ox.ac.uk

Finite volume WENO schemes for integro-differential equations
Barna Bihari (Lawrence Livermore National Laboratory, Center for Applied Scientific Computing, P.O. Box 808, L-560, Livermore 94551, USA)
with P N Brown
bihari@llnl.gov

Updating least squares solutions
Craig Brand (University of Strathclyde, Department of Mathematics, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK)
with P Knight
ta.cbra@maths.strath.ac.uk

Preconditioners for anisotropic problems in spherical geometry
David Brown (University of Reading, Department of Mathematics, Reading RG6 6AX, UK)
with N Nichols & M Bell
smr00deb@reading.ac.uk

On the divergence of collocation solutions for Volterra equations in smooth spaces
Hermann Brunner (Memorial University of Newfoundland, Department of Mathematics & Statistics, St. John's, NL A1C 5S7, Canada)
hermann@math.mun.ca

Computing invariant measures for stochastic differential equations
Alan Bryden (University of Strathclyde, Department of Mathematics, 26 Richmond Street, Glasgow G1 1XH, UK)
with D J Higham
ta.abry@maths.strath.ac.uk

A Reynolds–uniform numerical method for Prandtl's boundary layer problem for flow past a plate with mass transfer
John Butler (Trinity College Dublin, Maths Department, Dublin 2, Ireland)
with J J H Miller & G I Shishkin
butler@maths.tcd.ie

Numerical algorithm for solving phase conjugation problem
Raimondas Ciegis (Vilnius Gediminas Technical University, Department of Mathematical Modelling, Sauletekio 11, Vilnius LT-2040, Lithuania)
rc@fm.vtu.lt

Structured conditioning of matrix functions
Philip Davies (University of Manchester, Department of Mathematics, Oxford Road, Manchester M13 9PL, UK)
pdavies@maths.man.ac.uk

Convergence of collocation methods for time domain boundary integral equations
Penny Davies (University of Strathclyde, Department of Mathematics, 26 Richmond Street, Glasgow G1 1XH, UK)
with D B Duncan
penny@maths.strath.ac.uk

A modified Gram–Schmidt algorithm with iterative orthogonalization and column pivoting
Achiya Dax (Hydrological Service, Research Department, P.O.B. 36118, Jerusalem 91360, Israel)
dax@vms.huji.ac.il

Approximate solutions to a singular Volterra integral equation
Teresa Diogo (CEMAT/Instituto Superior Técnico, Department of Mathematics, Av. Rovisco Pais, 1, Lisbon 1049-001, Portugal)
with Pedro Lima, Neville Ford and Svilen Valtchev
tdiogo@math.ist.utl.pt

Overlapping grids for the heat equation
Dugald Duncan (Heriot-Watt University, Department of Mathematics, Riccarton, Edinburgh EH14 4AS, UK)
with Yiqi Qiu
d.b.duncan@ma.hw.ac.uk

Evaluating singular integrals using generalised quadrature
Gwynne Evans (De Montfort University, Dept of Computer Studies, James Went Building, The Gateway, Leicester LE1 9BH, UK)
gaevans@dmu.ac.uk

A constrained steady vortex flow in two dimension
Rahman Farnoosh (Iran University of Science and Technology, Department of Mathematics, Narmak, Tehran 16844, Iran)
rfarnoosh@iust.ac.ir

On the asymptotics of some new gradient methods
Roger Fletcher (University of Dundee, Mathematics Division, Dundee DD1 4HN, UK)
with Y-H Dai
fletcher@maths.dundee.ac.uk

A numerical remedy for time-space corner singularities
Natasha Flyer (National Center for Atmospheric Research, Division of Scientific Computing, P.O. Box 3000, Boulder, CO 80307, USA)
with B Fornberg
flyer@colorado.edu

Discrete wavelet transform preconditioning techniques
Judith Ford (UMIST, Department of Mathematics, PO Box 88, Manchester M60 1QD, UK)
j.ford@umist.ac.uk

Non–integrable resolvents and qualitative behaviour of solutions to integral equations
Neville Ford (Chester College, Department of Mathematics, Parkgate Road, Chester CH1 4BJ, UK)
njford@chester.ac.uk

High–order time stepping of ADI solutions to Maxwell's equations
Bengt Fornberg (University of Colorado, Department of Applied Mathematics, 526 UCB, Boulder, CO 80309, USA)
with Jongwoo Lee
fornberg@colorado.edu

A hybrid optimisation approach for water distribution network design
Eric Fraga (University College London, Department of Chemical Engineering, Torrington Place, London WC1E 7JE, UK)
with L G Papageorgiou
e.fraga@ucl.ac.uk

Space-time discretization of hyperbolic equations using special collocation points
Daniele Funaro (University of Modena, Department of Mathematics, Via Campi 213/B, Modena 41100, Italy)
funaro@unimo.it

Solving a class of nonlinear difference equations by using measure theory
Mortaza Gachpazan (Damghan University of Science, Department of Mathematics, Damghan, Iran)
with Akbar Hashmi Borzabadi
mgachpaz@math.um.ac.ir

Krylov methods and determinants for detecting bifurcations in partial differential equations
Bosco García-Archilla (Universidad de Sevilla, Matemática Aplicada II, Camino de los Descubrimientos, Sevilla 41092, Spain)
with J Sánchez & C Simó
bosco.garcia@esi.us.es

The numerical analysis of a reaction-diffusion system of $\lambda - \omega$ type
Marcus Garvie (University of Durham, Department of Mathematical Sciences, South Rd, Durham DH1 3LE, UK)
with J F Blowey
Marcus.Garvie@durham.ac.uk

An interior-point ℓ_1-penalty method for nonlinear optimization
Nicholas Gould (CCLRC – Rutherford Appleton Lab, Computational Science & Engineering, Chilton, Didcot OX11 0QX, UK)
with D Orban & Ph L Toint
N.I.M.Gould@rl.ac.uk

Object–oriented parallel interior point solver for structured nonlinear programs
Andreas Grothey (University of Edinburgh, School of Mathematics, Edinburgh EH9 3JZ, UK)
with J Gondzio
agr@maths.ed.ac.uk

A non–simplex active–set framework for basis-deficiency-allowing simplex variations
Pablo Guerrero–Garcia (University of Malaga, Dpto. Matematica Aplicada, Complejo Tecnologico, Campus Teatinos, Malaga 29071, Spain)
with A Santos–Palomo
pablito@lcc.uma.es

Hyper–sparsity in the revised simplex method and how to exploit it
Julian Hall (University of Edinburgh, Department of Mathematics, JCMB, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK)
with K I M McKinnon
J.A.J.Hall@ed.ac.uk

Stability of the hyperbolic QR factorization
Gareth Hargreaves (University of Manchester, Department of Mathematics, Oxford Road, Manchester M13 9PL, UK)
hargreaves@ma.man.ac.uk

Effective sparse preconditioners for sparse matrices arising from wavelet discretisations of boundary integral equations
Stuart Hawkins (University of Liverpool, Department of Mathematics, Peach Street, Liverpool L69 7ZL, UK)
with Ke Chen
stuarth@liv.ac.uk

Computing the polar decomposition in matrix groups
Nick Higham (University of Manchester, Department of Mathematics, Manchester M13 9PL, UK)
with D S Mackey, N Mackey & F Tisseur
higham@ma.man.ac.uk

A doubly–adaptive algorithm for globally adaptive numerical integration using Gauss–Kronrod formulas
Helmut Hlavacs (University of Vienna, Department for Computer Science and Bus, Lenaug. 2/8, Vienna 1080, Austria)
with A Kumar, S Raghavan & S Leopold
helmut.hlavacs@univie.ac.at

Necessary and sufficient conditions for positivity of Runge–Kutta methods and their relevance in practical situations
Zoltán Horváth (Széchenyi István University, Department of Mathematics, 1 Egyetem Square, Gyr 9026, Hungary)
horvathz@sze.hu

A moving mesh finite element method for nonlinear diffusion equations
Matthew Hubbard (University of Leeds, School of Computing, Leeds LS2 9JT, UK)
with M J Baines & P K Jimack
meh@comp.leeds.ac.uk

Improved preconditioners for the iterative solution of coupled 3–dimensional fluid–structure interaction problems
Martyn Hughes (University of Liverpool, Department of Mathematical Sciences, M and O Building, Peach Street, Liverpool L69 7L, UK)
with Ke Chen
mdhughes@liv.ac.uk

Phase space error control for adaptive time-stepping ODE solvers
Tony Humphries (McGill University, Department of Mathematics, 805 Sherbrooke West, Montreal, Quebec H3A 2K6, Canada)
tony.humphries@mcgill.ca

Viva Lobatto! Approximating highly–oscillatory integrals is easy...
Arieh Iserles (University of Cambridge, DAMTP, Wilberforce Rd, Cambridge CB3 0WA, UK)
ai@damtp.cam.ac.uk

A reduction of quaternion–valued matrices to upper Hessenberg form
Drahoslava Janovska (Institute of Chemical Technology, Prague, Department of Mathematics, Technicka 5, Prague 166 28, Czech Republic)
with G Opfer
janovskd@vscht.cz

The Runge–Kutta discontinuous Galerkin method for the morphodynamical equations
Paul Jelfs (University Of Reading, Department Of Mathematics, PO Box 220, Reading RG6 6AX, UK)
with P K Sweby
p.jelfs@rdg.ac.uk

Variational data assimilation in numerical weather prediction: a singular vector interpretation
Christine Johnson (University of Reading, Department of Mathematics, WhiteKnights, PO Box 220, Reading RG6 6AX, UK)
with N K Nichols, B J Hoskins, S P Ballard & A S Lawless
c.johnson@reading.ac.uk

High accuracy time stepping for stiff nonlinear PDEs
Aly–Khan Kassam (Oxford University, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK)
with L N Trefethen
akk@comlab.ox.ac.uk

Use of automatic differentiation in Matlab's boundary value solver bvp4c
Rob Ketzscher (Cranfield University, ESD, Applied Mathematics and Operational Research, RMCS, Shrivenham, Swindon SN6 8LA, UK)
with L F Shampine & S A Forth
R.Ketzscher@rmcs.cranfield.ac.uk

Two–point boundary value problems for systems of singularly perturbed nonlinear reaction–diffusion equations: numerical solution
Natalia Kopteva (National University of Ireland Cork, Mathematics Department, Cork, Ireland)
with M Stynes
n.kopteva@ucc.ie

An extrapolation technique for strong solutions of SDEs
Harbir Lamba (George Mason University, Department of Mathematics, 4400 University Drive, MS3F2, Fairfax, Virginia, 22030, USA)
with Ben Crain
hlamba@gmu.edu

Ten thousand digits for Trefethen's Problem #5
Dirk Laurie (University of Stellenbosch, Department of Mathematics, Matieland, Stellenbosch 7602, South Africa)
dpl@sun.ac.za

A Discrete Theory of Connections on Principal Bundles
Melvin Leok (California Institute of Technology, Control and Dynamical Systems, 107-81, CDS, Caltech, Pasadena, CA 91125–81, USA)
mleok@cds.caltech.edu

A faster implementation of the pivoted Cholesky factorization for semidefinite matrices
Craig Lucas (University of Manchester, Mathematics Department, Oxford Road, Manchester M13 9PL, UK)
clucas@maths.man.ac.uk

Automating the detection of small solutions to delay differential equations: Using arguments to argue the case
Patricia Lumb (Chester College of HE, Department of Mathematics, Parkgate Road, Chester CH1 4BJ, UK)
with N J Ford
p.lumb@chester.ac.uk

Uniform convergence of a finite difference scheme for a system of coupled reaction–diffusion equations
Niall Madden (NUI Galway, Department of Mathematics, Galway, Ireland)
with T Linss
niall.madden@nuigalway.ie

A priori error estimates for higher order finite element methods for an equation of mean curvature type
Gunar Matthies (Otto von Guericke University Magdeburg, Department of Mathematics, Magdeburg, Germany)
with L Tobiska
matthies@mathematik.uni-magdeburg.de

An alternate strip–based domain decomposition strategy for elliptic PDE's
L. Angela Mihai (University of Durham, Department of Mathematical Sciences, South Road, Durham DH1 3LE, UK)
with A. W. Craig
l.a.mihai@durham.ac.uk

The state of the art in software for SDP & SOCP problems
Hans Mittelmann (Arizona State University, Department of Mathematics and Statistics, Box 871804, Tempe, AZ 85287–18, USA)
mittelmann@asu.edu

A posteriori error estimators for a finite volume method with a linear parabolic equation
Lionel Nadau (Univerité de Pau et des Pays de l'Adou, Laboratoire de Mathèmatiques Appliquè, I.P.R.A.– L.M.A. Avenue de l'Université, Pau 64000, France)
with M Amara & D Trujillo
lionel.nadau@univ-pau.fr

Newton's method can look ahead
Jelena Nedic (Oxford University, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK)
with R Hauser
jelena@comlab.ox.ac.uk

Convergence of the incremental four dimensional variational method using non–tangent linear models
Nancy K. Nichols (The University of Reading, Department of Mathematics, Box 220 Whiteknights, Reading RG6 6AX, UK)
with A S Lawless
mikeandnancy@greenwood.u-net.com

On the solution of Hamilton–Jacobi equations using adaptive moving meshes
Aurelian Nicola (University of Strathclyde, Department of Mathematics, 26 Richmond Street, Glasgow G1 1XH, UK)
with J MacKenzie
aap00244@maths.strath.ac.uk

How to choose the step size when integrating an ordinary differential equation
Jitse Niesen (University of Cambridge, DAMTP, Wilberforce Road, Cambridge CB3 0WA, UK)
j.niesen@damtp.cam.ac.uk

Local error estimators based on hierarchical bases for the p version of the finite element method
Julia Novo (Universidad Autonoma de Madrid, Departmento de Matematicas, Cantoblanco, Madrid 28049, Spain)
with J de Frutos
julia.novo@uam.es

Convergence of q–Bernstein polynomials
Sofiya Ostrovska (Atilim University, Department of Mathematics, Kizilcasar Koyu, Incek, Ankara 06836, Turkey)
ostrovskasofiya@yahoo.com

Cluster robust error estimates for preconditioned gradient subspace iterations
Evgueni Ovtchinnikov (Harrow School of Computer Science, University of Westminster, London, UK)
e_ovtchinnikov@hotmail.com

A filtered matrix approach to multiwavelet image denoising
Serena Papi (University of Bologna, Department of Mathematics, Piazza di Porta San Donato, Bologna 40126, Italy)
with Silvia Bacchelli
spapi@csr.unibo.it

Diffraction in numerical solution of hyperbolic equations
Manouchehr Parsaei (University of Applied Science and Technology, Department of Information Technology, Tehran, Iran)
parsaei@khayam.ut.ac.ir

Regularization of weakly singular Volterra integral equations by fractional multistep methods
Robert Plato (Technical University of Berlin, Institute of Mathematics, Str. des 17. Juni 135, Berlin 10623, Germany)
plato@math.tu-berlin.de

An iterative algorithm for approximate orthogonalisation of symmetric matrices
Constantin Popa (University of Erlangen–Nurnberg, Lehrstuhl fuer Informatik 10, Cauerstrasse 6, Erlangen D–91058, Germany)
with Marcus Mohr & Ulrich Ruede
cpopa@immd10.informatik.uni-erlangen.de

A matrix factorization for least Frobenius norm updating
Michael Powell (University of Cambridge, DAMTP, CMS Building, Wilberforce Road, Cambridge CB3 0WA, UK)
mjdp@cam.ac.uk

Black–Box preconditioning for Raviart–Thomas formulation of second–order elliptic problems
Catherine Powell (UMIST, Department of Mathematics, PO Box 88, Sackville Street, Manchester M60 1QD, UK)
with D J Silvester
cp@fire.ma.umist.ac.uk

Applications of overlapping grids for welltest analysis
Yiqi Qiu (Edinburgh Petroleum Services Ltd, PanSystem Group, Research Park, Riccarton, Edinburgh EH14 4AP, UK)
with D Duncan & K Hutcheson
yiqi.qiu@e-petroleumservices.com

Bifurcations in numerical methods for Volterra integro–differential equations
Jason Roberts (Chester College of Higher Education, Department of Mathematics, Parkgate Road, Chester CH1 4BJ, UK)
with J T Edwards & N J Ford
j.roberts@chester.ac.uk

A fast convolution algorithm for non–reflecting boundary conditions
Achim Schaedle (Konrad–Zuse–Zentrum Berlin, Numerical Analysis and Modelling, Takustrasse 7, Berlin 14195, Germany)
with Ch Lubich
schaedle@zib.de

Spurious behaviour of a discretised van der Pol equation
Schalk Schoombie (University of the Free State, Department of Mathematics and Applied Ma, PO Box 339, Bloemfontein 9300, South Africa)
with E Mare
SchoomSW.sci@mail.uovs.ac.za

A numerical evaluation of HSL packages for the direct–solution of large sparse, symmetric linear systems
Jennifer Scott (CCLRC – Rutherford Appleton Laboratory, Computational Science & Engineering, Chilton, Didcot OX11 0QX, UK)
with N I M Gould
J.A.Scott@rl.ac.uk

Numerical simulation of dissipative particle dynamics
Tony Shardlow (Manchester University, Department of Mathematics, Oxford Road, Manchester M13 9PL, UK)
shardlow@maths.man.ac.uk

Moving mesh methods for problems with interior and boundary layers
Stephen Sikwila (University of Limerick, Department of Mathematics and Statistics, Limerick, Ireland)
stephen.sikwila@ul.ie

On the norms of inverses of pseudospectral differentiation matrices
David Sloan (University of Strathclyde, Department of Mathematics, Richmond Street, Glasgow G11 1XH, UK)
d.sloan@strath.ac.uk

Vorticity boundary effects on the global time–iteration matrix of discrete unsteady stream–function vorticity equations
Ercilia Sousa (Coimbra University, Department of Mathematics, Apartado 3008, Coimbra 3001–454, Portugal)
with I.J. Sobey
ecs@mat.uc.pt

Multigrid methods, Toeplitz matrices and image deblurring
Jochen Staudacher (University of Strathclyde, Department of Mathematics, 26 Richmond Street, Glasgow G1 1XH, UK)
with T Huckle
ra.jsta@maths.strath.ac.uk

Adaptive Mollifiers
Jared Tanner (University of California Davis, Department of Mathematics, One Shields Avenue, Davis CA 95616, USA)
with E Tadmor
jtanner@math.ucdavis.edu

Moving mesh finite element analysis of convective heat transfer and phase change
Rosen Tenchev (Strathclyde University, Department of Mathematics, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK)
with J MacKenzie
rosen.tenchev@strath.ac.uk

Solving symmetric quadratic eigenvalue problems
Francoise Tisseur (University of Manchester, Department of Mathematics, Manchester M13 9PL, UK)
ftisseur@ma.man.ac.uk

Multiple discretization multilevel method applied to the Stokes Problem
Lutz Tobiska (Otto von Guericke University, Department of Mathematics, Magdeburg, Germany)
with V John, P Knobloch & G Matthies
tobiska@mathematik.uni-magdeburg.de

Exponential fitted Runge–Kutta methods of collocation type: fixed or variable knot points?
Guido Vanden Berghe (Ghent University, Applied Mathematics and Computer Science, Krijgslaan 281–S9, Gent B–9000, Belgium)
with M Van Daele & H Vande Vyver
guido.vandenberghe@rug.ac.be

A high order adaptive collocation software for 1–D parabolic PDEs
Rong Wang (Dalhousie University, Department of Math & Stats, Halifax, NS B3H 3J5, Canada)
with P Keast & P Muir
wang@mathstat.dal.ca

New data storage formats for dense matrices lead to variety of high–performance algorithms
Jerzy Wasniewski (Danish Technical University, Informatics & Mathematical Modeling, IMM, DTU, Bldg. 305, Room number 232, Lyngby DK–2800, Denmark)
with Andersen, Gunnels, Gustavson & Reid
jw@imm.dtu.dk

Robust solutions to a general class of approximation problems
Alistair Watson (University of Dundee, Department of Mathematics, Dundee DD1 4HN, UK)
gawatson@maths.dundee.ac.uk

A matrix approximation problem in algebraic geometry
Joab Winkler (Sheffield University, Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK)
j.winkler@dcs.shef.ac.uk

Adaptive methods for piecewise polynomial collocation for ordinary differential equations
Kenneth Wright (University of Newcastle upon Tyne, Department of Computing Science, Claremont Tower, Claremont Road, Newcastle upon Tyne NE1 7RU, UK)
K.Wright@ncl.ac.uk

Adaptive steplength selections in gradient projection methods for quadratic programs
Gaetano Zanghirati (University of Ferrara, Department of Mathematics, via Machiavelli 35, Ferrara I–44100, Italy)
with T Serafini & L Zanni
g.zanghirati@unife.it