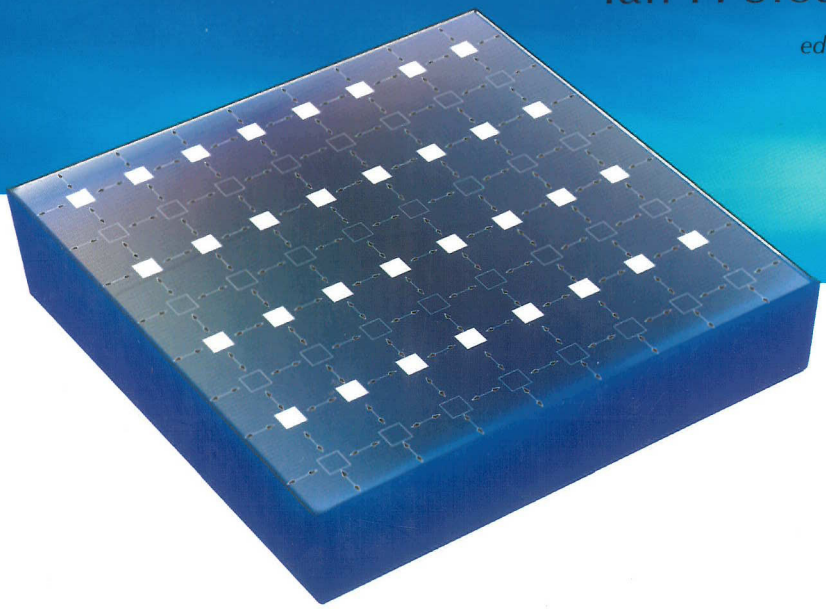
Series in Contemporary Applied Mathematics
CAM 8

Some Topics in Industrial and Applied Mathematics

Rolf Jeltsch
Ta-Tsien Li
Ian H Sloan
editors



Some Topics in Industrial and Applied Mathematics

Series in Contemporary Applied Mathematics CAM

Honorary Editor: Chao-Hao Gu (*Fudan University*)
Editors: P. G. Ciarlet (*City University of Hong Kong*),
Tatsien Li (*Fudan University*)

---

1. Mathematical Finance —— Theory and Practice
   (Eds. Yong Jiongmin, Rama Cont)
2. New Advances in Computational Fluid Dynamics
   —— Theory, Methods and Applications
   (Eds. F. Dubois, Wu Huamo)
3. Actuarial Science —— Theory and Practice
   (Eds. Hanji Shang, Alain Tosseti)
4. Mathematical Problems in Environmental Science and Engineering
   (Eds. Alexandre Ern, Liu Weiping)
5. Ginzburg-Landau Vortices
   (Eds. Haïm Brezis, Tatsien Li)
6. Frontiers and Prospects of Contemporary Applied Mathmetics
   (Eds. Tatsien Li, Pingwen Zhang)
7. Mathematical Methods for Surface and Subsurface Hydrosystems
   (Eds. Deguan Wang, Christian Duquennoi, Alexandre Ern)
8. Some Topics in Industrial and Applied Mathematics
   (Eds. Rolf Jeltsch, Tatsien Li, Ian Hugh Sloan)

Series in Contemporary Applied Mathematics CAM 8

Some Topics in Industrial and Applied Mathematics

editors

Rolf Jeltsch
ETH Zurich, Switzerland

Ta-Tsien Li
Fudan University, China

Ian H Sloan
University of New South Wales, Australia

Higher Education Press

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Rolf Jeltsch
Seminar of Applied Mathematics
ETH Zürich
CH-8092 Zürich
Switzerland

Tatsien Li
School of Mathematical Sciences
Fudan University
220, Handan Road
Shanghai, 200433
China

Ian Hugh Sloan
School of Mathematics
University of New South Wales
Sydney NSW 2052
Australia

图书在版编目 (CIP) 数据

工业与应用数学中的一些问题=Some Topics in Industrial and Applied Mathematics: 英文 / (瑞士) 杰尔奇 (Jeltsch, R.), 李大潜, (澳) 斯隆(Sloan, I. H.) 主编. —北京：高等教育出版社, 2007.7
(现代应用数学丛书)
ISBN 978-7-04-021903-6

I.工··· II.①杰···②李···③斯··· III.①工业工程—应用数学—英文 IV. TB11

中国版本图书馆 CIP 数据核字 (2007) 第 098434 号

Copyright © 2007 by
Higher Education Press
4 Dewai Dajie, Beijing 100011, P. R. China, and
World Scientific Publishing Co Pte Ltd
5 Toh Tuch Link, Singapore 596224

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.

ISBN 978-7-04-021903-6

Printed in P. R. China

Preface

On the occasion that the Officers' Meeting and the Board Meeting of ICIAM (International Council for Industrial and Applied Mathematics) was held in Shanghai from May 26 to May 27, 2006, many famous industrial and applied mathematicians gathered in Shanghai from different countries. The Shanghai Forum on Industrial and Applied Mathematics was organized from May 25 to May 26, 2006 at Shanghai Science Hall for the purpose of inviting some of them to present their recent results and discuss recent trends in industrial and applied mathematics. Sixteen invited lectures have been given for this activity. This volume collects the material covered by most of these lectures. It will be very useful for graduate students and researchers in industrial and applied mathematics.

The editors would like take this opportunity to express their sincere thanks to all the authors in this volume for their kind contribution. We are very grateful to the Shanghai Association for Science and Technology (SAST), Fudan University, the National Natural Science Foundation of China (NSFC), The China Society for Industrial and Applied Mathematics (CSIAM), the Shanghai Society for Industrial and Applied Mathmatics (SSIAM), the Institut Sino-Français de Mathématiques Appliquées (ISFMA) and the International Council for Industrial and Applied Mathematics (ICIAM) for their help and support. Our special thanks are also due to Mrs. Zhou Chunlian for her efficient assistance in editing this book.

Rolf Jeltsch, Ta-Tsien Li, Ian H. Sloan

April 2007

This page intentionally left blank

Contents

Preface

Samir Adly, Daniel Goeleven, Michel Théra: A Continuation Method for a Class of Periodic Evolution Variational Inequalities 1

Götz Alefeld: Complementarity Problems: An Overview on Existing Verification Procedures 29

Fernando Antoneli, Ana Paula S. Dias, Martin Golubitsky, Yunjiao Wang: Synchrony in Lattice Differential Equations 43

Alain Damlamian: The Periodic Unfolding Method for Quasi-convex Functionals 57

Iain S. Duff: Sparse System Solution and the HSL Library 78

Heinz W. Engl, Gerald Goekler, Andrea Schatz, Helmut Zeisel: Modelling and Numerics for the Transient Simulation of the Blast Furnace Process 95

Loredana Faienza, Antonio Fasano, Mario Primicerio: Gelification of Hydrocarbons: A Model Problem 120

Nguyen Van Huu, Vuong Quan Hoang: On the Martingale Representation Theorem and on Approximate Hedging a Contingent Claim in the Minimum Deviation Square Criterion 134

Rolf Jeltsch, Manuel Torrilhon: Flexible Stability Domains for Explicit Runge-Kutta Methods 152

Taketomo Mitsui, Yoshihiro Saito: MS-Stability Analysis for Numerical Solutions of Stochastic Differential Equations —Beyond Single-step Single Dim— 181

R. D. Russell, J. F. Williams, X. Xu: A Comparison of Direct Discretization of 4th Order Problems Versus System Reduction 195

Ya-xiang Yuan: Subspace Techniques for Nonlinear Optimization 206

This page intentionally left blank

A Continuation Method for a Class of Periodic Evolution Variational Inequalities

Samir Adly
DMI-XLIM, UMR-CNRS 6172, Université de Limoges
123 Avenue A. Thomas, 87060 Limoges Cedex, France.
E-mail: samir.adly@unilim.fr

Daniel Goeleven
IREMIA, Université de La Réunion
97400 Saint-Denis, France.
E-mail: goelven@univ-reunion.fr

Michel Théra
DMI-XLIM, UMR-CNRS 6172, Université de Limoges
123 Avenue A. Thomas, 87060 Limoges Cedex, France.
E-mail: michel.thera@unilim.fr

Abstract

In this paper, using the Brouwer topological degree, we prove an existence result for finite variational inequalities. This approach is also applied to obtain the existence of periodic solutions for a class of evolution variational inequalities.

2000 Mathematics Subject Classification 49J40, 49J20, 35K85

Key words and phrases Variational inequalities, differential inclusions, topological degree, guiding functions, periodic solutions

1 Introduction

It has been well recognised that variational inequalities offer the right framework to consider numerous applied problems in various areas such as economics and engineering. Throughout the paper we consider $\mathbb{R}^n$ equipped with the usual Euclidean scalar product $\langle \cdot, \cdot \rangle$. We start by considering a variational inequality $\mathrm{VI}(\Lambda, \varphi)$ that is the problem of finding $\overline{x} \in \mathbb{R}^n$ such that:

$$\langle \Lambda(\overline{x}), v - \overline{x} \rangle + \varphi(v) - \varphi(\overline{x}) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

In this formulation, $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous map, $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex function. When the operator Λ under consideration is supposed to be coercive, existence results for the problem $\mathrm{VI}(\Lambda, \varphi)$ are well known in the setting of reflexive Banach spaces. This study was initiated by G. Stampacchia in the 60's and we refer to the contributions of J.L. Lions [15], Brézis [4] and Kinderlehrer & Stampacchia [12] for various results and references therein.

In the first part of this paper we develop an original approach essentially based on the use of the Brouwer topological degree to prove some results related to the existence of a solution to problem $\mathrm{VI}(\Lambda, \varphi)$.

Then, we study a first order evolution variational inequality, that is a differential inclusion of the form: find a T-periodic function $u \in C^0([0,T];\mathbb{R}^n)$ such that:

$$\frac{du}{dt}(t) + F(u(t)) - f(t) \in -\partial\varphi(u(t)), \text{ a.e. } t \in [0,T],$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous map, $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $f \in C^0([0,+\infty);\mathbb{R}^n)$ is such that: $\frac{df}{dt} \in L^1_{\mathrm{loc}}(0,+\infty;\mathbb{R}^n)$, $T > 0$ is a prescribed period and $\partial\varphi$ is the convex subdifferential operator. This problem is studied by means of a continuation method. It is well known that the Brouwer topological degree plays a fundamental role in the theory of ordinary differential equations (ODE). M.A. Krasnosel'skii [13], [14] and H. Amann [3], developed a continuation method to compute this Brouwer topolgical degree associated to some gradient mapping (called the method of guiding function). This approach was useful for the study of the existence of periodic solutions for ODE's. Roughly speaking, if on some balls of $\mathbb{R}^n$ the Brouwer topological degree of the Poincaré translation operator (see e.g. [17]) associated to the ODE is different from zero, the problem has at least one periodic solution (for more details, references and possible extensions to the Leray-Schauder degree, we refer to the monograph of J. Mawhin [17]). With the emergence of many engineering disciplines and due to the lack of smoothness in many applications, it is not surprising that these classical mathematical tools require natural extension (for both analytical and numerical methods) to the class of unilateral dynamical systems. It is well known that the mathematical formulation of unilateral dynamical systems involves inequality constraints and hence contains natural non-smoothness. In mechanical systems, this non-smoothness could have its origin in the environment of the system studied (e.g. case of contact) in the dry friction, or in the discontinuous control term. Recently, new analytical tools have been developed for the study of unilateral evolution problems (see e.g. [1], [2], [7], [8], [9] and references cited therein). The study of periodic solutions for evolution variational inequalities is also important. The Krasnol'skii's original approach for ODE, has known

some extensions in order to obtain continuation methods for differential inclusions (see the article of L. Górniewicz [10] for more details and references). In the fourth section, we will be concerned with the existence of a T-periodic solution $u \in C^0([0,T];\mathbb{R}^n)$ such that:

$$\begin{gathered}\tfrac{du}{dt} \in L^\infty(0,T;\mathbb{R}^n);\\ u \text{ is right-differentiable on } [0,T);\\ u(0) = u(T);\\ \langle \tfrac{du}{dt}(t) + F(u(t)) - f(t), v - u(t)\rangle + \varphi(v) - \varphi(u(t)) \geq 0,\\ \forall v \in \mathbb{R}^n, \text{ a.e. } t \in [0,T]. \end{gathered} \tag{1}$$

Here $F : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous map, $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $f \in C^0([0,+\infty[;\mathbb{R}^n)$ is such that: $\frac{df}{dt} \in L^1_{\text{loc}}(0,+\infty;\mathbb{R}^n)$ and $T > 0$ is a prescribed period.

We prove (Corollary 5.1) that if F and φ satisfy some growth condition (see (36)), then problem (1) has at least one periodic solution. This approach is also applied to obtain the existence of a T-periodic solution of a second order dynamical system of the form:

$$M\ddot{q}(t) + C\dot{q}(t) + Kq(t) - F(t) \in -H_1\partial\Phi(H_1^T\dot{q}(t)), \tag{2}$$

where $q \in \mathbb{R}^m$ is the vector of generalized coordinate, $\Phi : \mathbb{R}^l \to \mathbb{R}$ is a convex function, $M \in \mathbb{R}^{m\times m}$ is a symmetric and positive definite matrix, $C \in \mathbb{R}^{m\times m}$ and $K \in \mathbb{R}^{m\times m}$ are given matrices and $H_1 \in \mathbb{R}^{m\times l}$ is a given matrix whose coefficients are determined by the directions of friction forces. The function $F \in C^0([0,+\infty);\mathbb{R}^m)$ is such that $\frac{dF}{dt} \in L^1_{\text{loc}}([0,+\infty);\mathbb{R}^m)$. The term $H_1\partial\Phi(H_1^T\cdot)$ is used to model the convex unilateral contact induced by friction forces. The paper is organized as follows: Section 2 contains some background materials on properties of the Brouwer topological degree and the concept of resolvent operator associated to a subdifferential operator. In Section 3, using an equivalent fixed point formulation as well as the Brouwer topological degree, we give some existence results for finite variational inequalities. In Section 4, we introduce the Poincaré operator associated to problem (1). Section 5 is devoted to the existence of a periodic solution of problem (1). In Section 6, we show that our approach could be applied to a special second order problem (2).

2 Brouwer topological degree and the resolvent operator $P_{\lambda,\varphi}$

It is well known that the degree theory is one of the most powerful tool in nonlinear analysis for the study of zeros of a continuous operator.

Let $\Omega \subset \mathbb{R}^n$ be an open and bounded subset with boundary $\partial\Omega$ and $f \in C^1(\Omega;\mathbb{R}^n) \cap C^0(\overline{\Omega},\mathbb{R}^n)$. The Jacobian matrix of f at $x \in \Omega$ is defined by $f'(x) = (\partial_{x_i} f_j(x))_{1\leq i,j\leq n}$ and the Jacobian determinant of f at $x \in \Omega$ is defined by

$$J_f(x) = \det(f'(x)).$$

We set

$$A_f(\Omega) = \{x \in \Omega \ : \ J_f(x) = 0\}.$$

Observing that if $f^{-1}(0) \cap A_f(\Omega) = \emptyset$ and $0 \notin f(\partial\Omega)$, then the set $f^{-1}(0)$ is finite. The quantity $\sum_{x\in f^{-1}(0)} \text{sign}\,(J_f(x))$ is therefore defined and is called the *Brouwer topological degree* of f with respect to Ω and 0 and is denoted by $\deg(f,\Omega,0)$. More generally, if $f \in C^0(\overline{\Omega};\mathbb{R}^n)$ and $0 \notin f(\partial\Omega)$, then the Brouwer topological degree of f with respect to Ω and 0, denoted by $\deg(f,\Omega,0)$, is well defined (see [16] for more details).

In the sequel, the scalar product on $\mathbb{R}^n$ is denoted as usual by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ the associated norm. For $r > 0$, we set $\mathbb{B}_r := \{x \in \mathbb{R}^n : \|x\| < r\}$, and respectively $\bar{\mathbb{B}}_r = \{x \in \mathbb{R}^n : \|x\| \leq r\}$, for the open (respectively closed) unit ball with radius $r > 0$. As usual, we use the notation $\partial\mathbb{B}_r$ to denote the boundary $\bar{\mathbb{B}}_r \setminus \overset{o}{\mathbb{B}}_r$ of $\mathbb{B}_r$, that is $\{x \in \mathbb{R}^n : \|x\| = r\}$. If $f : \bar{\mathbb{B}}_r \to \mathbb{R}^n$ is continuous and $0 \notin f(\partial\mathbb{B}_r)$, then the Brouwer topological degree of f with respect to $\mathbb{B}_r$ and 0 is well-defined (see e.g. [16]) and denoted by $\deg(f,\mathbb{B}_r,0)$. Let us now recall some properties of the topological degree that we will use later.

P1. If $0 \notin f(\partial\mathbb{B}_r)$ and $\deg(f,\mathbb{B}_r,0) \neq 0$, then there exist $x \in \mathbb{B}_r$ such that $f(x) = 0$.

P2. Let $\varphi : [0,1] \times \bar{\mathbb{B}}_r \to \mathbb{R}^n; (\lambda,x) \to \varphi(\lambda,x)$, be continuous such that, for each $\lambda \in [0,1]$, one has $0 \notin \varphi(\lambda,\partial\mathbb{B}_r)$, then the map $\lambda \to \deg(\varphi(\lambda,.),\mathbb{B}_r,0)$ is constant on $[0,1]$.

P3. Let us denote by $id_{\mathbb{R}^n}$ the identity mapping on $\mathbb{R}^n$. We have

$$\deg(id_{\mathbb{R}^n},\mathbb{B}_r,0) = 1.$$

P4. If $0 \notin f(\partial\mathbb{B}_r)$ and $\alpha > 0$, then

$$\deg(\alpha f,\mathbb{B}_r,0) = \deg(f,\mathbb{B}_r,0)$$

and

$$\deg(-\alpha f,\mathbb{B}_r,0) = (-1)^n \deg(f,\mathbb{B}_r,0).$$

P5. If $0 \notin f(\partial\mathbb{B}_r)$ and f is odd on $\mathbb{B}_r$ (i.e., $f(-x) = -f(x), \quad \forall x \in \mathbb{B}_r$), then $\deg(f,\mathbb{B}_r,0)$ is odd.

P6. Let $f(x) = Ax - b$, with $A \in \mathbb{R}^{n\times n}$ a nonsingular matrix and $b \in \mathbb{R}^n$. Then $\deg(f, A^{-1}b + \mathbb{B}_r, 0) = \text{sign}(\det A) = \pm 1$.

Let $V \in C^1(\mathbb{R}^n; \mathbb{R})$ and suppose that there exists $r_0 > 0$ such that for every $r \geq r_0$, $0 \notin \nabla V(\partial \mathbb{B}_r)$. Then $\deg(\nabla V, \mathbb{B}_r, 0)$ is constant for $r \geq r_0$ and one defines the index of V at infinity "$\text{ind}(V, \infty)$" by

$$\text{ind}(V, \infty) := \deg(\nabla V, \mathbb{B}_r, 0), \quad \forall r \geq r_0.$$

Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex mapping. It is well known that a) φ is continuous on $\mathbb{R}^n$; b) For all $x \in \mathbb{R}^n$, the convex subdifferential of φ at x is a nonempty compact and convex subset of $\mathbb{R}^n$ and defined by:

$$\partial\varphi(x) = \{w \in \mathbb{R}^n : \varphi(v) - \varphi(x) \geq \langle w, v - u\rangle, \quad \forall v \in \mathbb{R}^n\};$$

c) For all $x \in \mathbb{R}^n$, the directional derivative of φ at $x \in \mathbb{R}^n$ in the direction $\xi \in \mathbb{R}^n$, i.e.,

$$\varphi'(x; \xi) = \lim_{\alpha \downarrow 0} \frac{\varphi(x + \alpha\xi) - \varphi(x)}{\alpha}$$

exists (see e.g. [11] page 164). Since the subdifferential operator $\partial\varphi$ associated to φ, is maximal monotone (Brezis [5]), the operator $(I + \lambda\partial\varphi)^{-1}$ denoted by $P_{\lambda,\varphi}$ is a contraction eveywhere defined on $\mathbb{R}^n$, that is,

$$\|P_{\lambda,\varphi}(x) - P_{\lambda,\varphi}(y)\| \leq \|x - y\|, \quad \forall x,\ y \in \mathbb{R}^n.$$

This operator $P_{\lambda,\varphi}$ is called the resolvent of order $\lambda > 0$ associated to $\partial\varphi$ and for simplicity, we note P_φ instead of $P_{1,\varphi}$ when the parameter $\lambda = 1$. Let $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous mapping and consider the inequality problem: Find $\bar{x} \in \mathbb{R}^n$ such that

$$\langle \Lambda(\bar{x}), v - \bar{x}\rangle + \varphi(v) - \varphi(\bar{x}) \geq 0, \quad \forall v \in \mathbb{R}^n. \tag{3}$$

Clearly problem (3) is equivalent to the nonlinear equation: Find $\bar{x} \in \mathbb{R}^n$ such that

$$\bar{x} - P_\varphi(\bar{x} - \Lambda(\bar{x})) = 0. \tag{4}$$

In view of property P1 recalled earlier, it is important to compute the degree of the operator $id_{\mathbb{R}^n} - P_\varphi \circ (id_{\mathbb{R}^n} - \Lambda)$.

Remark 2.1. If $\bar{x}$ is a solution of problem (3), then

$$\langle \Lambda(\bar{x}), \xi\rangle + \varphi'(\bar{x}; \xi) \geq 0, \quad \forall \xi \in \mathbb{R}^n.$$

Indeed, let $\bar{x}$ be a solution of (3). Let $\xi \in \mathbb{R}^n$ and $\alpha > 0$ be given. Setting $v = \bar{x} + \alpha\xi$ in (3), we get

$$\langle \Lambda(\bar{x}), \alpha\xi\rangle + \varphi(\bar{x} + \alpha\xi) - \varphi(\bar{x}) \geq 0.$$

Thus, for all $\alpha > 0$, we have

$$\langle \Lambda(\bar{x}), \xi \rangle + \frac{\varphi(\bar{x} + \alpha\xi) - \varphi(\bar{x})}{\alpha} \geq 0.$$

Taking the limit as $\alpha \downarrow 0$ we obtain

$$\langle \Lambda(\bar{x}), \xi \rangle + \varphi'(\bar{x}; \xi) \geq 0.$$

Example 2.1. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$\varphi(x) = |x|, \quad \forall x \in \mathbb{R}.$$

We have

$$\partial\varphi(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and

$$P_\varphi(x) = (I + \partial\varphi)^{-1}(x) = \begin{cases} x - 1 & \text{if } x \geq 1 \\ 0 & \text{if } x \in [-1, 1] \\ x + 1 & \text{if } x \leq -1. \end{cases}$$

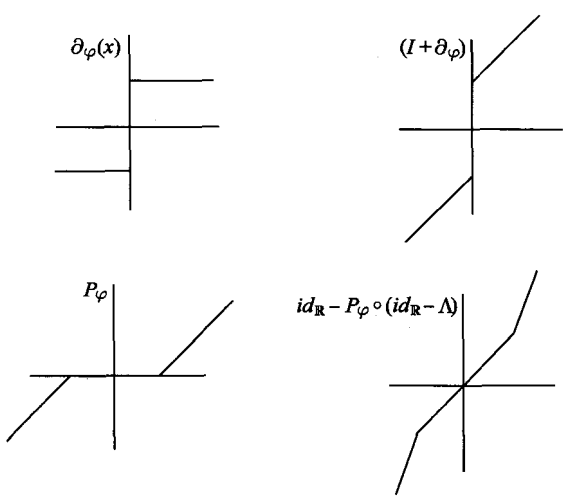


Figure 1 Example 2.1

Setting $\Lambda(x) = 2x$, we get

$$x - P_\varphi(x - \Lambda(x)) = \begin{cases} x & \text{if } |x| \leq 1 \\ 2x - 1 & \text{if } x \geq 1 \\ 2x + 1 & \text{if } x \leq -1. \end{cases}$$

We see that the operator $id_{\mathbb{R}} - P_\varphi \circ (id_{\mathbb{R}} - \Lambda)$ has a unique zero on $\mathbb{R}$.

Proposition 2.1. *Let $L > 0$ be given and assume that the mapping $G : [0, L] \times \mathbb{R}^n \to \mathbb{R}^n$ defined by $(\lambda, y) \mapsto G(\lambda, y)$ is continuous on $[0, L] \times \mathbb{R}^n$. Then, the mapping*

$$(\lambda, y) \mapsto P_{\lambda,\varphi}(G(\lambda, y))$$

is continuous on $[0, L] \times \mathbb{R}^n$.

Proof: Let $\{y_n\} \subset \mathbb{R}^n$ and $\{\lambda_n\} \subset [0, L]$ be given sequences converging respectively to $y^* \in \mathbb{R}^n$ and $\lambda_n \to \lambda^* \in \mathbb{R}$ as $n \to +\infty$. We claim that the sequence $\{P_{\lambda_n,\varphi}(G(\lambda_n, y_n))\}$ tends to $P_{\lambda^*,\varphi}(G(\lambda^*, y^*))$ as $n \to +\infty$. Indeed, setting $x_n := P_{\lambda_n,\varphi}(G(\lambda_n, y_n))$ and $x^* := P_{\lambda^*,\varphi}(G(\lambda^*, y^*))$, we have

$$\langle x_n - G(\lambda_n, y_n), v - x_n\rangle + \lambda_n\varphi(v) - \lambda_n\varphi(x_n) \geq 0, \quad \forall v \in \mathbb{R}^n \tag{5}$$

and

$$\langle x^* - G(\lambda^*, y^*), v - x^*\rangle + \lambda^*\varphi(v) - \lambda^*\varphi(x^*) \geq 0, \quad \forall v \in \mathbb{R}^n. \tag{6}$$

Let us first check that the sequence $\{x_n\}$ is bounded. Indeed, suppose on the contrary that the sequence $\{\|x_n\|\}$ is unbounded. Setting $v := 0$ in (5), we obtain

$$-\langle x_n - G(\lambda_n, y_n), x_n\rangle + \lambda_n[\varphi(0) - \varphi(x_n)] \geq 0,$$

and thus

$$\|x_n\|^2 \leq \|G(\lambda_n, y_n)\|\|x_n\| + \lambda_n[\varphi(0) - \varphi(x_n)].$$

It results that for n large enough, $\|x_n\| \neq 0$ and

$$1 \leq \frac{\|G(\lambda_n, y_n)\|}{\|x_n\|} + \frac{\lambda_n}{\|x_n\|^2}[\varphi(0) - \varphi(x_n)]. \tag{7}$$

As for n large enough, $\frac{1}{\|x_n\|} \in (0, 1]$ we use the convexity of φ, to obtain

$$\varphi\Big(\frac{x_n}{\|x_n\|}\Big) \leq \frac{1}{\|x_n\|}\varphi(x_n) + \Big(1 - \frac{1}{\|x_n\|}\Big)\varphi(0).$$

Thus,

$$\frac{\varphi(0) - \varphi(x_n)}{\|x_n\|} \leq \varphi(0) - \varphi\Big(\frac{x_n}{\|x_n\|}\Big).$$

From (7), we get

$$1 \leq \frac{\|G(\lambda_n, y_n)\|}{\|x_n\|} + \lambda_n\left[\frac{\varphi(0) - \varphi(\frac{x_n}{\|x_n\|})}{\|x_n\|}\right]. \tag{8}$$

The sequence $\{\frac{x_n}{\|x_n\|}\}$ remains in the compact set $\partial\mathbb{B}_1$ and from the continuity of φ, we derive that the sequence $\{\varphi(\frac{x_n}{\|x_n\|})\}$ is bounded in $\mathbb{R}$. Hence,

$$\lim_{n\to+\infty} \frac{\varphi(\frac{x_n}{\|x_n\|})}{\|x_n\|} = 0.$$

Taking now the limit as $n \to +\infty$ in (8), we obtain the contradiction $1 \leq 0$. The sequence $\{x_n\}$ is thus bounded. Setting $v := x^*$ in (5) and $v := x_n$ in (6), we obtain the relations

$$\langle x_n - G(\lambda_n, y_n), x_n - x^*\rangle - \lambda_n\varphi(x^*) + \lambda_n\varphi(x_n) \leq 0 \tag{9}$$

and

$$-\langle x^* - G(\lambda^*, y^*), x_n - x^*\rangle - \lambda^*\varphi(x_n) + \lambda^*\varphi(x^*) \leq 0. \tag{10}$$

Thus

$$\begin{aligned}\|x_n - x^*\|^2 \leq \|G(\lambda_n, y_n) - G(\lambda^*, y^*)\|\|x_n - x^*\| \\ +(\lambda_n - \lambda^*)\varphi(x^*) + (\lambda^* - \lambda_n)\varphi(x_n).\end{aligned} \tag{11}$$

Using the continuity of φ and the boundeness of $\{x_n\}$, we get that the sequence $\{\varphi(x_n)\}_n$ is bounded in $\mathbb{R}$. Moreover $\|G(\lambda_n, y_n) - G(\lambda^*, y^*)\| \to 0$ and $(\lambda_n - \lambda^*) \to 0$ in $\mathbb{R}$ as $n \to +\infty$. Relation (11) yields that $x_n \to x^*$ in $\mathbb{R}^n$ as $n \to +\infty$. Hence the operator $(\lambda, y) \mapsto P_{\lambda,\varphi}(G(\lambda, y))$ is continuous, which completes the proof. □

Proposition 2.2. *Suppose that $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex function. If there exists a continuous mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ and $r > 0$ such that*

$$\langle \Lambda(x), H(x)\rangle + \varphi'(x; H(x)) < 0, \quad \forall x \in \partial\mathbb{B}_r. \tag{12}$$

Then

$$\deg(H, \mathbb{B}_r, 0) = (-1)^n \deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - \Lambda), \mathbb{B}_r, 0).$$

Proof: Let $h : [0,1] \times \bar{\mathbb{B}}_r \to \mathbb{R}^n; (\lambda, y) \mapsto h(\lambda, y) := y - P_{\lambda,\varphi}(y - \lambda\Lambda(y) + (1-\lambda)H(y))$. Proposition 2.1 ensures that h is continuous. Let us now check that $h(\lambda, x) \neq 0, \quad \forall x \in \partial\mathbb{B}_r$. Indeed, suppose on the contrary that there exists $x \in \partial\mathbb{B}_r$ and $\lambda \in [0,1]$ such that $h(\lambda, x) = 0$, that is

$$x = P_{\lambda,\varphi}(x - \lambda\Lambda(x) + (1-\lambda)H(x)).$$

We first remark that $\lambda \neq 0$. Indeed, if we suppose, on the contrary, that $\lambda = 0$, then $x = P_0(x + H(x)) = x + H(x)$. This yields $H(x) = 0$ which contradicts assumption (12). Thus $\lambda > 0$ and

$$\langle \lambda\Lambda(x) - (1-\lambda)H(x), v - x\rangle + \lambda\varphi(v) - \lambda\varphi(x) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

It results that (see Remark 2.1):

$$\langle \lambda\Lambda(x) - (1-\lambda)H(x), \xi\rangle + \lambda\varphi'(x;\xi) \geq 0, \quad \forall \xi \in \mathbb{R}^n.$$

Setting $\xi := H(x)$, we obtain

$$\lambda[\langle \Lambda(x), H(x)\rangle + \varphi'(x; H(x))] \geq (1-\lambda)\|H(x)\|^2 \geq 0,$$

which contradicts assumption (12). Therefore,

$$\begin{aligned}\deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - \Lambda), \mathbb{B}_r, 0) &= \deg(h(1,.), \mathbb{B}_r, 0)\\ &= \deg(h(0,.), \mathbb{B}_r, 0)\\ &= \deg(id_{\mathbb{R}^n} - P_0(id_{\mathbb{R}^n} + H), \mathbb{B}_r, 0)\\ &= \deg(-H, \mathbb{B}_r, 0)\\ &= (-1)^n \deg(H, \mathbb{B}_r, 0),\end{aligned}$$

which completes the proof. □

3 Some existence results for finite variational inequalities

As a direct consequence of Proposition 2.2, we have the following existence results for finite dimensional variational inequalities.

Theorem 3.1. *Suppose that* 1) $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ *is a continuous operator;* 2) $\varphi : \mathbb{R}^n \to \mathbb{R}$ *is a convex function;* 3) *there exists* $r > 0$ *such that*

$$\langle \Lambda(x), x\rangle - \varphi'(x; -x) > 0, \quad \forall x \in \partial\mathbb{B}_r.$$

Then there exists $\bar{x} \in \mathbb{B}_r$ *such that*

$$\langle \Lambda(\bar{x}), v - \bar{x}\rangle + \varphi(v) - \varphi(\bar{x}) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

Proof: Just apply Proposition 2.2 with $H := -id_{\mathbb{R}^n}$. Indeed, here we have

$$\langle \Lambda(x), H(x)\rangle + \varphi'(x; H(x)) = -\langle \Lambda(x), x\rangle + \varphi'(x; -x).$$

□

Theorem 3.2. *Suppose that* 1) $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ *is continuous;* 2) $\varphi : \mathbb{R}^n \to \mathbb{R}$ *is convex and Lipschitz continuous with Lipschitz constant* $K > 0$, i.e.,

$$|\varphi(x) - \varphi(y)| \leq K\|x - y\|, \quad \forall x, y \in \mathbb{R}^n;$$

3) *there exists* $r > 0$ *such that*

$$\|\Lambda(x)\| > K, \quad \forall x \in \partial\mathbb{B}_r,$$

and

$$\deg(\Lambda, \mathbb{B}_r, 0) \neq 0.$$

Then there exists $\bar{x} \in \mathbb{B}_r$ *such that*

$$\langle \Lambda(\bar{x}), v - \bar{x} \rangle + \varphi(v) - \varphi(\bar{x}) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

Proof: Just apply Proposition 2.2 with $H := -\Lambda$. Indeed, we have

$$\begin{aligned}\langle \Lambda(x), H(x) \rangle + \varphi'(x; H(x)) &= -\|\Lambda(x)\|^2 + \varphi'(x; -\Lambda(x)) \\ &\leq -\|\Lambda(x)\|^2 + K\|\Lambda(x)\| \\ &= \|\Lambda(x)\|(K - \|\Lambda(x)\|).\end{aligned}$$

Therefore,

$$\langle \Lambda(x), H(x) \rangle + \varphi'(x; H(x)) < 0, \quad \forall x \in \partial\mathbb{B}_r.$$

Proposition 2.2 ensures that

$$\deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - \Lambda), \mathbb{B}_r, 0) = (-1)^n \deg(H, \mathbb{B}_r, 0) = \deg(\Lambda, \mathbb{B}_r, 0) \neq 0.$$

Hence, there exists $\bar{x} \in \mathbb{B}_r$ such that $\bar{x} = P_\varphi(\bar{x} - \Lambda(\bar{x}))$. The conclusion follows. □

Theorem 3.3. *Suppose that*

1) $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ *is continuous and there exists* $r > 0$ *such that*

$$\langle \Lambda x, x \rangle > 0, \quad \forall x \in \partial\mathbb{B}_r \text{ and } \deg(\mathrm{id}_{\mathbb{R}^n} + \Lambda, \mathbb{B}_r, 0) \neq 0.$$

2) $\varphi : \mathbb{R}^n \to \mathbb{R}$ *is a convex function satisfying,*

$$\varphi'(x; -x - \Lambda x) \leq 0, \quad \forall x \in \partial\mathbb{B}_r;$$

Then there exists $\bar{x} \in \mathbb{B}_r$ *such that*

$$\langle \Lambda(\bar{x}), v - \bar{x} \rangle + \varphi(v) - \varphi(\bar{x}) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

Proof: Just take $H := -\mathrm{id}_{\mathbb{R}^n} - \Lambda$ in Proposition 2.2. Indeed, we have

$$\begin{aligned}&\langle \Lambda(x), H(x) \rangle + \varphi'(x; H(x)) \\ &= -\|\Lambda(x)\|^2 - \langle \Lambda(x), x \rangle + \varphi'(x; -x - \Lambda(x)) < 0, \quad \forall x \in \partial\mathbb{B}_r.\end{aligned}$$

According to Proposition 2.2, we have

$$\begin{aligned}\deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - \Lambda), \mathbb{B}_r, 0) &= (-1)^n \deg(H, \mathbb{B}_r, 0) \\ &= \deg(id_{\mathbb{R}^n} + \Lambda, \mathbb{B}_r, 0) \neq 0.\end{aligned}$$

Hence, there exists $\bar{x} \in \mathbb{B}_r$ such that $\bar{x} = P_\varphi(\bar{x} - \Lambda(\bar{x}))$. The conclusion follows. □

Corollary 3.1. *Let $f \in \mathbb{R}^n$ be given. Suppose that* 1) *$A \in \mathbb{R}^{n\times n}$ is a real nonsingular matrix;* 2) *$\varphi : \mathbb{R}^n \to \mathbb{R}$ is convex and Lipschitz continuous with Lipschitz constant $K > 0$. Then there exists $\bar{x} \in \mathbb{R}^n$ such that*

$$\langle A\bar{x} - f, v - \bar{x}\rangle + \varphi(v) - \varphi(\bar{x}) \geq 0, \quad \forall v \in \mathbb{R}^n.$$

Proof: The result is a consequence of Theorem 3.2 with Λ defined by

$$\Lambda(x) = Ax - f, \quad \forall x \in \mathbb{R}^n.$$

The matrix A is nonsingular and thus there exists $c > 0$ such that $\|Ax\| \geq c\|x\|, \quad \forall x \in \mathbb{R}^n$. Let us choose

$$r > \max\left\{\frac{K + \|f\|}{c}, \|A^{-1}f\|\right\}.$$

We see that if $\|x\| = r$, then

$$\|\Lambda(x)\| \geq \|Ax\| - \|f\| \geq c\|x\| - \|f\| > K.$$

On the other hand, we remark that

$$h(\lambda, x) := Ax - \lambda f \neq 0, \quad \forall \lambda \in [0,1], x \in \partial\mathbb{B}_r.$$

Indeed, suppose on the contrary that there exists $\lambda \in [0,1]$ and $x \in \partial\mathbb{B}_r$ such that $Ax = \lambda f$. Then

$$\|x\| = \lambda\|A^{-1}f\| \leq \|A^{-1}f\|$$

and we obtain the contradiction $r \leq \|A^{-1}f\|$. Thus

$$\begin{aligned}\deg(A. - f, \mathbb{B}_r, 0) &= \deg(h(1,.), \mathbb{B}_r, 0) \\ &= \deg(h(0,.), \mathbb{B}_r, 0) \\ &= \deg(A., \mathbb{B}_r, 0) \\ &= \text{ sign } (\det A) \\ &\neq 0,\end{aligned}$$

which completes the proof. □

4 The Poincaré operator

Let us first recall some general existence and uniqueness result (see e.g. [18]).

Theorem 4.1. *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous operator such that for some $\omega \in \mathbb{R}$, $F + \omega I$ is monotone,* i.e.,

$$\langle F(x) - F(y), x - y\rangle \geq -\omega\|x - y\|^2, \quad \forall x,\ y \in \mathbb{R}^n.$$

Suppose that $f : [0, +\infty) \to \mathbb{R}^n$ satisfies

$$f \in C^0([0, +\infty); \mathbb{R}^n),\ \frac{df}{dt} \in L^1_{\text{loc}}(0, +\infty; \mathbb{R}^n).$$

Let $u_0 \in \mathbb{R}^n$ and $0 < T < +\infty$ be given. There exists a unique $u \in C^0([0, T]; \mathbb{R}^n)$ such that

$$\tfrac{du}{dt} \in L^\infty(0, T; \mathbb{R}^n); \tag{13}$$

$$u \text{ is right-differentiable on } [0, T); \tag{14}$$

$$u(0) = u_0; \tag{15}$$

$$\left\langle \tfrac{du}{dt}(t) + F(u(t)) - f(t), v - u(t)\right\rangle + \varphi(v) - \varphi(u(t)) \geq 0,$$
$$\forall v \in \mathbb{R}^n,\ \text{a.e. } t \in [0, T]. \tag{16}$$

Remark 4.1. Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is of the type

$$F(x) = Ax + \Psi'(x) + F_1(x), \quad \forall x \in \mathbb{R}^n,$$

where $A \in \mathbb{R}^{n\times n}$ is a real matrix, $\Psi \in C^1(\mathbb{R}^n; \mathbb{R})$ is convex and F_1 is Lipschitz continuous, i.e.,

$$\|F_1(x) - F_1(y)\| \leq k\|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

for some constant $k > 0$. Then F is continuous and $F + \omega I$ is monotone provided that

$$\omega \geq \sup_{\|x\|=1} \langle -Ax, x\rangle + k.$$

We note that if F is k-Lipschitz, then $F + kI$ is monotone.

Remark 4.2. i) The variational inequality in (16) can also be written as the differential inclusion

$$\frac{du}{dt}(t) + F(u(t)) - f(t) \in -\partial\varphi(u(t)),\ \text{a.e. } t \in [0, T], \tag{17}$$

ii) Let $u : [0, T] \to \mathbb{R}$ be the unique solution of (13)-(16). Then

$$\left\langle \frac{du}{dt}(t) + F(u(t)) - f(t), \xi\right\rangle + \varphi'(u(t); \xi) \geq 0, \quad \forall \xi \in \mathbb{R}^n,\ \text{a.e. } t \in [0, T].$$

Let $T > 0$ be given. Theorem 4.1 enables us to define the one parameter family $\{S(t) : 0 \leq t \leq T\}$ of operators from $\mathbb{R}^n$ into $\mathbb{R}^n$, as follows:

$$\forall y \in \mathbb{R}^n, \quad S(t)y = u(t), \tag{18}$$

u being the unique solution on $[0, T]$ of the evolution problem (13)-(16). Note that

$$\forall y \in \mathbb{R}^n, \quad S(0)y = y.$$

Lemma 4.1. (see e.g. [18]) *Let $T > 0$ be given and let $a, b \in L^1(0, T; \mathbb{R})$ with $b(t) \geq 0$* a.e. *$t \in [0, T]$. Let the absolutely continuous function $w : [0, T] \to \mathbb{R}_+$ satisfies*

$$(1 - \alpha)\frac{dw}{dt}(t) \leq a(t)w(t) + b(t)w^{\alpha}(t), \text{ a.e. } t \in [0, T],$$

where $0 \leq \alpha < 1$. Then

$$w^{1-\alpha}(t) \leq w^{1-\alpha}(0)e^{\int_0^t a(s)ds} + \int_0^t e^{\int_s^t a(q)dq} b(s)ds, \quad \forall t \in [0, T].$$

Theorem 4.2. *Suppose that the assumptions of Theorem 4.1 hold. Then*

$$\|S(t)y - S(t)z\| \leq e^{\omega t}\|y - z\|, \forall y, z \in \mathbb{R}^n, t \in [0, T].$$

Proof: Let $y, z \in \mathbb{R}^n$ be given. We have

$$\begin{aligned} -\Big\langle \frac{d}{dt}S(t)y + F(S(t)y) - f(t), S(t)z - S(t)y \Big\rangle \\ -\varphi(S(t)z) + \varphi(S(t)y) \leq 0, \text{ a.e. } t \in [0, T] \end{aligned}$$

and

$$\begin{aligned} \Big\langle \frac{d}{dt}S(t)z + F(S(t)z) - f(t), S(t)z - S(t)y \Big\rangle \\ -\varphi(S(t)y) + \varphi(S(t)z) \leq 0, \text{ a.e. } t \in [0, T]. \end{aligned}$$

It results that for almost every $t \in [0, T]$ we have

$$\begin{aligned} &\Big\langle \frac{d}{dt}(S(t)z - S(t)y), S(t)z - S(t)y \Big\rangle \\ &\leq \omega\|S(t)z - S(t)y\|^{-}\langle [F + \omega I](S(t)z) \\ &- [F + \omega I](S(t)y), S(t)z - S(t)y\rangle.2 \end{aligned}$$

Since by assumption, $F + \omega I$ is monotone, it results that

$$\frac{d}{dt}\|S(t)z - S(t)y\|^2 \leq 2\omega\|S(t)z - S(t)y\|^2, \text{ a.e. } t \in [0,T]. \tag{19}$$

Using Lemma 4.1 with $w(\cdot) := \|S(\cdot)z - S(\cdot)y\|^2$, $a(\cdot) := 2\omega$, $b(\cdot) = 0$, $\alpha = 0$, we get

$$\|S(t)z - S(t)y\|^2 \leq \|z - y\|^2 e^{2\omega t}, \quad \forall t \in [0,T].$$

The conclusion follows. □

Let us now consider the Poincaré operator $S(T) : \mathbb{R}^n \to \mathbb{R}^n$ given by $y \mapsto S(T)y$. Theorem 4.2 ensures that $S(T)$ is Lipschitz continuous, i.e.,

$$\|S(T)y - S(T)z\| \leq e^{\omega T}\|y - z\|, \quad \forall y, z \in \mathbb{R}^n.$$

Remark 4.3. i) Note that if F is continuous and monotone, then Theorem 4.2 holds with $\omega = 0$. In this case, the Poincaré operator $S(T)$ is nonexpansive, i.e.,

$$\|S(T)y - S(T)z\| \leq \|y - z\|, \quad \forall y, z \in \mathbb{R}^n.$$

ii) If F is continuous and strongly monotone, i.e., there exists $\alpha > 0$ such that

$$\langle F(x) - F(y), x - y\rangle \geq \alpha\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n,$$

then Theorem 4.2 holds with $\omega = -\alpha < 0$ and the Poincaré operator $S(T)$ is a contraction.

According to (18), the unique solution of the problem (13)-(16) satisfies, in addition, the periodicity condition

$$u(0) = u(T)$$

if and only if y is a fixed point of $S(T)$, that is

$$S(T)y = y.$$

Thus the problem of the existence of a periodic solution for the evolution problem (13)-(14), (16) reduces to the existence of a fixed point for $S(T)$.

5 Periodic solutions

Definition 5.1. Let $\Omega \subset \mathbb{R}^n$ be a given subset of $\mathbb{R}^n$. We say that $V \in C^1(\mathbb{R}^n; \mathbb{R})$ is a guiding function for (17) on Ω provided that

$$\langle F(x) - f(t), \nabla V(x)\rangle + \varphi'(x; \nabla V(x)) < 0, \quad \forall x \in \Omega, t \in [0,T]. \tag{20}$$

Remark 5.1. i) Suppose that there exists a guiding function $V \in C^1(\mathbb{R}^n;\mathbb{R})$ for (17) on $\partial\mathbb{B}_r$ $(r > 0)$, i.e.,

$$\langle F(x) - f(t), \nabla V(x)\rangle + \varphi'(x; \nabla V(x)) < 0, \quad \forall x \in \partial\mathbb{B}_r, t \in [0,T].$$

Then for any $\tau \in [0,T]$, we have

$$\deg(\nabla V, \mathbb{B}_r, 0) = (-1)^n \deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - F + f(\tau)), \mathbb{B}_r, 0). \quad (21)$$

Relation in (21) is a consequence of Proposition 2.2.

ii) Suppose that there exists a guiding function $V \in C^1(\mathbb{R}^n;\mathbb{R})$ for (17) on

$$\Omega_R := \{x \in \mathbb{R}^n : \|x\| \geq R\} \ \ (R > 0).$$

Then for $r \geq R$ and any $\tau \in [0,T]$, we have

$$\mathrm{ind}(V, \infty) = (-1)^n \deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - F + f(\tau)), \mathbb{B}_r, 0).$$

Proposition 5.1. *Suppose that there exists $R > 0$ such that*

$$\langle F(x) - f(t), \nabla V(x)\rangle < 0, \quad \forall x \in \mathbb{R}^n, \|x\| \geq R, t \in [0,T]. \quad (22)$$

Then for $r \geq R$ and any $\tau \in [0,T]$, we have

$$\mathrm{ind}(V, \infty) = \deg(f(\tau) - F, \mathbb{B}_r, 0).$$

Proof: Let $r \geq R$ be given and let $h : [0,1] \times \bar{\mathbb{B}}_r \to \mathbb{R}^n$ be the mapping defined by $(\lambda, y) \mapsto h(\lambda, y) := \lambda\nabla V(y) + (1-\lambda)(f(0) - F(y))$. We claim that $h(\lambda, y) \neq 0, \quad \forall y \in \partial\mathbb{B}_r, \lambda \in [0,1]$. Indeed, suppose by contradiction that

$$\lambda\nabla V(y) + (1-\lambda)(f(\tau) - F(y)) = 0$$

for some $y \in \partial\mathbb{B}_r$ and $\lambda \in [0,1]$. Then

$$\lambda\langle \nabla V(y), f(\tau) - F(y)\rangle = -(1-\lambda)\|f(\tau) - F(y)\|^2. \quad (23)$$

If $\lambda = 0$, relation (23) implies that $f(\tau) - F(y) = 0$, and since $y \in \partial\mathbb{B}_r$ and $r \geq R$, we derive a contradiction from relation (22). Hence, $\lambda \neq 0$ and relation (23) yields $\langle \nabla V(y), f(\tau) - F(y)\rangle < 0$ and a contradiction to (22). Thus,

$$\begin{aligned}\deg(\nabla V, \mathbb{B}_r, 0) &= \deg(h(1,.), \mathbb{B}_r, 0) = \deg(h(0,.), \mathbb{B}_r, 0)\\ &= \deg(f(\tau) - F, \mathbb{B}_r, 0).\end{aligned}$$

□

Theorem 5.1. *Suppose that $f \in C^0([0,+\infty);\mathbb{R}^n)$ and $\frac{df}{dt} \in L^1_{\text{loc}}(0,+\infty;\mathbb{R}^n)$. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping satisfying the conditions of Theorem 4.1. Suppose that there exist constants $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, $C_1 \geq 0$, $C_2 \geq 0$ and a function $W \in C^1(\mathbb{R}^n;\mathbb{R})$ such that*

$$c_1\|x\|^2 \leq W(x) \leq c_3\|x\|^2, \quad \forall x \in \mathbb{R}^n, \tag{24}$$

$$c_2\|\nabla W(x)\|^2 \leq W(x), \quad \forall x \in \mathbb{R}^n \tag{25}$$

and

$$\langle F(x), \nabla W(x)\rangle + \varphi'(x;\nabla W(x)) \leq C_1\|\nabla W(x)\|^2 + C_2\|\nabla W(x)\|, \ \forall x \in \mathbb{R}^n. \tag{26}$$

Let $T > 0$ be given. Assume that there exists a (guiding) function $V \in C^1(\mathbb{R}^n;\mathbb{R})$ and $R > 0$ such that

$$\langle F(x) - f(t), \nabla V(x)\rangle + \varphi'(x;\nabla V(x)) < 0, \quad \forall x \in \mathbb{R}^n, \|x\| \geq R, t \in [0,T]. \tag{27}$$

Then there exists $r_0 > R$ such that for any $\tau \in [0,T]$, we have

$$\begin{aligned}\deg(id_{\mathbb{R}^n} - S(T)(\cdot), \mathbb{B}_r, 0) &= \deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - F + f(\tau)), \mathbb{B}_r, 0)\\ &= (-1)^n \text{ind}(V,\infty), \quad \forall r \geq r_0.\end{aligned}$$

Proof: Let us first remark that without any loss of generality, we may assume that $C_1 > 0$. We set

$$r_0 := R\frac{\sqrt{c_3}}{\sqrt{c_1}}e^{\frac{C_1}{2c_2}T} + \frac{\sqrt{c_2}C_2}{\sqrt{c_1}C_1}(e^{\frac{C_1}{2c_2}T} - 1) + \frac{1}{2\sqrt{c_1c_2}}\int_0^T \|f(s)\|e^{\frac{C_1}{2c_2}s}ds.$$

(1) We claim that if $y \in \mathbb{R}^n$, $\|y\| = r$ with $r \geq r_0$, then

$$\|S(t)y\| \geq R, \quad \forall t \in [0,T].$$

Suppose by contradiction that there exists $t^* \in [0,T]$ such that $\|S(t^*)y\| < R$. We know that $u(.) \equiv S(.)y$ satisfies (17), i.e.,

$$\frac{du}{dt}(t) + F(u(t)) - f(t) \in -\partial\varphi(u(t)), \text{ a.e. } t \in [0,T], \tag{28}$$

and thus

$$\frac{du}{dt}(t^*-t) + F(u(t^*-t)) - f(t^*-t) \in -\partial\varphi(u(t^*-t)), \text{ a.e. } t \in [0,t^*]. \tag{29}$$

Setting

$$Y(t) = u(t^* - t), t \in [0,t^*],$$

we derive

$$-\frac{dY}{dt}(t) + F(Y(t)) - f(t^* - t) \in -\partial\varphi(Y(t)), \text{ a.e. } t \in [0, t^*]. \tag{30}$$

Thus, for every $v \in \mathbb{R}^n$ and almost every $t \in [0, t^*]$, we have

$$\varphi(v) - \varphi(Y(t)) + \left\langle -\frac{dY}{dt}(t), v - Y(t) \right\rangle \geq \langle -F(Y(t)) + f(t^* - t), v - Y(t) \rangle.$$

Hence,

$$\left\langle \frac{dY}{dt}(t), \xi \right\rangle \leq \langle F(Y(t)) - f(t^* - t), \xi \rangle + \varphi'(Y(t); \xi), \quad \forall \xi \in \mathbb{R}^n, \text{ a.e. } t \in [0, t^*].$$

Set $v = \nabla W(Y(t))$ to obtain

$$\begin{aligned}
\left\langle \frac{dY}{dt}(t), \nabla W(Y(t)) \right\rangle &\leq \langle F(Y(t)) - f(t^* - t), \nabla W(Y(t)) \rangle \\
&\quad + \varphi'(Y(t); \nabla W(Y(t))) \\
&\leq C_1 \|\nabla W(Y(t))\|^2 + C_2 \|\nabla W(Y(t))\| \\
&\quad + \|f(t^* - t)\| \|\nabla W(Y(t))\| \\
&\leq \frac{C_1}{c_2} W(Y(t)) + \frac{C_2}{\sqrt{c_2}} \sqrt{W(Y(t))} \\
&\quad + \frac{1}{\sqrt{c_2}} \|f(t^* - t)\| \sqrt{W(Y(t))}, \quad \text{a.e. } t \in [0, t^*].
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{1}{2} \frac{d}{dt} W(Y(t)) &\leq \frac{C_1}{2c_2} W(Y(t)) + \frac{C_2}{2\sqrt{c_2}} \sqrt{W(Y(t))} \\
&\quad + \frac{1}{2\sqrt{c_2}} \|f(t^* - t)\| \sqrt{W(Y(t))}, \text{ a.e. } t \in [0, t^*].
\end{aligned}$$

Using Lemma 4.1 with $w(\cdot) := W(Y(\cdot))$, $a(\cdot) := \frac{C_1}{2c_2}$, $b(\cdot) := \frac{1}{2\sqrt{c_2}}(C_2 + \|f(t^* - \cdot)\|)$ and $\alpha := \frac{1}{2}$, we obtain for every $t \in [0, t^*]$

$$\begin{aligned}
\sqrt{W(Y(t))} &\leq \sqrt{W(Y(0))} e^{\frac{C_1}{2c_2} t} + \int_0^t \frac{C_2}{2\sqrt{c_2}} e^{\frac{C_1}{2c_2}(t-s)} ds \\
&\quad + \int_0^t \frac{1}{2\sqrt{c_2}} \|f(t^* - s)\| e^{\frac{C_1}{2c_2}(t-s)} ds.
\end{aligned}$$

Thus, for every $t \in [0, t^*]$

$$\begin{aligned}
\|Y(t)\| &\leq \frac{1}{\sqrt{c_1}} \Big(\sqrt{c_3} \|Y(0)\| e^{\frac{C_1}{2c_2} t} + \int_0^t \frac{C_2}{2\sqrt{c_2}} e^{\frac{C_1}{2c_2}(t-s)} ds \\
&\quad + \int_0^t \frac{1}{2\sqrt{c_2}} \|f(t^* - s)\| e^{\frac{C_1}{2c_2}(t-s)} ds \Big).
\end{aligned}$$

Since $Y(t^*) = u(0) = S(0)y = y$ and $Y(0) = u(t^*) = S(t^*)y$, we get

$$\begin{aligned}\|y\| &\leq \frac{1}{\sqrt{c_1}}\Big(\sqrt{c_3}\|S(t^*)y\|e^{\frac{c_1}{2c_2}t^*} + \int_0^{t^*} \frac{C_2}{2\sqrt{c_2}} e^{\frac{c_1}{2c_2}(t^*-s)}ds \\ &\quad + \int_0^{t^*} \frac{1}{2\sqrt{c_2}}\|f(t^*-s)\|e^{\frac{c_1}{2c_2}(t^*-s)}ds\Big) \\ &< R\frac{\sqrt{c_3}}{\sqrt{c_1}}e^{\frac{c_1}{2c_2}T} + \frac{\sqrt{c_2}C_2}{\sqrt{c_1}C_1}(e^{\frac{c_1}{2c_2}T} - 1) \\ &\quad + \frac{1}{2\sqrt{c_1c_2}}\int_0^T \|f(s)\|e^{\frac{c_1}{2c_2}s}ds = r_0.\end{aligned}$$

Hence, $\|y\| < r_0$, a contradiction. Let $r \geq r_0$ be given.

(2) We claim that there exists $\varepsilon > 0$ and $T^* \in (0, T]$ such that

$$\langle F(x) - f(t), \nabla V(y)\rangle + \varphi'(x; \nabla V(y)) < 0,$$

$$\forall x \in \mathbb{R}^n, y \in \mathbb{R}^n, \|y\| = r, \|x - y\| \leq \varepsilon, t \in [0, T^*].$$

Indeed, recalling that the mapping $(z, \xi) \mapsto \varphi'(z; \xi)$ is upper semicontinuous (see e.g. [11]), we note that the mapping $(t, x, y) \mapsto \langle F(x) - f(t), \nabla V(y)\rangle + \varphi'(x; \nabla V(y))$ is upper semicontinuous on $[0, T] \times \mathbb{R}^n \times \mathbb{R}^n$ and if $y \in \mathbb{R}^n$, $\|y\| = r \geq r_0 \geq R$ then (by condition (27)):

$$\langle F(y) - f(0), \nabla V(y)\rangle + \varphi'(y; \nabla V(y)) < 0.$$

Thus, for $t > 0$ close to 0, let us say $t \leq T^*$ and x closed to y, let us say $\|x - y\| \leq \varepsilon$, $\varepsilon > 0$, small enough , we have

$$\langle F(x) - f(0), \nabla V(y)\rangle + \varphi'(x; \nabla V(y)) < 0.$$

(3) We claim that there exists $\bar{T} \in (0, T^*]$ such that

$$\|S(t)y - y\| \leq \varepsilon, \quad \forall y \in \partial\mathbb{B}_r, \quad \forall t \in [0, \bar{T}].$$

Indeed, by contradiction suppose that there exists sequences $\{t_n\}$ and $\{y_n\}$ with $t_n \in [0, \frac{T^*}{n}]$ $(n \in \mathbb{N}, n \geq 1)$, $\|y_n\| = r$ and such that $\|S(t_n)y_n - y_n\| > \varepsilon$. Taking a subsequence, if necessary, we may assume that $t_n \to 0+$ and $y_n \to y^* \in \partial\mathbb{B}_r$. On the other hand, we have

$$\begin{aligned}\|S(t_n)y_n - y_n\| &= \|S(t_n)y_n - S(t_n)y^* + S(t_n)y^* - y_n\| \\ &\leq \|S(t_n)y_n - S(t_n)y^*\| + \|S(t_n)y^* - y_n\|.\end{aligned}$$

Then, using Theorem 4.2, we obtain

$$\|S(t_n)y_n - y_n\| \leq \sqrt{e^{2wt_n}}\|y_n - y^*\| + \|S(t_n)y^* - y_n\|.$$

Using the continuity of the map $t \mapsto S(t)y$, we see that $\|S(t_n)y_n - y_n\| \to 0$, a contradiction.
(4) Let $H_{\bar{T}} : [0,1] \times \bar{C}_r \to \mathbb{R}^n; (\lambda, y) \to H_{\bar{T}}(\lambda, y) := y - (1-\lambda)\nabla V(y) - S(\lambda\bar{T})y$. We claim that the homotopy $H_{\bar{T}}$ is such that $0 \neq H_{\bar{T}}(\lambda, y), \forall y \in \partial\mathbb{B}_r, \lambda \in [0,1]$. By contradiction, suppose that there exists $y \in \mathbb{R}^n$, $\|y\| = r$ and $\lambda \in [0,1]$ such that

$$y - (1-\lambda)\nabla V(y) - S(\lambda\bar{T})y = 0.$$

Then

$$S(\lambda\bar{T})y - y = -(1-\lambda)\nabla V(y)$$

and thus

$$\langle S(\lambda\bar{T})y - y, \nabla V(y)\rangle = -(1-\lambda)\|\nabla V(y)\|^2 \leq 0. \tag{31}$$

On the other hand, we know that

$$\begin{aligned} &\left\langle \frac{d}{dt}S(t)y, v - S(t)y \right\rangle + \varphi(v) - \varphi(S(t)y) \\ &\geq \langle -F(S(t)y) + f(t), v - S(t)y\rangle, \forall v \in \mathbb{R}^n, \quad \text{a.e. } t \in [0,T]. \end{aligned} \tag{32}$$

Thus

$$\begin{aligned} &\left\langle \frac{d}{dt}S(t)y, \nabla V(y) \right\rangle + \varphi'(S(t)y; \nabla V(y)) \\ &\geq \langle -F(S(t)y) + f(t), \nabla V(y)\rangle, \quad \text{a.e., } t \in [0,T]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\left\langle \int_0^{\lambda\bar{T}} \frac{d}{ds}S(s)y ds, \nabla V(y) \right\rangle \\ &\geq \int_0^{\lambda\bar{T}} \langle -F(S(s)y) + f(s), \nabla V(y)\rangle - \varphi'(S(s)y; \nabla V(y))\, ds. \end{aligned}$$

Part (1) of this proof ensures that $\|S(t)y\| \geq R, \quad \forall t \in [0, \lambda\bar{T}] \subset [0,T]$. Part (3) of this proof garantees that $\|S(t)y - y\| \leq \varepsilon, \quad \forall t \in [0, \lambda\bar{T}] \subset [0, \bar{T}]$. Then using part (2) of this proof, we may assert that the map $s \mapsto \langle F(S(s)y) - f(s), \nabla V(y)\rangle + \varphi'(S(s)y; \nabla V(y))$ is upper semicontinuous and strictly negative on $[0, \lambda\bar{T}]$. Thus

$$\int_0^{\lambda\bar{T}} \langle -F(S(s)y) + f(s), \nabla V(y)\rangle - \varphi'(S(s)y; \nabla V(y)) ds > 0$$

and we obtain

$$\langle S(\lambda\bar{T})y - y, \nabla V(y)\rangle = \langle \int_0^{\lambda\bar{T}} \frac{d}{ds}S(s)y ds, \nabla V(y)\rangle > 0.$$

This contradicts relation (31).
(5) Thanks to part (4) of this proof, we may use the invariance by homotopy property of the topological degree and observe that

$$\begin{aligned}\deg(id_{\mathbb{R}^n} - S(\bar{T})\cdot, \mathbb{B}_r, 0) &= \deg(H_{\bar{T}}(1,\cdot), \mathbb{B}_r, 0)\\ &= \deg(H_{\bar{T}}(0,\cdot), \mathbb{B}_r, 0)\\ &= \deg(-\nabla V, \mathbb{B}_r, 0)\\ &= (-1)^n \deg(\nabla V, \mathbb{B}_r, 0).\end{aligned}$$

(6) Let $H : [0,1] \times \bar{\mathbb{B}}_r \to \mathbb{R}^n; (\lambda, y) \to H(\lambda, y) := y - S((1-\lambda)T + \lambda\bar{T})y$. We claim that $H(\lambda, y) \neq 0, \quad \forall y \in \partial\mathbb{B}_r, \lambda \in [0,1]$. By contradiction, suppose that there exists $y \in \mathbb{R}^n$, $\|y\| = r$ and $\lambda \in [0,1]$ such that $y = S((1-\lambda)T + \lambda\bar{T})y$. Let us set $h := (1-\lambda)T + \lambda\bar{T}$. We have

$$y = S(h)y$$

and thus,

$$V(y) = V(S(h)y). \tag{33}$$

On the other hand,

$$\begin{aligned}&\left\langle \frac{d}{dt}S(t)y, v - S(t)y \right\rangle + \varphi(v) - \varphi(S(t)y) \\ &\geq \langle -F(S(t)y) + f(t), v - S(t)y\rangle, \quad \forall v \in \mathbb{R}^n, \text{a.e.}, \quad t \in [0,T].\end{aligned} \tag{34}$$

Thus,

$$\begin{aligned}&\left\langle \frac{d}{dt}S(t)y, \nabla V(S(t)y) \right\rangle + \varphi'(S(t)y; \nabla V(S(t)y)) \\ &\geq \langle -F(S(t)y) + f(t), \nabla V(S(t)y)\rangle, \quad \text{a.e.}, \quad t \in [0,T].\end{aligned} \tag{35}$$

Part (1) of this proof ensures that $\|S(t)y\| \geq R, \quad \forall t \in [0,T]$. The map $s \mapsto \langle F(S(s)y) - f(s), \nabla V(S(s)y)\rangle + \varphi'(S(s)y; \nabla V(S(s)y))$ is upper semi-continuous and (by condition (27)) strictly negative on $[0,T]$. Thus, using (35), we obtain

$$\begin{aligned}V(S(h)y) - V(y) &= \int_0^h \frac{d}{ds}V(S(s)y)ds \\ &= \int_0^h \langle \frac{d}{ds}S(s)y, \nabla V(S(s)y)\rangle \\ &\geq \int_0^h \langle -F(S(s)y) + f(s), \nabla V(S(s)y)\rangle \\ &\qquad -\varphi'(S(s)y; \nabla V(S(s)y))ds \\ &> 0\end{aligned}$$

This contradicts (33).
(7) Thanks to part (6) of this proof, we may use the invariance by homotopy property of the topological degree and see that

$$\begin{aligned}\deg(id_{\mathbb{R}^n} - S(T)\cdot, \mathbb{B}_r, 0) &= \deg(H(0,\cdot), \mathbb{B}_r, 0)\\ &= \deg(H(1,\cdot), \mathbb{B}_r, 0)\\ &= \deg(id_{\mathbb{R}^n} - S(\bar{T})\cdot, \mathbb{B}_r, 0).\end{aligned}$$

In conclusion, for all $r \geq r_0$, we have

$$\deg(id_{\mathbb{R}^n} - S(T)\cdot, \mathbb{B}_r, 0) = \deg(id_{\mathbb{R}^n} - S(\bar{T})\cdot, \mathbb{B}_r, 0)$$

and

$$\deg(id_{\mathbb{R}^n} - S(\bar{T})\cdot, \mathbb{B}_r, 0) = (-1)^n \deg(\nabla V, \mathbb{B}_r, 0).$$

Thus

$$\deg(id_{\mathbb{R}^n} - S(T)\cdot, \mathbb{B}_r, 0) = (-1)^n \mathrm{ind}(V, \infty).$$

Finally, for any $\tau \in [0, T]$, we have also (see Proposition 2.2 and Remark 5.1):

$$(-1)^n \mathrm{ind}(V, \infty) = \deg(id_{\mathbb{R}^n} - P_\varphi(id_{\mathbb{R}^n} - F + f(\tau)), \mathbb{B}_r, 0).$$

□

Corollary 5.1. *Suppose that $f \in C^0([0, +\infty); \mathbb{R}^n)$ and $\frac{df}{dt} \in L^1_{\mathrm{loc}}(0, +\infty; \mathbb{R}^n)$. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping satisfying the conditions of Theorem 4.1. Suppose in addition that there exist constants $C_1 \geq 0, C_2 \geq 0$ such that*

$$\langle F(x), x\rangle + \varphi'(x; x) \leq C_1\|x\|^2 + C_2\|x\|, \quad \forall x \in \mathbb{R}^n. \tag{36}$$

Let $T > 0$ be given. Assume that there exists $V \in C^1(\mathbb{R}^n; \mathbb{R})$ and $R > 0$ such that

$$\langle F(x) - f(t), \nabla V(x)\rangle + \varphi'(x; \nabla V(x)) < 0, \quad \forall x \in \mathbb{R}^n, \|x\| \geq R, t \in [0, T]. \tag{37}$$

and

$$\mathrm{ind}(V, \infty) \neq 0.$$

Then there exists at least one $u \in C^0([0, T]; \mathbb{R}^n)$ such that $\frac{du}{dt} \in L^\infty(0, T; \mathbb{R}^n)$,

$$u(0) = u(T); \tag{38}$$

$$\left\langle \frac{du}{dt}(t) + F(u(t)) - f(t), v - u(t)\right\rangle + \varphi(v) - \varphi(u(t)) \geq 0,$$
$$\forall v \in \mathbb{R}^n, \text{ a.e., } t \in [0, T]. \tag{39}$$

Proof: We may apply Theorem 5.1 with $W(x) = \frac{1}{2}\|x\|^2$. It results that for $r > 0$ large enough, we have $\deg(id_{\mathbb{R}^n} - S(T)\cdot, \mathbb{B}_r, 0) \neq 0$ and the existence of a fixed point for the Poincaré operator follows from the existence property of the topological degree. □

Corollary 5.2. *Suppose that $f \in C^0([0,+\infty);\mathbb{R}^n)$ and $\frac{df}{dt} \in L^1_{\text{loc}}(0,+\infty;\mathbb{R}^n)$. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex and Lipschitz continuous function. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping satisfying the conditions of Theorem 4.1. Suppose in addition that there exists $\Phi \in C^1(\mathbb{R}^n;\mathbb{R})$ such that*

$$\text{ind}(\Phi, \infty) \neq 0.$$

$$F(x) = \nabla\Phi(x), \quad \forall x \in \mathbb{R}^n,$$

and

$$c_1\|x\|^2 \leq c_2\|\nabla\Phi(x)\|^2 \leq \Phi(x) \leq c_3\|x\|^2, \quad \forall x \in \mathbb{R}^n$$

for some constants $c_1 > 0$, $c_2 > 0$ and $c_3 > 0$. Let $T > 0$ be given. There exists at least one $u \in C^0([0,T];\mathbb{R}^n)$ such that $\frac{du}{dt} \in L^\infty(0,T;\mathbb{R}^n)$,

$$u(0) = u(T); \tag{40}$$

$$\left\langle \frac{du}{dt}(t) + F(u(t)) - f(t), v - u(t) \right\rangle + \varphi(v) - \varphi(u(t)) \geq 0,$$
$$\forall v \in \mathbb{R}^n, \text{ a.e., } t \in [0,T]. \tag{41}$$

Proof: We may apply Theorem 5.1 with $W = \Phi$ and $V = -\Phi$. Indeed, let $K > 0$ denote the Lipschitz constant of φ. We have

$$\begin{aligned}\langle F(x), W(x)\rangle + \varphi'(x; W(x)) &= \|\nabla\Phi(x)\|^2 + \varphi'(x; \nabla\Phi(x)) \\ &\leq \|\nabla\Phi(x)\|^2 + K\|\nabla\Phi(x)\|, \quad \forall x \in \mathbb{R}^n.\end{aligned}$$

Moreover, let $C := K + \max_{t\in[0,T]} \|f(t)\|$ and $R > \frac{C}{c_1}\sqrt{\frac{c_3}{c_2}}$. We have

$$\begin{aligned}\langle F(x) - f(t), V(x)\rangle + \varphi'(x; V(x)) &= -\|\nabla\Phi(x)\|^2 + \langle f(t), \nabla\Phi(x)\rangle \\ &\quad + \varphi'(x; -\nabla\Phi(x)) \\ &\leq -\|\nabla\Phi(x)\|^2 + C\|\nabla\Phi(x)\| \\ &\leq -c_1\|x\|^2 + C\sqrt{\frac{c_3}{c_2}}\|x\| \\ &\leq \|x\|\left(C\sqrt{\frac{c_3}{c_2}} - c_1\|x\|\right) \\ &< 0, \forall x \in \mathbb{R}^n, \|x\| \geq R, t \in [0,T].\end{aligned}$$

□

Corollary 5.3. *Suppose that* $f \in C^0([0,+\infty);\mathbb{R}^n)$ *and* $\frac{df}{dt} \in L^1_{\text{loc}}(0,+\infty;\mathbb{R}^n)$. *Let* $\varphi : \mathbb{R}^n \to \mathbb{R}$ *be a convex and Lipschitz continuous function and let* $T > 0$ *be given. Let* $A \in \mathbb{R}^{n\times n}$ *be a real matrix and denote by* $\sigma(A)$ *the set of eigenvalues of* A. *If*

$$\mathrm{Re}(\lambda) > 0, \quad \forall \lambda \in \sigma(A)$$

then there exists at least one $u \in C^0([0,T];\mathbb{R}^n)$ *such that* $\frac{du}{dt} \in L^\infty(0,T;\mathbb{R}^n)$,

$$u(0) = u(T); \tag{42}$$

$$\left\langle \frac{du}{dt}(t) + Au(t) - f(t), v - u(t)\right\rangle + \varphi(v) - \varphi(u(t)) \geq 0,$$
$$\forall v \in \mathbb{R}^n, \text{ a.e., } t \in [0,T]. \tag{43}$$

Proof: Our assumption $\mathrm{Re}(\sigma(A)) \subset]0,+\infty[$ together with Lyapunov's Theorem ensure the existence of a positive definite matrix G such that

$$GA + A^T G = I.$$

Let us now define $V \in C^1(\mathbb{R}^n;\mathbb{R})$ by

$$V(x) := -\frac{1}{2}\langle (G+G^T)x, x\rangle, \quad \forall x \in \mathbb{R}^n.$$

Then

$$\nabla V(x) = -(G+G^T)x,$$

and

$$\begin{aligned}\langle Ax, \nabla V(x)\rangle &= -\langle Ax, Gx\rangle - \langle Ax, G^T x\rangle \\ &= -\langle x, A^T Gx\rangle - \langle GAx, x\rangle \\ &= -\|x\|^2.\end{aligned}$$

If we set $C = (\max_{s\in[0,T]} \|f(s)\| + K)\|G+G^T\|$ and if K denotes the Lipschitz constant of φ, then we have

$$\langle Ax - f(t), \nabla V(x)\rangle + \varphi'(x;\nabla V(x)) \leq -\|x\|^2 + C\|x\|, \quad \forall x \in \mathbb{R}^n, t \in [0,T].$$

Thus, for $R > 0$ large enough, condition (37) is satisfied. It is also clear that condition (36) holds (see Remark 5.2) and that all assumptions of Theorem 4.1 are satisfied. Moreover

$$\mathrm{ind}(V,\infty) = \mathrm{ind}(G+G^T,\infty) = \mathrm{sign}\det(G+G^T) \neq 0,$$

since $G+G^T$ is positive definite and thus nonsingular.

The conclusion follows from Corollary 5.1. □

Remark 5.2. Suppose that $\varphi : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant K. Then,

i) If F has a linear growth, i.e., there exist $c_1 \geq 0$ and $c_2 \geq 0$ such that

$$\|F(x)\| \leq c_1\|x\| + c_2, \quad \forall x \in \mathbb{R}^n,$$

then condition (36) of Corollary 5.1 holds with $C_1 = c_1$ and $C_2 = c_2 + K$.

ii) If $\langle F(x), x\rangle \leq 0, \quad \forall x \in \mathbb{R}^n$, then condition (36) of Corollary 5.1 holds with $C_1 = 0$ and $C_2 = K$.

Examples 5.1. i) Let $F : \mathbb{R}^n \to \mathbb{R}^n$ and $\varphi : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F(x) = x \text{ and } \varphi(x) = \|x\|, x \in \mathbb{R}^n.$$

We have

$$\langle F(x), x\rangle + \varphi'(x; x) \leq \|x\|^2 + \|x\|.$$

Hence condition (36) of Corollary 5.1 is satisfied with $C_1 = C_2 = 1$.

ii) If we take $F(x) = -x$ and $\varphi(x) = \|x\|$, then

$$\langle F(x), x\rangle + \varphi'(x; x) \leq -\|x\|^2 + \|x\| \leq \|x\|.$$

Hence condition (36) of Corollary 5.1 is satisfied with $C_1 = 0$ and $C_2 = 1$.

6 Second order periodic dynamical system with friction

Let us consider the following second order dynamical system with periodic conditions: For $(q_0, \dot{q}_0) \in \mathbb{R}^m \times \mathbb{R}^m$, we consider the problem $P(q_0, \dot{q}_0)$ of finding a function $t \mapsto q(t)$ $(t \in [0, T])$ with $q \in C^1([0, T]; \mathbb{R}^m)$, such that:

$$\frac{d^2q}{dt^2} \in L^\infty(0, T; \mathbb{R}^m), \tag{44}$$

$$\frac{dq}{dt} \text{ is right-differentiable on } [0, T], \tag{45}$$

$$q(0) = q(T) \text{ and } \dot{q}(0) = \dot{q}(T), \tag{46}$$

$$M\frac{d^2q}{dt^2}(t) + C\frac{dq}{dt}(t) + K(q(t)) - F(t) \in -H_1\partial\Phi(H_1^T\frac{dq}{dt}(t)), \text{ a.e. } t \in [0, T]. \tag{47}$$

In this problem $\Phi : \mathbb{R}^l \to \mathbb{R}$ is a convex function, $M \in \mathbb{R}^{m\times m}$ is a symmetric and positive definite matrix, $C \in \mathbb{R}^{m\times m}$ and $K \in \mathbb{R}^{m\times m}$ are given matrices and $H_1 \in \mathbb{R}^{m\times l}$ is a given matrix whose coefficients are related to the directions of friction forces. The function $F \in C^0([0, +\infty); \mathbb{R}^m)$ is such that $\frac{dF}{dt} \in L^1_{\text{loc}}([0, +\infty); \mathbb{R}^m)$.

The second order dynamical system (47) is useful for the study of many problems in unilateral mechanics. Indeed, the motion of various mechanical systems with frictional contact can be studied within the framework of equation (47). For such problems m is the number of degrees of freedom, M is the mass matrix, C is the viscous damping matrix and K is the stiffness matrix. The vector $q \in \mathbb{R}^m$ is the vector of generalised coordinates. The term $H_1 \partial\Phi(H_1^T \cdot)$ is used to model the unilaterality of the contact induced by the friction forces.

Since the matrix M is symmetric and positive definite, then problem (47) is equivalent to the first order variational inclusion:

$$\begin{cases} \dot{x}(t) + A(x(t)) - f(t) \in -\partial\varphi(x(t)), \\ x(0) = x(T), \end{cases}$$

where the vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n$ $(n = 2m)$ and the matrix $A \in \mathbb{R}^{n\times n}$ is defined by

$$A = \begin{pmatrix} 0_{m\times m} & -I_{m\times m} \\ M^{-\frac{1}{2}}KM^{-\frac{1}{2}} & M^{-\frac{1}{2}}CM^{-\frac{1}{2}} \end{pmatrix}, \tag{48}$$

with

$$x(0) = \begin{pmatrix} M^{-\frac{1}{2}}q(0) \\ M^{-\frac{1}{2}}\dot{q}(0) \end{pmatrix}, \quad x(T) = \begin{pmatrix} M^{-\frac{1}{2}}q(T) \\ M^{-\frac{1}{2}}\dot{q}(T) \end{pmatrix}, \tag{49}$$

$$f(t) = \begin{pmatrix} 0_{m\times m} \\ F(t) \end{pmatrix} \tag{50}$$

and the convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\varphi(x) = (\Phi \circ H_1^T M^{-\frac{1}{2}})(x_2). \tag{51}$$

In this case, let us observe that the subdifferential of φ is given by:

$$\partial\varphi(x) = \begin{pmatrix} 0_{m\times 1} \\ \partial(\Phi \circ H_1^T M^{-\frac{1}{2}})(x_2) \end{pmatrix} = \begin{pmatrix} 0_{m\times 1} \\ M^{-\frac{1}{2}} H_1 \partial\Phi(H_1^T M^{-\frac{1}{2}} x_2) \end{pmatrix}. \tag{52}$$

It is clear that A is continuous and $A + \omega I_{n\times n}$ is monotone provided that $\omega \geq \sup_{\|x\|=1} \langle -Ax, x\rangle$. A direct consequence of Corollary 5.3, we obtain an existence result for second order periodic systems.

Theorem 6.1. *If the function Φ is convex and Lipschitz continuous and* $\mathrm{Re}(\sigma(A)) \subset]0, +\infty[$, *then there exists at least one $q \in C^1(0, T; \mathbb{R}^n)$ such that $\frac{d^2q}{dt^2} \in L^\infty(0, T; \mathbb{R}^n)$ satisfying (46)-(47).*

Remark 6.1. We note that the conclusions of Theorem 6.1 hold under the key assumption Φ is convex and Lipschitz continuous. Let us now give a counterexample when Φ is convex but not Lipschitz. Let us take $m = 1$, $\Phi : \mathbb{R} \to \mathbb{R}$, $x \mapsto \Phi(x) = x^2$, $M = C = K = H_1 = 1$ and $F(t) = -t, \quad \forall t \in [0,1]$. The differential inclusion (47) then reduces to

$$\begin{cases} \ddot{q}(t) + 2\dot{q}(t) + q(t) = t, \\ q(0) = q(1), \\ \dot{q}(0) = \dot{q}(1). \end{cases} \tag{53}$$

We let the reader check that problem (53) has no solutions.

Example 6.1.

Let us take $m = 1$, $\Phi : \mathbb{R} \to \mathbb{R}$; $x \mapsto \Phi(x) = |x|$. In this case, we have

$$\partial\Phi(x) = \mathrm{Sign}(x)$$

where

$$\mathrm{sign}(x) := \begin{cases} -1 & \text{if } x < 0, \\ [-1,+1] & \text{if } x = 0, \\ +1 & \text{if } x > 0. \end{cases}$$

We consider the following problem

$$\begin{cases} m\ddot{q}(t) + c\dot{q}(t) + kq(t) - F(t) \in -\mathrm{Sign}(\dot{q}(t)), \ t \in [0,T], \\ q(0) = q(T), \\ \dot{q}(0) = \dot{q}(T), \end{cases} \tag{54}$$

with $F \in C^0([0,+\infty[;\mathbb{R})$ such that $\frac{dF}{dt} \in L^1_{\mathrm{loc}}([0,+\infty[;\mathbb{R})$. The matrix A in (48) is given by

$$A = \begin{pmatrix} 0 & -1 \\ \frac{k}{m} & \frac{c}{m} \end{pmatrix}.$$

We suppose that $m,\ c,\ k > 0$ and we set $\Delta = \frac{c^2}{m^2} - 4\frac{k}{m}$ and we have

$$\sigma(A) = \begin{cases} \{\frac{c}{2m} - \frac{\sqrt{\Delta}}{2}\ ,\ \frac{c}{2m} + \frac{\sqrt{\Delta}}{2}\} & \text{if } c \geq 2k\sqrt{m}, \\ \{\frac{c}{2m} - i\frac{\sqrt{-\Delta}}{2}\ ,\ \frac{c}{2m} + i\frac{\sqrt{-\Delta}}{2}\} & \text{if } c < 2k\sqrt{m}. \end{cases}$$

We note that in both cases $\mathrm{Re}(\sigma(A)) \subset]0,+\infty[$ and hence by Theorem 6.1, problem (54) has at least a solution.

References

[1] S. Adly, *Attractivity theory for second order dynamical systems with application to dry friction*, to appear in Journal of Math. Anal. and Appl. (2006).

[2] S. Adly, D. Goeleven, *A stability theory for second order non-smooth dynamical systems with application to friction problems*, J. Math. Pures Appl. **83** (2004), pp. 17-51.

[3] H. Amann, *A note on degree theory for gradient mappings*, Proc. Amer. Math. Soc. **85** (1982), pp. 591-595.

[4] H. Brézis, *Problèmes unilatéraux*, J. Math. Pures Appl. **51** (1972), 1-168.

[5] H. Brézis, Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert, North Holland, Amsterdam, 1973.

[6] F. de Blasi, L. G´rniewicz, G. Pianigiani, *Topological degree and periodic solutions of differential inclusions*, Nonlinear Anal. **37** (1999), no. 2, Ser. A: Theory Methods, pp. 217–243.

[7] D. Goeleven, B. Brogliato, *Necessary conditions of asymptotic stability for unilateral dynamical systems*, Nonlinear Analysis, **61**, (2005) pp. 961-1004.

[8] D. Goeleven, D. Motreanu, V. Motreanu, *On the stability of stationary solutions of first order evolution variational inequalities*, Advances in Nonlinear Variational Inequalities, **6**, (2003), pp. 1-30.

[9] D. Goeleven and B. Brogliato, 2002 "*Stability and Unstability Matrices for Linear Evolution Variational Inequalities*", IEEE Transaction on Automatic Control, **49**, (2004) pp. 1-14.

[10] L. Górniewicz, Topological Approach to Differential Inclusions, in Topological Methods in Differential Equations and Inclusions (ed. A. Granas and M. Frigon), NATO ASI Series, Mathematical and Physical Sciences, Vol. 472, Kluwer Academic Publishers, 1994.

[11] J.B. Hiriart-Urruty and C. Lemaréchal, Fundamentals of Convex Analysis, Springer, 1993.

[12] D. Kinderlehrer and G. Stampacchia, An introduction to variational inequalities and their applications, Academic Press, 1980.

[13] M.A. Krasonel'skii, The Operator of Translation along the Trajectories of Differential Equations, Nauka, Moscow (1966) (Russian); English translation: American Math. Soc., Translations of math. Monographs, vol. 19, Providence, 1968.

[14] M.A. Krasonel'skii and P.P. Zabreiko, Geometrical Methods of Nonlinear Analysis, Nauka, Moscow (1975) (Russian); English translation: Springer, Berlin, 1984.

[15] J.L. Lions, Quelques méthodes de résolution des problèmes aux limites non-linéaires, Dunod, 1969.

[16] N.G. Lloyd, Degree Theory, Cambridge University Press, 1978.

[17] J. Mawhin, Continuation Theorems and Periodic Solutions of Ordinary Differential Equations, in Topological Methods in Differential Equations and Inclusions (ed. A. Granas and M. Frigon), NATO ASI Series, Mathematical and Physical Sciences, Vol. 472, Kluwer Academic Publishers, 1994.

[18] R.E. Showalter, Monotone Operators in Banach Spaces and Nonlinear Partial Differential Equations, American Mathematical Society, USA, 1997.

[19] E. Zeidler, Nonlinear Functional Analysis and its Applications, I: Fixed-Point Theorems, Springer-Verlag, New York, 1986.

Complementarity Problems: An Overview on Existing Verification Procedures

Götz Alefeld
Institut für Angewandte Mathematik,
Universität Karlsruhe, 76128 Karlsruhe, Germany.
Email:goetz.alefeld@math.uni-karlsruhe.de

Abstract

In this paper we give an overview on verification procedures for the solution of complementarity problems.

2000 MR Subject Classification 90C33, 65G30, 65K10
Keywords Complementarity problem, enclosure of solutions, verification, interval arithmetic

1 Introduction

Complementarity problems are finding more and more attention in applications as well as from a mathematical point of view. Especially numerical methods for solving these problems are of great interest. As it is usually the case with numerical methods the computed result delivers an approximation and it has to be confirmed somehow that a solution of the given problem really exists close to the output of the computer. This confirmation is done by the computer and is called verification.

In this survey paper we are mainly concerned with this last aspect. After formulating the complementarity problem and mentioning some well-known applications, we show by a simple example, that verification is a must. We continue by introducing a general verification procedure for complementarity problems. After that we show that for a special class of problems there exists a simple verification procedure. Finally we consider problems with interval data which from a practical point of view always have to be considered if one is taking into account rounding errors, for example.

2 Complementarity Problems and Applications

Given a (nonlinear) mapping $l : \mathbb{R}^n_+ \to \mathbb{R}^n$, where $\mathbb{R}^n_+$ denotes the set of vectors with nonnegative components, the problem

"Find $z \in \mathbb{R}^n$ such that

$$\left.\begin{aligned} z &\geq 0 \\ l(z) &\geq 0 \\ z^T l(z) &= 0 \end{aligned}\right\} \tag{2.1}$$

(or to show that no such z exists)" is called (nonlinear) complementarity problem (NCP).

Equivalent formulations are:

"Find $w, z \in \mathbb{R}$" such that

$$\left.\begin{aligned} w \geq 0, z \geq 0 \\ w = l(z) \\ z^T w = 0 \end{aligned}\right\} \tag{2.2}$$

(or to show that no such w, z exist)",

or:

"Find $z \geq 0$ such that

$$g(z) = \min\,(z, l(z)) = 0\,, \tag{2.3}$$

where the minimum is taken componentwite (or to show that no such z exists)".

If

$$l(z) = Mz + q,$$

where $M \in \mathbb{R}^{n \times n}$ and $q \in \mathbb{R}^n$ is a given matrix and a given vector, respectively, then the problem is called linear (LCP).

Many problems in science, engineering and economics either arise naturally or can be reformulated as a complementarity problem. We consider a problem from optimization.

Example 1 (Quadratic programming (QP))

"Minimize

$$f(x) = c^T x + \frac{1}{2}\, x^T Q x$$

subject to

$$Ax \geq b$$

$$x \geq 0",$$

where $Q \in \mathbb{R}^{n \times n}$ is symmetric, $c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. If $Q = O$, then we have a so-called linear programming problem.

It is well-known that if $x \in \mathbb{R}^n$ is a locally optimal solution, then there exists a $y \in \mathbb{R}^n$ such that $(x, y)^T$ satisfies the Karush-Kuhn-Tucker conditions

$$\left.\begin{array}{l} u = c + Qx - A^T y \geq 0 \,,\, x \geq 0 \,,\; x^T u = 0 \\ v = -\, b + Ax \geq 0 \qquad\;\; ,\, y \geq 0 \,,\; y^T v = 0 \end{array}\right\}. \tag{2.4}$$

Defining the block matrix

$$M = \begin{pmatrix} Q & A^T \\ A & O \end{pmatrix}$$

and the block vectors

$$q = \begin{pmatrix} c \\ -\, b \end{pmatrix} \,,\; z = \begin{pmatrix} x \\ y \end{pmatrix},$$

respectively, (2.4) can be written as

$$\left.\begin{array}{l} z \geq 0 \\ Mz + q \geq 0 \\ z^T(Mz + q) = 0 \end{array}\right\},$$

which is an (LCP).

If Q is not only symmetric but also positive semi-definite (i.e., $f(x)$ is convex), then (2.4) is not only necessary but also sufficient for x to be a globally optimal solution of the (QP).

A series of further problems which lead to or which can be formulated as a complementarity problem, are the following: contact problem, porous flow problem, obstacle problem, journal bearing problem, elastic plastic torsion problem. For details see [7].

3 Verification of Solutions of Complementarity Problems

We start with a simple example which shows that verification of solutions is a must if one has computed an approximation.

Example 2 (see [1])

Let $\quad l(z) = Mz + q$ where

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix} \text{ and } q = \begin{pmatrix} 2 \\ 1 \\ -\, 10^{-6} \end{pmatrix}.$$

We consider the formulation (2.2) of the complementarity problem.

For the "approximate solution"

$$z = \begin{pmatrix} 10^{-6} \\ 10^{-6} \\ 1 \end{pmatrix}, \quad w = \begin{pmatrix} 3 \\ 2 \\ 10^{-6} \end{pmatrix}, \tag{3.1}$$

one obtains

$$\| \, l(z) \; - \; w \, \|_\infty = \; \| \, Mz + q - w \, \|_\infty = 4 \cdot 10^{-\,6}$$

and

$$z^T w = 6 \cdot 10^{-\,6}.$$

In many iterative methods (e.g. for interior-point-methods) the condition

$$\max \{ z^T w \; , \; \| \, Mz + q \; - \; w \, \|_\infty \} \leq \epsilon$$

is used as a stopping criteria for some fixed given ϵ. A pair $(z, w)^T$ which fulfils this inequality is then called an "ϵ-approximate solution". In this sense the given vectors z, w form a $6 \cdot 10^{-\,6}$-approximate solution. However, it can be shown that there is no exact solution of the (LCP) within an $\| \cdot \|_\infty$ distance of 0.25 from this ϵ-approximate solution with $\epsilon = 6 \cdot 10^{-\,6}$. We will come back to this statement later, again.

Our starting point for the verification of solutions is the equivalent formulation (2.3) of a complementarity problem.

However, we start by explaining the general idea, which is independent of the underlying equation (2.3):

Assume that we have given a continuous mapping

$$H : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$$

and an interval vector $[x] \subseteq D$. A mapping

$$\delta H : [x] \times [x] \to \mathbb{R}^{n \times n}$$

is called slope, if

(A) $\qquad H(x) - H(y) = \delta H(x, y)(x - y) \; , \; x, y \in [x]$

holds. Assume that there exists an interval matrix $\delta H(x, [x])$ such that

(B) $\qquad \delta H(x, y) \in \delta H(x, [x])$

for all $y \in [x]$ and some fixed $x \in [x]$.

Define for a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ the interval vector

$$L(x, A, [x]) := x - A^{-1}H(x) + (I - A^{-1}\delta H(x, [x]))([x] - x).$$

Then, the following hold:

a) If $L(x, A, [x]) \subseteq [x]$, then there exists an $x^* \in [x]$ such that $H(x^*) = 0$.

b) If $L(x, A, [x]) \cap [x] = \emptyset$ (empty set), then $H(x) \neq 0$ for all $x \in [x]$.

In Moore [9] these results were obviously the first time applied to verify the existence of a zero x^* of a mapping H in a given interval vector $[x]$. A proof of the statements a) and b), can also be found in [2]. We are now going to apply these statements to the problem (2.3), that is we replace the general case $H(x) = 0$
by

$$g(z) = \min(z, l(z)) = 0,$$

where $l(z)$ is given. First of all we have to find a slope $\delta g(x, y)$ for which

$$\text{(A)} \qquad g(x) - g(y) = \delta g(x, y)$$

for all $x, y \in [x]$ holds, where $[x]$ is some given interval vector. Forthermore, we have to bound the slope by an interval matrix $\delta g(x, [x])$:

$$\text{(B)} \qquad \delta g(x, y) \in \delta g(x, [x])$$

for some fixed $x \in [x]$ and all $y \in [x]$.

We first consider (A): Define for each $i \in \{1, 2, \ldots, n\}$ the sets

$$S_i^+ = \{x \in [x] |\ l_i(x) > x_i\},$$
$$S_i^- = \{x \in [x] |\ l_i(x) < x_i\},$$
$$S_i^0 = \{x \in [x] |\ l_i(x) = x_i\},$$

where $l(x) = (l_i(x))$. Then

$$g_i(x) = \begin{cases} x_i, & x \in S_i^+ \\ l_i(x), & x \in S_i^- \\ l_i(x) = x_i, & x \in S_i^0. \end{cases}$$

Using this representation of $g_i(x)$ we obtain the nine cases from table 1 for the *i-th* row $\delta g_i(x, y)$ of the slope matrix $\delta g(x, y)$, where e_i denotes the *i-th* unit vector.

Table 1: $\delta g_i(x,y)$

$x \setminus y$	S_i^+	S_i^-	S_i^0
S_i^+	e_i^T	$\alpha_i(\delta l_i(x,y) - e_i^T) + e_i^T$	e_i^T
S_i^-	$\beta_i(\delta l_i(x,y) - e_i^T) + e_i^T$	$\delta l_i(x,y)$	$\delta l_i(x,y)$
S_i^0	e_i^T	$\delta l_i(x,y)$	e_i^T

Furthermore, $\delta l_i(x,y)$ denotes a slope (vector) of the *i-th* component of the mapping $l(x)$:

$$l_i(x) - l_i(y) = \delta l_i(x,y)(x-y),$$

and

$$\alpha_i = \frac{y_i - l_i(y)}{(\delta l_i(x,y) - e_i^T)(x-y)},$$

$$\beta_i = \frac{l_i(x) - x_i}{\delta l_i(x,y) - e_i^T)(x-y)}.$$

We prove this for the case $x \in S_i^+, y \in S_i^-$, for example. The proof for the remaining cases can be performed similarly. If $x \in S_i^+, y \in S_i^-$, then

$$\begin{aligned} g_i(x) - g_i(y) &= x_i - l_i(y) \\ &= y_i - l_i(y) + e_i^T(x-y) \\ &= \frac{y_i - l_i(y)(\delta l_i(x,y) - e_i^T)(x-y)}{(\delta l_i(x,y) - e_i^T)(x-y)} + e_i^T(x-y) \\ &= (\alpha_i(\delta l_i(x,y) - e_i^T) + e_i^T)(x-y) \\ &= \delta g_i(x,y)(x-y). \end{aligned}$$

Concerning (B), let $x \in [x]$ be fixed and consider the nonlinear programming problems

$$\min_{y \in [x]} \{y_i - l_i(y)\}$$

and

$$\max_{y\in[x]}\{y_i - l_i(y)\}.$$

Let $y^{i,1}$ and $y^{1,2}$, respectively, be solutions of these problems, then if

$$\delta g_i(x,[x]) = \begin{cases} e_i^T\ , & y^{i,2} \in S_i^+ \cup S_i^0 \\ \delta l_i(x,[x])\ , & y^{i,1} \in S_i^- \cup S_i^0 \\ [0,\alpha_i](\delta l_i(x,[x]) - e_i^T) + e_i^T, & x \in S_i^+ \cup S_i^0, \\ & y^{i,2} \in S_i^- \\ [\beta_i,1](\delta l_i(x,[x]) - e_i^T) + e_i^T, & x \in S_i^-, \\ & y^{i,1} \in S_i^+ \end{cases}$$

(B) holds. For details see [1].

Exploiting the preceding ideas we get the following algorithm for the verification of a solution:

Algorithm:

Let $r > 0$ be a given tolerance and let x be an approximate solution of

$$g\ (z) = \min(z, l(z)) = 0.$$

Define the interval vector

$$[x] = (x + r\ [-e, e]) \cap \mathbb{R}_+$$

where $e = (1, 1, \ldots, 1)^T$.

Choose a nonsingular A and compute $L(x, A, [x])$.

If $L(x, A, [x]) \subseteq [x]$ then $g(z^*) = 0$ for some $z^* \in [x]$.

If $L(x, A, [x]) \cap [x] = \emptyset$(empty set), then $g(z) \neq 0$ for all $z \in [x]$.

In [1] this algorithm was applied in order to prove the conclusion from example 1. In [3] there are also numerical results for the nonlinear case.

4 Special Complementarity Problems

In certain cases it is possible to avoid the laborous bounding of the slope given by the elements of table 1. Starting again with the formulation (2.3), we easily see that a nonnegative solution of (2.3) is a fixed point of the mapping

$$p\ :\ \mathbb{R}_+^n \to \mathbb{R}_+^n$$

where

$$p(x) = \max\{0, x - D \cdot l(x)\}$$

and where $D = \text{diag}\ (d_i)$ is a diagonal matrix with positive elements in the main diagonal. The maximum is again taken componentwise. Please note that the preceding statement is not true in general for a nonsingular, nonnegative matrix, which is not a diagonal matrix.

Assume now that l' has a so-called interval arithmetic evaluation $l'([x])$, that is $l'(x) \in l'([x])$ holds for all $x \in [x]$ and an interval vector [x]. (Such an $l'([x])$ exists and can be computed if the Jacobian of l exists and can be evaluated for the given interval vector. For details see any book on interval arithmetic.)

Let now [x] be given. Let $l'([x])$ denote the interval arithmetic evaluation of l' over $[x]$. If

$$\Gamma(x, [x], D) := \max\{0, x - D \cdot l[x] + (I - Dl'([x]))([x] - x)\}$$

is contained in $[x], \Gamma(x, [x], D) \subseteq [x]$, where $x \in [x]$ is fixed, then there exists a solution z^* of (2.3) in $\Gamma(x, [x], D)$ (and therefore also in $[x]$). All solutions of (2.3) contained in $[x]$ are also contained in $\Gamma(x, [x], D)$. From the last statement it follows that there is no solution of (2.3) in $[x]$ if $\Gamma(x, [x], D) \cap [x] = \emptyset$. (In the definition of $\Gamma(x, [x], D)$, there appears an interval vector $[x]$ and the maximum of the zero vector 0 and an interval vector, say $[y]$, has to be performed. This maximum is defined as follows: Let the interval vector $[y]$ have the lower bound $\underline{y}$ and the upper bound $\overline{y}$, respectively. Then $\max\{0, [y]\}$ is the interval vector with lower bound $\max\{0, \underline{y}\}$ and upper bound $\max\{0, \overline{y}\}$, respectively. The maximum of two real vectors is formed componentwise.)

If the problem is linear,

$$l(z) = Mz + q$$

with a given H-matrix M with positive diagonal elements, it is easy to find an interval vector $[x]$ and a diagonal matrix with positive elements in the main diagonal such that $\Gamma(x, [x], D) \subseteq [x]$. Furthermore we have a simple iterative method which computes a sequence of interval vectors, all containing the (in this case for arbitrary q) unique solution z^*, and converging to z^*.

We first repeat some properties of H-matrices. Then we consider the iterative method and finally we construct an interval vector containing z^*. Define for the given matrix $M = (m_{ij})$ the so-called comparison

matrix $\tilde{M} = (\tilde{m}_{ij})$
by

$$\tilde{m}_{ij} = \begin{cases} |m_{ii}| \text{ if } i = j \\ -|m_{ij}| \text{ if } i \neq j \end{cases}.$$

M is called an H-matrix iff there exists a positive vector d such that $\tilde{M}d > 0$.

The diagonal elements of an H-matrix are different from zero. Therefore, if M is real, they are either positive or negative. If the diagonal elements of a (real) H-matrix are all positive, then this matrix belongs to the set of so-called P-matrices (see [7]). If M is a P-matrix then the (LCP) has a unique solution for each q.

We now assume that M is an H-matrix with positive diagonal elements $m_{ii} > 0, i = 1, 2, \ldots, n$. Since $l'(x) = M$ in this case, we obtain

$$\Gamma(x, [x], D) = \max\{0, x - D(Mx + q) + (I - DM)([x] - x)\}.$$

Assume now that

$$\Gamma(x, [x], D) \subseteq [x]$$

for some given $[x]$. Define $[x^0] := [x]$ and consider the iterative method

$$[x^{k+1}] = \Gamma(x^k, [x^k], D) \cap [x^k] \tag{4.1}$$

where

$$x^k = m([x^k])$$

is the center of $[x^k]$.

Then the following hold:

a) If the (unique) solution z^* of (2.3) is contained in $[x^0]$ then the iteration method (4.1) is well-defined and $\lim_{k\to\infty} [x^k] = z^*$.

b) If $z^* \notin [x^0]$ then the intersection becomes empty after a finite number of steps.

For details see [5], where also the method for finding $[x^0]$ with $z^* \in [x^0]$ is discussed. In that paper also a couple of numerical examples are presented.

5 Interval Data

Representing the given data of a given complementarity problem on a computer usually implies rounding of the data. Hence on the computer

a problem different from the given one has to be solved. One can overcome this difficulty by enclosing the data by intervals on the computer and by considering a whole set of complementarity problems. Then the problem consists of computing the set of the solutions of all problems.

Another interesting case, in which interval data have to be considered, was given by Schäfer [10,11]. He showed that discretizing a certain free boundary problem leads in a natural manner to a complementarity problem with interval data if the discretization error is taken into account.

After the formulation of the problem we consider some iterative methods for the inclusion of the solution set and show convergence of these methods for a certain class of problems.

Let there be given an interval matrix $[M] = ([m_{ij}]) \in \mathbb{IR}^{n\times n}$ and an interval vector $[q] \in ([q_i]) \in \mathbb{IR}^n$. Then we consider the set $\sum$ of the solutions of all possible linear complementarity problems:

$$\sum := \{z \in \mathbb{R}^n \mid z \geq 0\,,\ q+Mz \geq 0\,,\ z^T(q+Mz) = 0\,,\ M \in [M]\,,\ q \in [q]\}.$$

The problem consists of finding an (as small as possible) interval vector $[x] \in \mathbb{IR}^n$, such that $\sum \subseteq [x]$.

We now consider the case, that the given interval matrix is a so-called (interval) H-matrix. For the definition of an H-matrix in the interval case we generalize first the definition of the comparison matrix of a given matrix (see section 3).

Given the interval matrix $[M] = ([m_{ij}])$, where $[m_{ij}] = [\underline{m}_{ij}, \overline{m}_{ij}]$, the comparison matrix $\tilde{M} \in \mathbb{R}^{n\times n}$ is a real matrix with

$$\tilde{m}_{ij} = \begin{cases} min\{|m_{ij}| \mid m_{ij} \in [m_{ij}]\}, & i = j \\ -\ |[m_{ij}]|, & i \neq j, \end{cases}$$

where the absolute value of the interval $[m_{ij}]$ is a real number defined by $|\ [m_{ij}]\ | = \max\ \{|\ \underline{m}_{ij}\ |\ ,\ |\ \overline{m}_{ij}\ |\}$. The interval matrix $[M]$ is called interval H-matrix if the comparison matrix is an H-matrix, that is, if there exists a positive vector, such that $\tilde{M}d > 0$.

Now assume that the given interval matrix $[M] = ([m_{ij}])$ is an interval H-matrix with $\underline{m}_{ii} > 0, i = 1, 2, \ldots, n$. We split $[M]$ into its diagonal part $[D]$ and its off-diagonal part $-[R] : [M] = [D] - [R]$. Define the diagonal matrix $[\hat{D}]$ by

$$[\hat{D}]\ := \text{diag}\ (\frac{1}{[m_{ii}]}).$$

Then we consider the iteration method

$$[x^0] = [\underline{x}^0, \overline{x}^0] \in \mathbb{IR}^n, \underline{x}^0 \geq 0$$

$$[x^{k+1}] = \max\{0\ ,\ [\hat{D}]([R]\ [x^k] - [q]\}\ ,\ k = 0, 1, 2, \ldots\ . \qquad \text{(T)}$$

(T) has the following properties:

a) $\lim_{k\to\infty} [x^k] = [x^*]$ with

$[x^*] = \max\{0\ ,\ [\hat{D}]\ ([R]\ [x^*]\ -\ [q])\}$;

$[x^*]$ is unique.

b) $\sum \subseteq [x^*]$.

The proof of a) follows by application of the Banach fixed point theorem to the mapping

$$g([x]) = \max\{0, [\hat{D}]\ ([R]\ [x]\ -\ [q])\}$$

using the fact that the spectral radius of the real matrix $|\,[\hat{D}]|\,|\,[R]|$ is less than one. The second part b) follows by simple interval arithmetic manipulations.

The introduced method (T) can be considered as a straight forward generalization of the well-known total step method to complementarity problems (with interval data). Another choice would be to consider the Gauss-Seidel-method (GS) and its generalization. It has been shown in [4] that under the conditions assumed for (T), (GS) is also convergent (to the same limit $[x^*]$ as (T)). Introducing a so-called relaxation parameter ω in order to speed up (GS) by the successive over relaxation method (SOR) is not advantageous if interval data are considered since in this case the fixed point $[x^*]$ of (T) may be inflated for certain values of ω. For details see [4].

It is interesting to raise the question whether the limit $[x^*]$ of (T) is the smallest interval vector (with respect to inclusion) for which $\sum \subseteq [x^*]$ holds. In general this is not the case under our assumption that $[M]$ is an interval matrix with $\underline{m}_{ii} > 0$, $i = 1, 2, \ldots, n$. However, by restricting the set of matrices further, this is actually true: If we assume that for the given interval H-matrix not only $\underline{m}_{ii} > 0$ but also $\overline{m}_{ij} \leq 0$, $i \neq j$ (such a matrix is called an interval M(inkowski)-matrix), then $[x^*]$ is optimal in the sense that there exists no smaller interval vector with

respect to inclusion such that b) holds.

For a proof see [4], where also a couple of numerical examples with interval data can be found.

6 Newton's Method

In this section we apply the idea of Newton's method to the (NCP) from (2.1).

Define the linearized mapping l^k by

$$l^k(z) = l(z^k) + l'(z^k)(z - z^k).$$

Then we consider the following algorithm, which may be considered as a generalization of Newton's method for the solution of nonlinear equations:

$$(NM) \begin{cases} \text{Choose } z^0; \\ \text{For } k = 0, 1, \ldots \text{ compute the solution } z^{k+1} \text{ of the (LCP)} \\ z \geq 0 \ , \ l^k(z) \geq 0 \ , \ z^T l^k(z) = 0. \end{cases}$$

Concerning the existence and convergence of the sequence $\{z^k\}_{k=0}^{\infty}$ the following result has been proven by Z. Wang [12]:

Let $D \subseteq \mathbb{R}^n$ be open and $\mathbb{R}^n_+ \subseteq D$. Assume that $l : D \longrightarrow \mathbb{R}^n$ is differentiable, $D_0 \subseteq D$ is convex and that

$$||l'(x) - l'(y)||_\infty \leq \gamma ||x - y||_\infty, x, y \in D_0.$$

Suppose there exists a starting point $z^0 \in D_0$ such that $l'(z^0)$ is an H-matrix with positive diagonal elements and

$$||\overline{l'(z^0)}^{-1}||_\infty \leq \beta,$$

where the bar denotes the comparison matrix. Denote by z^1 the first iterate of (NM) starting with z^0. Let $||z^1 - z^0||_\infty \leq \eta$. If $h = \beta\gamma\eta \leq \frac{1}{2}$ and $\overline{S}(z^0, r^*) \subseteq D_0$, where $r^* = (1 - \sqrt{1 - 2h})/\beta\gamma$, then the sequence $\{z^k\}$ is well defined, remains in $\overline{S}(z^0, r^*)$ and converges to a solution z^* of (NCP) which exists in the ball $\overline{S}(z^0, r^*)$. z^* is unique in the ball $S(z^0, r^{**})$, where $r^{**} = (1 + \sqrt{1 - 2h})/\beta\gamma$. The error estimation

$$||z^k - z^*||_\infty \leq \frac{1}{2^{k-1}}(2h)^{2^k - 1}\eta = (\beta\gamma 2^k)^{-1}(2h)^{2^k}$$

holds. The preceding statements also hold for the 1-norm.

If $l'(x)$ is positive definite, let $\widetilde{l'(x)} = \frac{1}{2}(l'(x) + (l'(x))^T)$. If $\overline{l'(x^0)}$ is replaced by $\widetilde{l'(x^0)}$, the preceding statements also hold in the 2-norm.

Using the error estimation for $z^k - z^*$, the verification of a solution in a ball can be verified.

7 Further Results

In the paper [8], Mathias and Pang have shown that for the unique solution z^* of the (linear) complementarity problem (2.1) with $l(z) = Mz + q$, where M is a so-called P-matrix, the estimation

$$||z - z^*||_\infty \leq \frac{1 + ||M||_\infty}{c(M)} ||r(z)||_\infty$$

with

$$r(z) = \min\{z, Mz + q\},$$

$$c(M) = \min_{||z||_\infty = 1} \{\max_{1 \leq i \leq n} z_i (Mz_i)_i\}$$

holds for all $z \in \mathbb{R}^n$. However, $c(M)$ is not easy to find for a given approximation z of z^*. Therefore the following estimation by X. Chen and S. Xiang [6] is of great importance: For each $p \geq 1$ it holds for all $z \in \mathbb{R}^n$

$$||z - z^*||_p \leq \max_{D \in [0,1]^n} ||(I - D + DM)^{-1}||_p ||r(z)||_p,$$

where $D = \text{diag}\,(d_1, \ldots, d_n)$. Furthermore

$$\max_{D \in [0,1]^n} ||(I - D + DM)^{-1}||_\infty$$

$$\leq \frac{\max\{1, ||M||_\infty\}}{c(M)} = \frac{1 + ||M||_\infty}{c(M)} - \frac{\min\{1, ||M||_\infty\}}{c(M)}.$$

8 Acknowledgement

The author is grateful to Dr. U. Schäfer for reading the paper and for helpful comments which improved the presentation.

References

[1] G. E. Alefeld, X. Chen, F. A. Potra: Numerical validation of solutions of linear complementarity problems. Numer. Math. 83, 1-23 (1999).

[2] G E. Alefeld, G. Mayer: Interval Analysis: Theory and Applications. Journal of Comp. and Appl. Math. 121, 421-464 (2000).

[3] G. E. Alefeld, X. Chen, F. A. Potra: Numerical Validation of solutions of complementarity problems: The nonlinear case. Numer. Math 92, 1-16 (2002).

[4] G. E. Alefeld, U. Schäfer: Iterative Methods for Linear Complementarity Problems with Interval Data. Computing 70, 25-259 (2003).

[5] G. E. Alefeld, Z. Wang, Z. Shen: Enclosing solutions of linear complementarity problems for H-matrices. Reliable Computing 10, 423-435 (2004).

[6] X. Chen, S. Xiang: Computation of error bounds for P-matrix linear complementarity problems. Math. Programming, 106, 513-525 (2006).

[7] R. W. Cottle, J.-S. Pang, R. E. Stone: The linear complementarity problem. Academic Press, 1992.

[8] R. Mathias, J. S. Prang: Error bounds for the linear complementarity problem with a P-matrix. Linear Algebra Appl., 132, 123-136 (1990).

[9] R. E. Moore: A Test for Existence of Solutions to Nonlinear Systems. SIAM J. Numer. Anal. 14, 611-615 (1977).

[10] U. Schäfer: Das lineare Komplementaritätsproblem mit Intervalleinträgen. Dissertation. University of Karlsruhe (1999).

[11] U. Schäfer: An enclosure method for free boundary problems based on a linear complementarity problem with interval data. Numer. Funct. Anal. Optim. 22, 991-1011 (2001).

[12] Z. Wang: Semilocal Convergence of Newton's Method for Finite-Dimensional Variational Inequalities and Nonlinear Complementarity Problems. PhD-Thesis. Fakultät für Mathematik. University of Karlsruhe (2005).

Synchrony in Lattice Differential Equations*

Fernando Antoneli
Department of Applied Mathematics, University of São Paulo
São Paulo, SP 05508-090, Brazil.
E-mail: antoneli@ime.usp.br

Ana Paula S. Dias
Centro de Matemática, Universidade do Porto
Porto, 4169-007, Portugal.
E-mail: apdias@fc.up.pt

Martin Golubitsky
Department of Mathematics, University of Houston
Houston, TX 77204-3008, USA.
E-mail: mg@uh.edu

Yunjiao Wang
Department of Mathematics, University of Houston
Houston, TX 77204-3008, USA.
E-mail: yunjiao@math.uh.edu

Abstract

We survey the recent results on patterns of synchrony in lattice differential equations on a square lattice. Lattice differential equations consist of choosing a phase space $\mathbf{R}^m$ for each point in a lattice and a system of differential equations on each of these phase spaces such that the whole system is translation invariant. The architecture of a lattice differential equation is the specification of which sites are coupled to which (the nearest neighbor coupling is a standard example). A polydiagonal is a finite-dimensional subspace obtained by setting coordinates in different phase spaces equal. A polydiagonal Δ has k colors if points in Δ have at most k unequal cell coordinates. A *pattern of synchrony* is a polydiagonal that is flow-invariant for every lattice differential equation with a given architecture. We survey two main results: the classification of two-color patterns of synchrony and the fact that every pattern of synchrony for a fixed architecture is spatially doubly

*This work was supported in part by NSF Grant DMS-0244529. The work of FA was supported in part by a FAPESP Grant 03/12631-3.

periodic assuming that the architecture includes both the nearest and the next nearest neighbor couplings.

1 Introduction

In this paper we survey the recent results on lattice differential equations (LDE) [6, 11, 12, 2, 1] that show that certain characteristics of the dynamics of LDE depend significantly on whether the coupling is the nearest neighbor coupling or a more general coupling. In our exposition we focus on planar square lattices. We show that patterns of synchrony that occur naturally in these systems can be aperiodic in the case of the nearest neighbor (NN) coupling and must be spatially doubly periodic when both the nearest and the next nearest coupling (NNN) are present. The framework of our work is a general theory of patterns of synchrony in coupled systems developed by Stewart *et al.* [10, 9]. The dynamics of LDE are also discussed in Chow *et al.* [3, 4, 5].

We index the planar square lattice $\mathcal{L} \cong \mathbf{Z}^2$ by pairs of integers (i, j) and call each lattice point a *cell.* The architecture of a lattice dynamical system specifies which cells are coupled to which, which cells have the same dynamics, and which couplings are the same. The set $I(c)$, the *input set* of a cell $c \in \mathcal{L}$, is the set of all cells coupled to c. For example, with NN coupling the input set of cell (i, j) is

$$I(i,j) = \{(i+1,j),(i-1,j),(i,j+1),(i,j-1)\}$$

whereas, with NNN coupling the input set is

$$\begin{aligned} I(i,j) = {} & \{(i+1,j),(i-1,j),(i,j+1),(i,j-1)\} \cup \\ & \{(i+1,j+1),(i-1,j+1),(i+1,j-1),(i-1,j-1)\} \end{aligned}$$

See Figure 1.1.

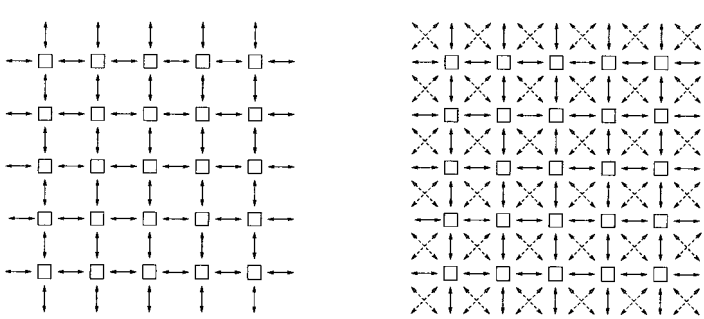


Figure 1.1 (Left) square lattice network with NN coupling (—). (Right) square lattice network with NN and NNN couplings (- - -).

We assume that the phase space of cell c is $\mathbf{R}^m$ and we denote the coordinates in this phase space by x_c. We consider only those LDE that are equivariant with respect to the symmetries of the lattice. Square lattice symmetries are generated by translations within the lattice ($\mathcal{L}$ itself) and rotations and reflections of the origin that preserve the lattice. The latter group is isomorphic to $\mathbf{D}_4$, the eight element symmetry group of the square. Thus the group of symmetries of the square lattice is

$$\mathcal{G}_{\mathcal{L}} = \mathbf{D}_4 \dot{+} \mathcal{L}$$

Equivariance with respect to $\mathcal{L}$ requires that m is constant for all cells.

Suppose that the input set $I(c) = \{c_1, \ldots, c_p\}$. Then we denote the coordinates in the input set, the *coupling variables*, by

$$x_{I(c)} = (x_{c_1}, \ldots, x_{c_p}) \in (\mathbf{R}^m)^p$$

With this notation a *lattice differential equation* is a coupled system of differential equations of the form

$$\dot{x}_{i,j} = g_{i,j}(x_{i,j}, x_{I(i,j)})$$

where $(i,j) \in \mathcal{L}$, $x_{i,j} \in \mathbf{R}^m$, and $g_{i,j}$ is a function that depends on the internal cell variable $x_{i,j}$ and the coupling variables $x_{I(i,j)}$. Moreover, equivariance with respect to translations implies that $g_{i,j} = g$ is independent of the cell.

With NN coupling the square symmetry $\mathbf{D}_4$ forces g to be invariant under permutations of the coupling variables. Thus square lattice differential equations with the nearest neighbor coupling have the form

$$\dot{x}_{i,j} = g(x_{i,j}, \overline{x_{i+1,j}, x_{i-1,j}, x_{i,j+1}, x_{i,j-1}}) \tag{1.1}$$

where g is invariant under all permutations of the coupling variables under the bar.

Synchrony is one of the most interesting features of coupled cell systems and in order to study it, we need to formalize the concept. We use a strong form of network synchrony, which we now define. A *polydiagonal* Δ is a subspace of the phase space of a coupled cell system that is defined by equality of cells coordinates.

Definition 1.1. A polydiagonal Δ is a *pattern of synchrony* if Δ is flow-invariant for every coupled cell system with the given network architecture.

Solutions in a flow-invariant polydiagonal Δ have sets of coordinates that remain equal for all time, and hence have coordinates that are synchronous in this very strong way. Coloring two cells the same when

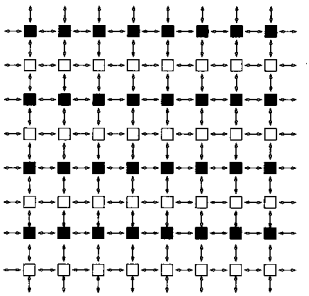


Figure 1.2 Horizontal two-color pattern of synchrony in square lattice with NN coupling.

the coordinates in Δ are required to be equal permits us to visualize the pattern of synchrony associated with a flow-invariant polydiagonal. A k-color pattern of synchrony is one which is defined by exactly k colors.

Symmetry generates many patterns of synchrony. Let $\Sigma \subset \mathcal{G}_{\mathcal{L}}$ be a subgroup and define the *fixed-point subspace*

$$\mathrm{Fix}(\Sigma) = \{x : \sigma x = x \quad \forall\, \sigma \in \Sigma\}.$$

Fixed-point subspaces are well-known to be flow-invariant for any equivariant system. See [8, Lemma XIII 2.1] or [7, Theorem 1.17]. Moreover, fixed-point subspaces of subgroups of $\mathcal{G}_{\mathcal{L}}$ are polydiagonals; so, for lattice differential equations, fixed-point subspaces are patterns of synchrony. As an example, let Σ be generated by two translations $\sigma_1(i,j) = (i+1,j)$ and $\sigma_2(i,j) = (i,j+2)$. Then $\mathrm{Fix}(\Sigma)$ is $2m$-dimensional and consists of points where $x_{i,j} = x_{k,l}$ whenever $j \equiv l \mod 2$. In this case $\mathrm{Fix}(\Sigma)$ is a 2-color pattern of synchrony on the square lattice with NN coupling. See Figure 1.2.

It is not the case that every pattern of synchrony in a coupled cell system is a fixed-point subspace. Stewart *et al.* [10, Theorem 6.1] proved that a polydiagonal is a pattern of synchrony if and only if the coloring associated to the polydiagonal is balanced.

Definition 1.2. A coloring is *balanced* if for for every pair of cells c and d that have the same color there is a bijection from $I(c)$ to $I(d)$ that preserves both color and coupling type.

In the horizontal two-color pattern with the nearest neighbor coupling in Figure 1.2, all couplings are identical and every black cell receives two black inputs and two white inputs and every white cell receives two black inputs and two white inputs. So this pattern is balanced.

In Section 2 we describe a result from [6] that shows that with NN coupling there are continua of balanced two-colorings most of which are

spatially aperiodic. We also describe a result of Wang [11, 12] that classifies all balanced two-colorings up to symmetries in $\mathcal{G}_{\mathcal{L}}$.

There are some subtleties concerning the definition of square lattice LDE with NNN coupling, which we now describe. First we define *square lattice differential equations with the nearest and the next nearest neighbor coupling*, the NNN case, to be those equations that have the form

$$\begin{aligned}\dot{x}_{i,j} = g(x_{i,j}, \overline{x_{i+1,j}, x_{i-1,j}, x_{i,j+1}, x_{i,j-1}}, \\ \overline{x_{i+1,j+1}, x_{i-1,j+1}, x_{i+1,j-1}, x_{i-1,j-1}})\end{aligned} \tag{1.2}$$

where $(i,j) \in \mathbf{Z}^2$, $x_{i,j} \in \mathbf{R}^m$, and g is invariant under independent permutation of the nearest neighbor cells and of the next nearest neighbor cells.

Lattice differential equations can arise from the discretization of systems of partial differential equations. For example, the discretization of a planar reaction-diffusion system leads to

$$\dot{u}_{i,j} = -\beta^+ \Delta^+ u_{i,j} - \beta^\times \Delta^\times u_{i,j} - f(u_{i,j}), \quad (i,j) \in \mathbf{Z}^2 \tag{1.3}$$

where $\beta^+ > \beta^\times > 0$ and Δ^+ and $\Delta^\times$ are the *discrete 2-dimensional Laplace operators* on $\mathbf{Z}^2$ based on the nearest neighbors and the next nearest neighbors, respectively, and are given by

$$\begin{aligned}\Delta^+ u_{i,j} &= u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} \\ \Delta^\times u_{i,j} &= u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1} - 4u_{i,j}\,.\end{aligned}$$

In these equations the parameters β^+ and $\beta^\times$ are coupling parameters, and f is a nonlinear function that represents the "internal dynamics". When $\beta^\times = 0$ this system has only the nearest neighbor coupling, which is a discretization that is frequently used. When the nine-point footprint of the discretized Laplacian is used (β^+ and $\beta^\times$ are both nonzero), then we have an example of square lattice differential equations with the nearest and the next nearest the neighbor coupling, since this system is independently invariant under permutation of the nearest neighbor cells and the next nearest neighbor cells.

Note that (1.3) is equivariant with respect to rotations, reflections, and translations that preserve the lattice. There are, however, examples of $\mathcal{G}_{\mathcal{L}}$-equivariant systems that are not in the form of square lattice differential equations with the nearest and the next nearest neighbor coupling; that is, g is not invariant under independent permutation of the nearest and the next nearest neighbor cells. For example, consider the $\mathbf{D}_4$-invariant function

$$\begin{aligned}g = x_{1,0}(x_{1,1} + x_{1,-1}) + x_{0,1}(x_{-1,1} + x_{1,1}) + \\ x_{-1,0}(x_{-1,-1} + x_{-1,1}) + x_{0,-1}(x_{1,-1} + x_{-1,-1})\end{aligned} \tag{1.4}$$

where g is centered at cell $(0,0)$, i.e. $\dot{x}_{0,0} = g$. So, in fact, there are two reasonable and distinct classes of LDE with the nearest and the next nearest neighbor coupling: NNN equations and $\mathbf{D}_4$-invariant g. Each NNN system is $\mathbf{D}_4$-equivariant, but the converse is false.

In Section 3 we review results in [2] that show that every balanced k-coloring for square lattice NNN equations is spatially doubly periodic and that there are only a finite number of k-colorings for each k. These results show that there are huge differences in patterns of synchrony between the NN and NNN cases. The techniques we developed are general enough to prove similar theorems for other lattices (in particular the other planar lattices); the general principle seems to be that if there is enough coupling, then balanced k-colorings are spatially periodic and finite in number.

Note that every balanced coloring for the $\mathbf{D}_4$-equivariant case is automatically a balanced coloring for the NNN case. So once the finiteness and spatial periodicity theorems about balanced k-colorings are proved for the NNN case, they are automatically valid for the $\mathbf{D}_4$-equivariant case as well. But there may be fewer balanced colorings for the $\mathbf{D}_4$ case than exist for the NNN case, and in fact this happens. See Example 3.11.

2 Square Lattice with NN Coupling

Examples show that with NN coupling there are a continuum of balanced two-colorings, most of which are spatially aperiodic. These examples are easily understood using the trick of interchanging colors along diagonals discovered in [6].

Consider the balanced horizontal two-coloring in Figure 1.2 and the diagonal line drawn in Figure 2.1 (left). If we interchange colors along that diagonal we arrive at the two-coloring in Figure 2.1 (right). Observe that this new coloring is still balanced. For example each black cell still receives two black inputs and two white inputs. The only difference is that the black inputs are now not necessarily along the horizontal. Also note that the new pattern of synchrony is not spatially doubly periodic, even though the original pattern is.

Next we observe that this diagonal trick can be repeated independently on as many diagonals as one wishes. This process leads to patterns of synchrony like those in Figure 2.2. In fact there are a continuum of such patterns of synchrony. To see this, fix any horizontal line H and any binary sequence of colors black and white along H. Now by performing the diagonal trick along each diagonal, if needed, we can arrange for the balanced coloring restricted to H to be the designated binary sequence. Since the number of bi-infinite binary sequences is the same as the real numbers there are a continuum of balanced two-colorings. Of

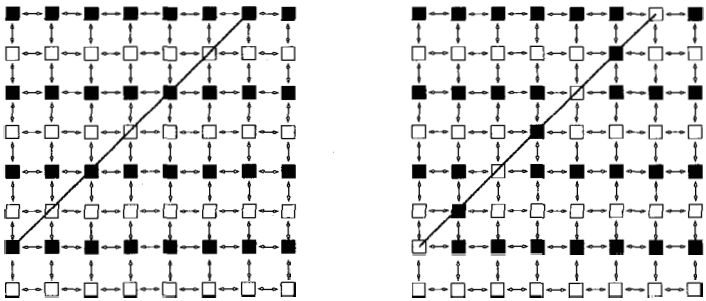


Figure 2.1 Alternating diagonal trick for horizontal two-coloring.

course many of these are identical up to symmetry. But there are only a countable number of symmetries in $\mathcal{G}_{\mathcal{L}}$; so, up to symmetry, there are still a continuum of different balanced two-colorings.

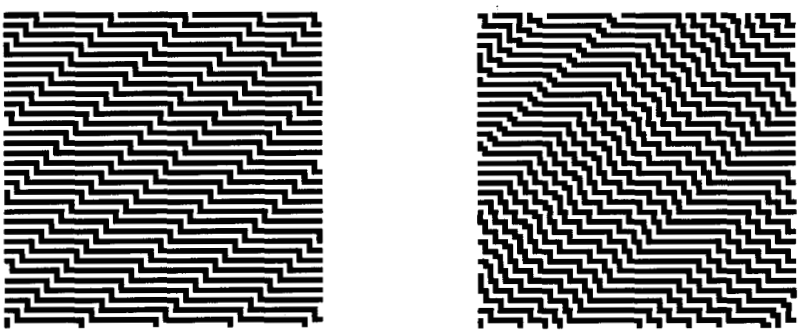


Figure 2.2 Repeating the alternating diagonal trick for horizontal two-coloring.

There is an interesting observation that follows from these results. Suppose that an LDE has an equilibrium inside the horizontal pattern of synchrony in Figure 1.2. Then there is a corresponding equilibrium in every one of the continuum of balanced two-colorings that can obtained from the horizontal one using this diagonal trick. The reason is that LDE's restricted to each of these flow-invariant spaces are identical, and hence each has an equilibrium. (This point can be checked directly from the LDE, but it also follows from the notion of quotient networks discussed in [10, 9].) So symmetry forces a countable number of equilibria with patterns of synchrony that are identical up to symmetry, but network architecture (NN coupling) forces a continuum of equilibria corresponding to patterns of synchrony that are quite different.

Wang [11, 12] classifies, up to symmetry, all balanced two-colorings with NN coupling. There are eight isolated examples, shown in Figure 2.3, and two infinite families, shown in Figure 2.4.

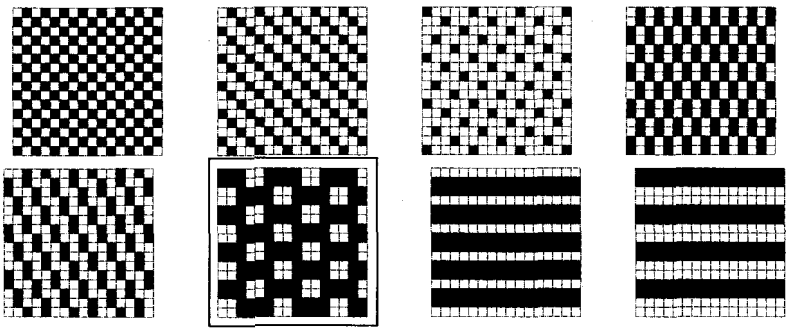


Figure 2.3 Isolated balanced two-colorings on square lattice with NN coupling. The pattern in the box is not a fixed-point subspace of a subgroup of $\mathcal{G}_{\mathcal{L}}$.

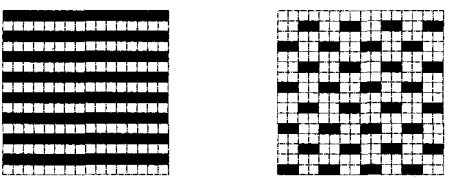


Figure 2.4 Balanced two-colorings on square lattice with NN coupling that generate a continuum of patterns of synchrony by the diagonal trick of interchanging colors on diagonals along which colors alternate.

There is an interesting question: Does the existence of a continuum of balanced two-colorings in the NN case tell us about spurious solutions to discretizations of planar systems of PDE. By themselves, the answer is probably no since two-color solutions are highly oscillatory solutions. If there is to be some such correspondence, we would have to find the continua of balanced k-colorings that would, in some sense, converge to a variety of patterns in a sensible limit. For now we just present a continuum of balanced k-colorings in Figure 2.5.

3 Square Lattice with NNN Coupling

In this section we discuss the following result proved in [2].

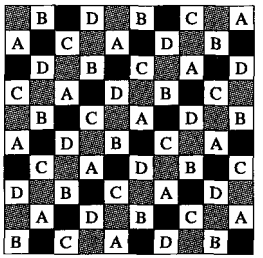


Figure 2.5 A double continua of balanced 6-colorings where diagonals with colors AB and diagonals with colors CD can be interchanged. Similarly, a balanced 5-coloring can be obtained by setting $C = D$. Both examples generalize to examples for any k.

Theorem 3.1. *On the square lattice with both the nearest and the next nearest coupling, every pattern of synchrony is spatially doubly periodic. Moreover, for fixed k, there are, up to symmetry, at most a finite number of balanced k-colorings.*

Definition 3.2. Let $U \subset \mathcal{L}$ be a subset. The *closure* of U consists of all cells that are connected by some arrow to a cell in U. The *boundary* of U is the set

$$\mathrm{bd}(U) = \mathrm{cl}(U) \setminus U.$$

There is a natural expanding sequence of finite subsets that covers the lattice $\mathcal{L}$ and which depends on the kind of coupling. Let

$$W_0 = \{0\} \quad \text{and} \quad W_{i+1} = \mathrm{cl}(W_i) \tag{3.1}$$

for $i \geqslant 0$. Note that for any coloring of a lattice $\mathcal{L}$ by k colors, there is some j such that all k colors are represented by cells in W_j. In fact, more is true for balanced colorings.

Lemma 3.3. *In any balanced k-coloring W_{k-1} contains all k colors.*

Proof. We claim that if $\ell < k$, then W_ℓ contains at least $\ell + 1$ colors. The proof proceeds by induction on W_ℓ. $W_0 = \{0\}$ contains one cell and one color.

Assume that the statement is true for $\ell < k - 1$; we prove that it is also true for $\ell + 1$. Suppose that the number m of colors contained in $W_{\ell+1} = \mathrm{cl}(W_\ell)$ is the same as the number of colors in W_ℓ. Then every cell $c \in W_{\ell+1}$ has a color that is the same as the color of a cell d in W_ℓ. So, all cells connected to d lie in $W_{\ell+1}$ and are colored by the m colors. Therefore, balanced implies that the cells connected to c must also be colored by one of the m colors. It follows that the cells in

$W_{\ell+2} = \mathrm{cl}(W_{\ell+1})$ are also colored by these m colors. By induction the entire lattice is colored by m colors; hence $m = k$. So if $m < k$, the number of colors in $W_{\ell+1}$ must be greater than the number of colors in W_ℓ. That is, $W_{\ell+1}$ contains at least $\ell + 2$ colors. It follows that W_{k-1} contains all k colors. □

It follows from Lemma 3.3 that if we fix a balanced k-coloring to W_k, then there exists a cell e with that color in W_{k-1}. Moreover, all the neighbors of e are in W_k and so all their colors are known. We can ask now about the extension of the pattern from W_k to the whole lattice, more specifically, about the extension of the pattern from W_k to W_{k+1}. Recall that in a balanced k-coloring, for any two cells c and d of the same color, there is a bijection between $I(c)$ and $I(d)$ that preserves arrow type and color. In particular, if we know the color of all cells in $I(c)$ of a certain coupling type and we know the color of all cells of the same coupling type except one in $I(d)$, then since the coloring is balanced we can determine the color of the last cell with that coupling type in $I(d)$.

Definition 3.4. Let $U \subset \mathcal{L}$ be a finite set.

(a) Every cell $c \in U$ is called 0-*determined.*

(b) A cell $c \in \mathrm{bd}(U)$ is *p-determined*, where $p \geq 1$ if there is a cell $d \in U$ such that c is in the input set of d and each cell in the input set of d that has the same coupling type as c, except c itself, is q-determined for some $q < p$.

(c) A cell $c \in \mathrm{bd}(U)$ is *determined* if it is *p-determined* for some p.

(d) The set U *determines* its boundary if all cells in $\mathrm{bd}(U)$ are determined.

Definition 3.5. The set W_{i_0} is a *window* if W_i determines its boundary for all $i \geqslant i_0$.

Remark 3.6. Note that if there are no 1-determined cells then, by induction, there are no p-determined cells for any p. In particular, if there are no 1-determined cells, then windows do not exist.

Example 3.7. Let $\mathcal{L}$ be the square lattice with the nearest neighbor coupling. Then this network has no window.

We claim that no set W_i is a window. By Remark 3.6 it is sufficient to show that there are no 1-determined cells. For example, consider W_2 and its boundary as shown in Figure 3.1. Since the cells on the boundary are in a diagonal line it is not possible for them to be the only cell in the input set of a cell in W_2 that is not in W_2. Note that when $i > 2$ the set W_i has the same "diamond shape" as W_2. So there are no 1-determined cells in $\mathrm{bd}(W_i)$. By Remark 3.6, this network has no window. ◇

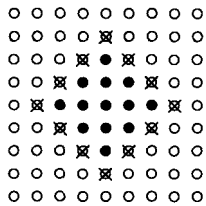


Figure 3.1 The set W_2 (black cells) and its boundary (white cells with a cross).

Example 3.8. Consider $\mathcal{L}$ with the four nearest neighbors and four next nearest neighbors. In this case the set W_i is a square of size $2i+1$. Note that all the cells on each side, except the last two on both extremes, are 1-determined, since they are the only nearest neighbor cells outside the square (Figure 3.2). We show that the sets W_i for $i \geqslant 2$ determine their boundaries. To do this, we need (by symmetry) to analyze just one of the four corners.

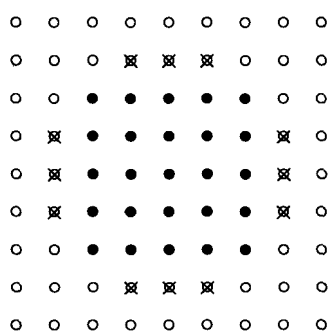


Figure 3.2 The set W_2 (black cells) and the 1-determined cells in its boundary (white cells with a cross).

The three cells in the corners of the square are 2-determined using the next nearest neighbors coupling as long as the square has size greater than 3. See Figure 3.3. ◇

Lemma 3.9. *Assume that W_{i_0} is a window. Suppose that a balanced k-coloring restricted to W_{i-1} for some $i \geqslant i_0$ contains all k colors. Then the k-coloring is uniquely determined on the whole lattice by its restriction to W_i.*

The proof of this lemma [2, Lemma 3.11] uses determinacy.

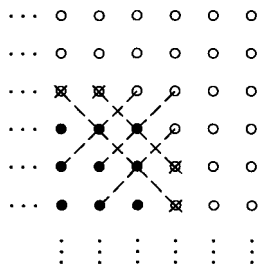


Figure 3.3 The corner of a set W_i (black cells), the 1-determined cells (white cells with a cross) and 2-determined cells (white cells connected to black cells by dashed lines).

Theorem 3.10. *Suppose that a lattice network has a window. Fix $k \geqslant 1$. Then there are a finite number of balanced k-colorings on $\mathcal{L}$ and each balanced k-coloring is spatially multiply-periodic.*

Proof. Let W_j be a window for $G_{\mathcal{L}}$ where $j \geqslant k$. By Lemma 3.3, the interior of W_j contains all k colors. Then by Lemma 3.9, a balanced k-coloring is uniquely determined by its restriction to W_j. Since there is only a finite number of possible ways to distribute k colors on the cells in W_j it follows that there are only a finite number of balanced k-colorings.

Let K be a balanced k-coloring on $G_{\mathcal{L}}$ and let $v \in \mathcal{L}$. Let $T_v(K)$ be the coloring obtained by shifting the coloring K by v, that is, the color of cell c in $T_v(K)$ is the same as the color of cell $c - v$ in K. Since translations are symmetries of the lattice network $T_v(K)$ is also a balanced coloring.

Let v be a generator of the lattice and consider all translates of K in the direction of v. Since there are only a finite number of balanced k-colorings and an infinite number of translates of K, there must exist $N \in \mathbf{Z}^+$, such that K and $T_{Nv}(K)$ exhibit exactly the same coloring. It follows that K is invariant under the translation T_{Nv}. Hence K is periodic in the direction of v. The same argument can be applied to all the generators of the lattice, thus all balanced k-colorings are spatially multiply-periodic. □

Example 3.11. It is straightforward to check that the two-coloring in Figure 3.4 (left) is balanced in the NNN case. Note that up to symmetry there are three different kinds of black cells; see Figure 3.4 (right). Each of these black cells has two white and two black nearest neighbors and three white and one black next nearest neighbors; hence the black cells are NNN balanced. An analogous calculation works for the three types of white cell.

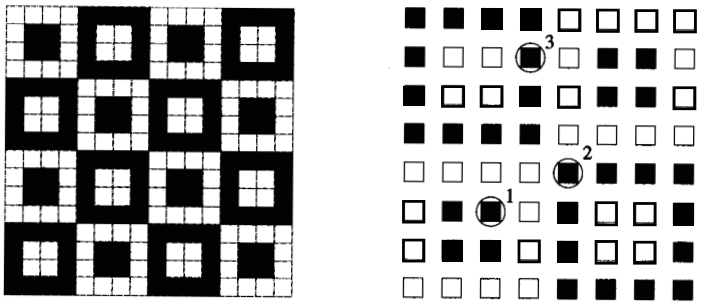


Figure 3.4 A pattern of synchrony of square lattices with NNN coupling that is not flow-invariant for all $\mathbf{D}_4$-equivariant LDE.

We claim that this pattern is not flow-invariant for all $\mathbf{D}_4$-equivariant LDE. Indeed, we can just check this statement using the $\mathbf{D}_4$-invariant function g given in (1.4), which we recall here as

$$\begin{aligned}\dot{x}_{ij} = {} & x_{i+1,j}(x_{i+1,j+1} + x_{i+1,j-1}) + x_{i,j+1}(x_{i+1,j+1} + x_{i-1,j+1}) + \\ & x_{i-1,j}(x_{i-1,j+1} + x_{i-1,j-1}) + x_{i,j-1}(x_{i-1,j-1} + x_{i+1,j-1}).\end{aligned} \tag{3.2}$$

Let the black cells have the coordinate x_B and the white cells have the coordinate x_W. Then the differential equation associated to black cell 1 in Figure 3.4 (right) has the form

$$\begin{aligned}\dot{x}_B = {} & x_W(x_W + x_W) + x_W(x_W + x_W) + \\ & x_B(x_W + x_B) + x_B(x_B + x_W)\end{aligned} \tag{3.3}$$

and the differential equation associated to the black cell 2 has the form

$$\begin{aligned}\dot{x}_B = {} & x_B(x_W + x_W) + x_W(x_W + x_B) + \\ & x_W(x_B + x_W) + x_B(x_W + x_W).\end{aligned} \tag{3.4}$$

However, the right hand side of (3.4) is not equal to the right hand side of (3.3). Hence, the subspace corresponding to Figure 3.4 (left) is not a flow invariant subspace of (3.2). ◇

References

[1] F. Antoneli, A.P.S. Dias, M. Golubitsky, and Y. Wang. Flow invariant subspaces for lattice dynamical systems. In: *Workshop on Bifurcation Theory and Spatio-Temporal Pattern Formation in PDE* (W. Nagata and N.S. Namachchivaya, eds.) Fields Institute Communications. To appear.

[2] F. Antoneli, A.P.S. Dias, M. Golubitsky, and Y. Wang. Patterns of synchrony in lattice dynamical systems. *Nonlinearity* **18** (2005) 2193–2209.

[3] S.-N. Chow. Lattice Dynamical Systems. In: *Dynamical Systems* (S.-N. Chow, eds.) Lecture Notes in Math. **1822** (2003) 1–102.

[4] S.-N. Chow and J. Mallet-Paret. Pattern formation and spatial chaos in lattice dynamical systems: I. *IEEE Trans. Circuits Systems* **42** (1995) 746–751.

[5] S.-N. Chow, J. Mallet-Paret and E.S. Van Vleck. Dynamics of lattice differential equations. *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **6** (9) (1996) 1605–1621.

[6] M. Golubitsky, M. Nicol, and I. Stewart. Some curious phenomena in coupled cell systems. *J. Nonlinear Sci.* **14** (2) (2004) 119–236.

[7] M. Golubitsky and I. Stewart. *The Symmetry Perspective*, Progress in Mathematics **200**, Birkhäuser, Basel 2002.

[8] M. Golubitsky, I.N. Stewart and D.G. Schaeffer. *Singularities and Groups in Bifurcation Theory: Vol. 2.* Applied Mathematical Sciences **69**, Springer-Verlag, New York 1988.

[9] M. Golubitsky, I. Stewart, and A. Török. Patterns of synchrony in coupled cell networks with multiple arrows. *SIAM J. Appl. Dynam. Sys.* **4** (1) (2005) 78–100.

[10] I. Stewart, M. Golubitsky, and M. Pivato. Symmetry groupoids and patterns of synchrony in coupled cell networks. *SIAM J. Appl. Dynam. Sys.* **2** (4) (2003) 609–646.

[11] Y. Wang. *Patterns of Synchrony in Lattice Dynamical Systems.* PhD Thesis, Department of Mathematics, University of Houston, August, 2006.

[12] Y. Wang and M. Golubitsky. Two-color patterns of synchrony in lattice dynamical systems. *Nonlinearity* **18** (2005) 631–657.

The Periodic Unfolding Method for Quasi-convex Functionals*

Alain Damlamian
Laboratoire d'Analyse et de Mathématiques Appliqu
CNRS UMR 8050
niversités de Marne-la-Vallée et Paris 12 Val-de-Marne
94010 Créteil Cedex, France.
E-mail: damla@univ-paris12.fr

1 Homogenization

The theory of *homogenization* arose in the second third of the 20th Century in order to solve numerical problems associated with partial differential equations with highly oscillarory coefficients (composite materials).

When the fine structure is *very fine,* any discretization requires too many points for a computer to handle. The idea of homogenization is to imbed the problem into a family with a small parameter (ε, corresponding to the small scale or scales), look for a limit problem as ε goes to zero, and hope that the limit is

(a) easier to compute numerically, and

(b) a good approximation of the original problem.

This approach has been very successful in many cases. Several methods are commonly used:

— *Asymptotic expansion:* with the Ansatz

$$u^\varepsilon(x) = u_0\left(x, \frac{x}{\varepsilon}\right) + \varepsilon u_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon^2 u_2\left(x, \frac{x}{\varepsilon}\right) + \cdots$$

with the formal substitution of

$$\frac{\partial}{\partial x_j} \text{ by } \frac{\partial}{\partial x_j} + \frac{1}{\varepsilon}\frac{\partial}{\partial y_j}.$$

— *Oscillating test functions* (Luc Tartar's method using compensated compactness.)

This method applies beyond the periodic case also

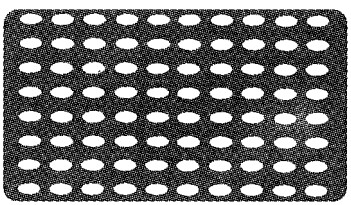
*This is a report on joint with Doina Cioranescu and Riccardo de Arcangelis.

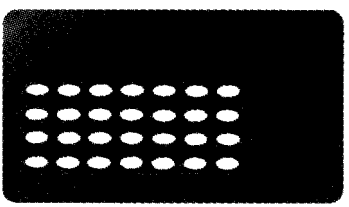
— *Two-scale convergence* (due originally to G.Nguetseng)

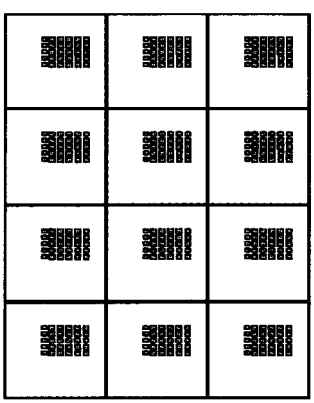
In the same spirit, the theory of Γ-convergence was introduced to deal specifically with real-valued functionals. It is very closely related to Homogenization when the latter considers partial differential equations which are the Euler equations for the minimization of an energy functional.

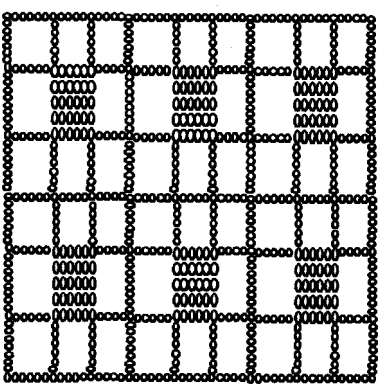
More recently (2002-2003), a new approach was devised by D.Cioranescu, G. Griso and A.D. for the periodic case (it also applies to the multi-periodic case).

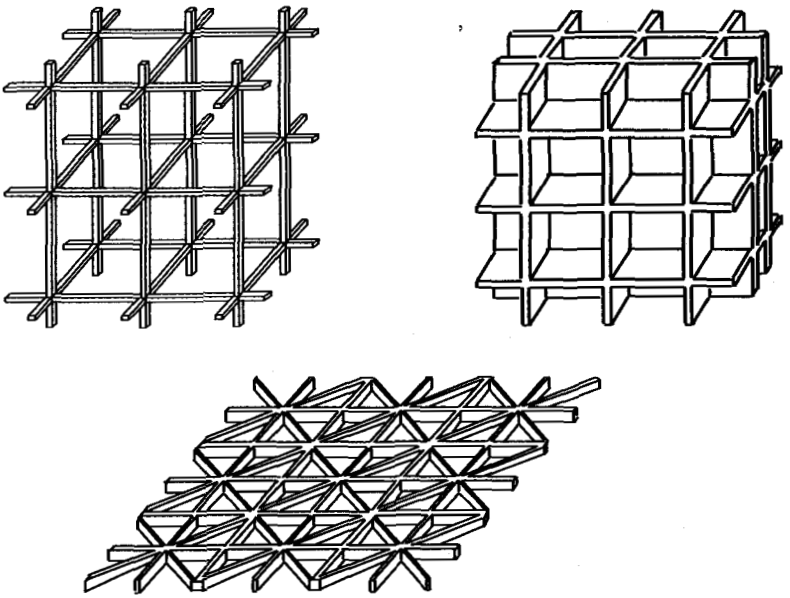
Some examples of periodic or multi-periodic materials and structures











2 The periodic unfolding method[CDG]

The *periodic unfolding method* is a "fixed domain" method which is well-suited to treat periodic homogenization problems of many types.

The basic idea is that if the proper scaling is used, oscillatory behaviour can be turned into weak or even strong convergence, at the price of an increase in the dimensionality of the problem, but with significant simplifications in the proofs and explicit formulas. It also gives a most elementary proof for the results of the theory of two-scale convergence due to G. Nguetseng.

The *periodic unfolding method* is being successfully applied to many cases:

Originally, it was applied for linear problems in the standard periodic case of homogenization, then to the case with holes.

It also applies equally to

– Elasticity and composite materials,

– Truss-like structures, as well as rods, plates and composites thereof,

– Multiscale periodic homogenization.

It gives an elegant alternate proof for the case of "small holes", where so called "strange terms" appear, and can treat the mixed case of periodic homogenization and small holes as efficiently.

In each case, the method is elementary in spirit, and provides proofs

which are both simple and rigorous and which look like a mix of formal asymptotic expansions and two-scale convergence.

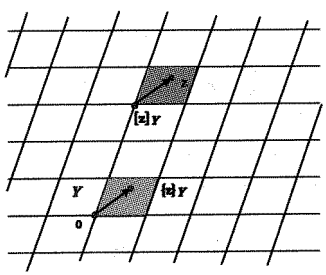
More recently, the first proof of the speed of convergence of the solutions of the ε-problems to the solution of the limit problem (in the model case) was obtained using the *periodic unfolding method* (G.Griso).

We begin with an overview of the unfolding method in its simplest form, then, focus on the proof of the Γ-convergence of non-linear functionals in the periodic case.

3 The unfolding operator $\mathcal{J}_\varepsilon$

$\Omega \subset \mathbf{R}^n$ open and Y a reference cell (ex. $[0,1[^n$, or more generally a set having the paving property with respect to the basis $(b_1, \cdots, b_n)$ defining the periods).

By analogy with the 1D notations, for $z \in \mathbf{R}^n$, $[z]_Y$ denotes the unique integer combination $\sum_{j=1}^n k_j b_j$ of the periods such that $z - [z]_Y$ belongs to Y, and $\{z\}_Y$ denotes the unique element $z - [z]_Y$ of Y.



For $w \in L^p(\Omega)$, $p \in [1, \infty]$ extended by zero outside of Ω define $\mathcal{J}_\varepsilon(w) \in L^p(\Omega \times Y)$ by

$$\mathcal{J}_\varepsilon(w)(x,y) = w\left(\varepsilon\left[\frac{x}{\varepsilon}\right]_Y + \varepsilon y\right) \quad \text{for } x \in \Omega \text{ and } y \in Y.$$

$\Longrightarrow \forall x \in \Omega \qquad \mathcal{J}_\varepsilon(w)\left(x, \left\{\frac{x}{\varepsilon}\right\}_Y\right) = w(x).$

$\Longrightarrow$ Integration formula

$$\int_\Omega w\, dx = \frac{1}{|Y|}\int_{\Omega_\varepsilon \times Y} \mathcal{J}_\varepsilon(w) dx dy, \qquad \forall w \in L^1(\Omega)$$

where

$$\Omega_\varepsilon = \bigcup \{\varepsilon\xi + \varepsilon Y\}; \forall \xi \in \mathbf{Z}^{\mathbf{N}}, \{\varepsilon\xi + \varepsilon Y\} \cap \Omega \neq \varnothing$$

Properties of $\mathcal{J}_\varepsilon$ $\quad 1 \le p < \infty$

- $\|\mathcal{J}_\varepsilon(w)\|_{L^p(\Omega\times Y)} \le |\,Y\,|^{\frac{1}{p}}\,\|w\|_{L^p(\Omega)}, \quad \forall w \in L^p(\Omega)$
- $\mathcal{J}_\varepsilon(vw) = \mathcal{J}_\varepsilon(v)\,\mathcal{J}_\varepsilon(w), \quad \forall v, w \in L^2(\Omega)$
- $\forall w \in L^p(\Omega) \Longrightarrow \mathcal{J}_\varepsilon(w) \to w$ in $L^p(\Omega \times Y)$
- $\forall w_\varepsilon \to w$in $L^p(\Omega) \Longrightarrow \mathcal{J}_\varepsilon(w_\varepsilon) \to w$ in $L^p(\Omega \times Y)$
- $\{w_\varepsilon\} \in L^p(\Omega)$. If $\mathcal{J}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ in $L^p(\Omega \times Y) \Longrightarrow$

  $w_\varepsilon \rightharpoonup w$ in $L^p(\Omega)$ where $w = \dfrac{1}{|\,Y\,|}\displaystyle\int_Y \widehat{w}\,dy$
- If $\mathcal{J}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ in $L^p(\Omega \times Y)$ and

  $\limsup |Y|^{\frac{1}{p}}\|w_\varepsilon\|_{L^p(\Omega)} \le \|\hat{w}\|_{L^p(\Omega\times Y)} \Longrightarrow \mathcal{J}_\varepsilon(w_\varepsilon) \to \hat{w}$

Two-scale Convergence (cf. Nguetseng, Allaire)

Definition *Let $\{w_\varepsilon\} \subset L^p(\Omega)$, be a bounded sequence. It is said that $\{w_\varepsilon\}_\varepsilon$ two-scale converges in the $L^p(\Omega \times Y)$-sense to some w in $L^p(\Omega \times Y)$, whenever, for every smooth function φ on $\Omega \times Y$, the following convergence holds:*

$$\int_\Omega w_\varepsilon(x)\varphi\left(x, \frac{x}{\varepsilon}\right)\,dx \to \frac{1}{|Y|}\int\int_{\Omega\times Y} w(x,y)\,\varphi(x,y)\,dxdy.$$

Theorem *From a bounded sequence $\{w_\varepsilon\} \subset L^p(\Omega)$, one can always extract a subsequence which two-scale converges in the $L^p(\Omega \times Y)$-sense.*

Relation with Two-scale Convergence

Proposition *Let $\{w_\varepsilon\} \subset L^p(\Omega)$, be a bounded sequence. The following are equivalent:*

- *$\{\mathcal{J}_\varepsilon(w_\varepsilon)\}_\varepsilon$ converges weakly to w in $L^p(\Omega \times Y)$;*
- *$\{w_\varepsilon\}_\varepsilon$ two-scale converges to w (in the $L^p(\Omega \times Y)$ sense).*

Proof: It is enough to check that for every element φ in a set of admissible test-functions for two-scale convergence (in this case, $\mathcal{D}(\Omega, L^q(Y))$ is enough), $\mathcal{J}_\varepsilon[\varphi(x, \frac{x}{\varepsilon})]$ converges strongly to φ in $L^q(\Omega \times Y)$).

This follows from the fact that

$$\mathcal{J}_\varepsilon\left[\varphi\left(x, \frac{x}{\varepsilon}\right)\right](x,y) = \varphi\left(\varepsilon\left[\frac{x}{\varepsilon}\right]_Y + \varepsilon y, y\right). \qquad \square$$

Properties of $\mathcal{J}_\varepsilon$ with respect to gradients:

- $\forall w \in W^{1,p}(\Omega)$,

$$\nabla_y(\mathcal{J}_\varepsilon(w)) = \varepsilon \mathcal{J}_\varepsilon(\nabla_x w).$$

Lemma *Let $\{w_\varepsilon\}$ be a sequence in $W^{1,p}(\Omega)$, which is bounded in $L^p(\Omega)$ and such that $\mathcal{J}_\varepsilon(w_\varepsilon) \rightharpoonup \widehat{w}$ in $L^p(\Omega \times Y)$. Suppose furthermore that*

$$\varepsilon \|\nabla w_\varepsilon\|_{L^p(\Omega)} \leq C.$$

$\Longrightarrow$ *Then the following convergence holds in $L^p(\Omega; W^{1,p}(Y))$:*

$$\nabla_y(\mathcal{J}_\varepsilon(w)) = \varepsilon \mathcal{J}_\varepsilon(\nabla_x w_\varepsilon) \rightharpoonup \nabla_y \hat{w} \quad \text{in } L^p(\Omega \times Y).$$

Furthermore

$$\hat{w} \in L^p(\Omega; W^{1,p}_{\text{per}}(Y)).$$

4 Macro–Micro decomposition of functions

The scale-splitting operators $\mathcal{Q}_\varepsilon$ and $\mathcal{R}_\varepsilon$, every $\varphi \in H^1(\Omega)$ can be split as

$$\varphi = \mathcal{Q}_\varepsilon(\varphi) + \mathcal{R}_e(\varphi).$$

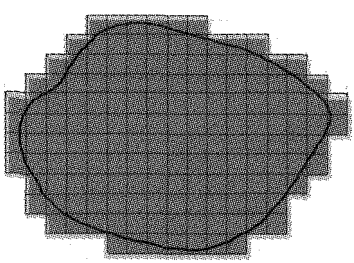
For $k \in \mathbf{Z}^n$, set $\xi_k = \sum_{j=1}^n k_j b_j$ and

$$\widetilde{\Omega^\varepsilon} = \bigcup \left\{ \varepsilon \xi_k + \varepsilon \overline{Y} \right\},$$
$$\varepsilon \xi_k + \varepsilon \overline{Y} \cap \overline{\Omega} \neq \varnothing.$$

$\partial\Omega$ is assumed to be bounded and Lipschitz. It is then well-known that there exists a continuous extension operator $\mathcal{P} : H^1(\Omega) \longmapsto H^1(\mathbf{R}^n)$ such that $\forall \varphi \in H^1(\Omega)$

$$\|\mathcal{P}(\varphi)\|_{L^2(\mathbf{R}^n)} + \|\nabla \mathcal{P}(\varphi)\|_{L^2(\mathbf{R}^n)} \leq C \, \|\varphi\|_{H^1(\Omega)},$$

(where $C = C(\partial\Omega)$).



The set $\widetilde{\Omega^\varepsilon}$

Definition of $\mathcal{Q}_\varepsilon$

$\mathcal{Q}_\varepsilon$ is a Q_1-interpolation operator of order ε in the sense of numerical Analysis.

For a given function φ on Ω, one starts by defining $\mathcal{Q}_\varepsilon(\varphi)$ at the nodes $\varepsilon\xi_k$ included in $\widetilde{\Omega^\varepsilon}$, using any reasonable averaging method. For example (this is enough in our case here),

$$\mathcal{Q}_\varepsilon(\varphi)(\varepsilon\xi_k) = \frac{1}{|Y|}\int_Y \mathcal{P}(\varphi)(\varepsilon\xi_k + \varepsilon y)\, dy.$$

Then $\mathcal{Q}_\varepsilon(\varphi)$ is the restriction to Ω of the Q_1- interpolate of the discrete function $\mathcal{Q}_\varepsilon(\varphi)(\varepsilon\xi_k)$.

It is well known, from Numerical Analysis, that
$\forall\phi \in H^1(\Omega)$:

- $\|\mathcal{Q}_\varepsilon(\varphi)\|_{H^1(\Omega)} \leq C\|\varphi\|_{H^1(\Omega)}$,
- $\|\phi - \mathcal{Q}_\varepsilon(\varphi)\|_{L^2(\Omega)} \leq \varepsilon\, C\|\nabla\varphi\|_{L^2(\Omega)}$, (Poincaré-Wirtinger).

Definition of $\mathcal{R}_\varepsilon$

It is clear that $\mathcal{Q}_\varepsilon$ is designed NOT to capture any oscillations of order ε.

What must capture these oscillations is therefore

$$\mathcal{R}_\varepsilon(\varphi) = \phi - \mathcal{Q}_\varepsilon(\varphi).$$

The two inequalities above for $\mathcal{Q}_\varepsilon$ translate as:

- $\|\mathcal{R}_\varepsilon(\varphi)\|_{L^2(\Omega)} \leq \varepsilon C\|\varphi\|_{H^1(\Omega)}$,
- $\|\mathcal{R}_\varepsilon(\varphi)\|_{H^1(\Omega)} \leq C\|\varphi\|_{H^1(\Omega)}$.

Similarly, one has: $\forall\phi \in W^{1,p}(\Omega)$,

- $\|\mathcal{Q}_\varepsilon(\varphi)\|_{W^{1,p}(\Omega)} \leq C\|\varphi\|_{W^{1,p}(\Omega)}$,
- $\|\phi - \mathcal{Q}_\varepsilon(\varphi)\|_{L^p(\Omega)} \leq \varepsilon\, C\|\nabla\varphi\|_{L^p(\Omega)}$, ($p$-Poincaré-Wirtinger)
- $\|\mathcal{R}_\varepsilon(\varphi)\|_{L^p(\Omega)} \leq \varepsilon C\|\varphi\|_{W^{1,p}(\Omega)}$,
- $\|\mathcal{R}_\varepsilon(\varphi)\|_{W^{1,p}(\Omega)} \leq C\|\varphi\|_{W^{1,p}(\Omega)}$.

Convergence for sequences in $W^{1,p}(\Omega), p \in [1,\infty)$

Theorem *Let $\{w_\varepsilon\}$ be a weakly convergent sequence in $W^{1,p}(\Omega)$*

$$w_\varepsilon \rightharpoonup w \quad \text{in } W^{1,p}(\Omega).$$

Then there exists a subsequence (still denoted ε) and a $\widehat{w} \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))$ such that:

- $\mathcal{Q}_\varepsilon(w_\varepsilon) \rightharpoonup w$ *in* $W^{1,p}(\Omega)$,
- $\mathcal{J}_\varepsilon(w_\varepsilon) \rightharpoonup w$ *in* $L^p(\Omega; W^{1,p}(Y))$,
- $\frac{1}{\varepsilon}\mathcal{J}_\varepsilon(\mathcal{R}_\varepsilon(w_\varepsilon)) \rightharpoonup \widehat{w}$ *in* $L^p(\Omega; W^{1,p}(Y))$,
- $\mathcal{J}_\varepsilon(\nabla w_\varepsilon) \rightharpoonup \nabla_x w + \nabla_y \widehat{w}$ *in* $L^p(\Omega \times Y)$.

5 Homogenization of nonlinear convex energies

The homogenization in the general case of nonlinear convex integral energies is well established, and goes back to [Ma] (1978) and [CSb] (1979).

The use of periodic unfolding method simplifies the treatment of the homogenization process, and allows the deduction of different types of limit formulas, by reducing it again to a weak convergence problem in an appropriate L^p space, cf. [CDDA].

$\mathcal{A}_0$ is the class of the bounded open subsets of $\mathbf{R}^n$ having Lipschitz boundary.

f Carathéodory energy density

$$(H_0) \quad \begin{cases} f : (x,z) \in \mathbf{R}^n \times \mathbf{R}^n \mapsto f(x,z) \in [0,+\infty[\\ f(\cdot,z) \text{ Lebesgue measurable, and } Y\text{-periodic for every } z \in \mathbf{R}^n \\ f(x,\cdot) \text{ convex for a.e. } x \in \mathbf{R}^n. \end{cases}$$

For $p \in [1,+\infty[$, and $M > 0$

$$(H_1) \quad f(x,z) \leq M(1+|z|^p) \text{ for a.e. } x \in \mathbf{R}^n, \text{ and every } z \in \mathbf{R}^n,$$

$$(H_2) \quad |z|^p \leq f(x,z) \text{ for a.e. } x \in \mathbf{R}^n, \text{ and every } z \in \mathbf{R}^n.$$

$W^{1,p}_{\text{per}}(Y)$ is the Banach space of Y-periodic functions in $W^{1,p}_{\text{loc}}(\mathbf{R}^n)$ endowed with the $W^{1,p}(Y)$-norm.

$$f_{\text{hom}} : z \in \mathbf{R}^n \mapsto \inf\left\{\int_Y f(y, z + \nabla v(y))dy : v \in W^{1,p}_{\text{per}}(Y)\right\}.$$

Theorem *Let f satisfy (H_0), and assume that (H_1) holds for some $p \in [1,+\infty[$. Let $\Omega \in \mathcal{A}_0$, and let $\{\varepsilon_h\} \subset]0,+\infty[$ converge to 0. Then,*

for every $u \in W^{1,p}(\Omega)$,

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h},\nabla u_h(x)\right)dx : \{u_h\}\subset W^{1,p}(\Omega),\right.\\
&\qquad\left. u_h \rightharpoonup u \text{ in } W^{1,p}(\Omega)\right\}\\
&=\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h},\nabla u_h(x)\right)dx : \{u_h\}\subset W^{1,p}(\Omega),\right.\\
&\qquad\left. u_h \rightharpoonup u \text{ in } W^{1,p}(\Omega)\right\}\\
&=\inf\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W^{1,p}_{\text{per}}(Y))\right\}.
\end{aligned}$$

If in addition, (H_2) *holds with* $p \in]1,+\infty[$, *then this common value equals*

$$\int_\Omega f_{\text{hom}}(\nabla u(x))dx.$$

The proof is contained in the following three lemmas.

Lemma 1 *Assume that* f *satisfies* (H_0). *Let* $\Omega \in \mathcal{A}_0$, $p \in [1,+\infty[$, *and* $\{\varepsilon_h\} \subset]0,+\infty[$ *converge to 0 as* $h \to +\infty$. *Then, for every* $u \in W^{1,p}(\Omega)$,

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h},\nabla u_h(x)\right)dx : \{u_h\}\subset W^{1,p}(\Omega),\right.\\
&\qquad\left. u_h \rightharpoonup u \text{ in } W^{1,p}(\Omega)\right\}\\
&\geq\inf\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W^{1,p}_{\text{per}}(Y))\right\}
\end{aligned}$$

Lemma 2 *Assume that* f *satisfies* (H_0), *and that* (H_1) *holds for some* $p \in [1,+\infty[$. *Let* $\Omega \in \mathcal{A}_0$, *and let* $\{\varepsilon_h\} \subset]0,+\infty[$ *converge to 0 as* $h \to +\infty$. *Then, for every* $u \in W^{1,p}(\Omega)$,

$$\begin{aligned}
&\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h},\nabla u_h(x)\right)dx : \{u_h\}\subset W^{1,p}(\Omega),\right.\\
&\qquad\left. u_h \rightharpoonup u \text{ in } W^{1,p}(\Omega)\right\}\\
&\leq\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W^{1,p}_{\text{per}}(Y))\right\}.
\end{aligned}$$

Lemma 3 *Assume that f satisfies (H_0). Let $\Omega \in \mathcal{A}_0$, $p \in]1,+\infty[$, and suppose that (H_1) and (H_2) hold. Then, for every $u \in W^{1,p}(\Omega)$,*

$$\inf\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))\right\}$$

$$= \int_\Omega f_{\text{hom}}(\nabla u(x))dx.$$

The proof of each of these lemmas is elementary!!

Proof (of Lemma 1, easy): $u \in W^{1,p}(\Omega)$, $u_h \rightharpoonup u$,

$$\lim_{h\to+\infty}\int_\Omega f(\frac{x}{\varepsilon_h},\nabla u_h(x))dx$$

exists and is finite. Then there exists $\{\varepsilon_{h_k}\} \subset \{\varepsilon_h\}$ and $U \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))$ with

$$\mathcal{T}_{\varepsilon_{h_k}}(\nabla u_{h_k}) \rightharpoonup \nabla u + \nabla_y U \text{ in } (L^p(\Omega\times Y))^n.$$

$$\int_\Omega f\left(\frac{x}{\varepsilon_{h_k}},\nabla u_{h_k}(x)\right)dx$$

$$= \int_{\Omega_{\varepsilon_{h_k}}\times Y} \mathcal{T}_{\varepsilon_{h_k}}\left(f\left(\frac{\cdot}{\varepsilon_{h_k}},\nabla u_{h_k}(\cdot)\right)\right)(x,y)dxdy$$

$$\geq \int_{\Omega\times Y} f\left(\frac{1}{\varepsilon_{h_k}}\left(\varepsilon_{h_k}\left[\frac{x}{\varepsilon_{h_k}}\right]+\varepsilon_{h_k}y\right), \mathcal{T}_{\varepsilon_{h_k}}(\nabla u_{h_k})(x,y)\right)dxdy$$

$$= \int_{\Omega\times Y} f\left(\left[\frac{x}{\varepsilon_{h_k}}\right]+y, \mathcal{T}_{\varepsilon_{h_k}}(\nabla u_{h_k})(x,y)\right)dxdy$$

$$= \int_{\Omega\times Y} f(y, \mathcal{T}_{\varepsilon_{h_k}}(\nabla u_{h_k})(x,y))dxdy,$$

where the last functional is sequentially weakly $(L^p(\Omega))^n$-lower semicontinuous (well-known consequence of Fatou's Lemma under hypothesis (H_0)).

$$\liminf_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h},\nabla u_h(x)\right)dx \geq \int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y U(x,y))dxdy$$

$$\geq \inf\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))\right\}. \ \square$$

Proof (of Lemma 2): $u \in W^{1,p}(\Omega)$, $U \in C^1(\mathbf{R}^n \times \mathbf{R}^n)$ with $U(x,\cdot)$ Y-periodic for every $x \in \Omega$. For every $h \in \mathbf{N}$ and $x \in \Omega$, we set

$u_h(x) = u + \varepsilon_h(x, \frac{x}{\varepsilon_h})$.Clearly,

$$\nabla u_h(x) = \nabla u(x) + \varepsilon_h \nabla_x U\Big(x, \frac{x}{\varepsilon_h}\Big) + \nabla_y U\Big(x, \frac{x}{\varepsilon_h}\Big)$$
$$\text{for every } h \in \mathbf{N}, x \in \Omega.$$

Then, from the periodicity of U, one has

$$\begin{aligned}
&\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h(x)\Big)dx \\
&= \int_{\Omega_{\varepsilon_h}\times Y} f(y, \mathcal{T}_{\varepsilon_h}(\nabla u_h)(x,y))dxdy \\
&= \int_{\Omega_{\varepsilon_h}\times Y} f\Big(y, \mathcal{T}_{\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big) \\
&\qquad + \nabla_y U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big)\Big)dxdy.
\end{aligned}$$

Due to continuity properties of $\nabla_x U$ and to the periodicity of $\nabla_y U$,

$$\begin{aligned}
&\varepsilon_h \nabla_x U\Big(\cdot, \frac{\cdot}{\varepsilon_h}\Big) \to 0 \text{ uniformly in } \Omega, \\
&\nabla_y U\Big(\cdot, \frac{\cdot}{\varepsilon_h}\Big) \to \int_Y \nabla_y U(\cdot, y)dy = 0 \text{ weakly* in } (L^\infty(\Omega))^n.
\end{aligned}$$

This implies that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega)$. Moreover,

$$\begin{aligned}
&\inf\Big\{\limsup_{h\to+\infty} \int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla v_h(x)\Big)dx : \{v_h\} \subset W^{1,p}(\Omega), \\
&\qquad v_h \rightharpoonup u \text{ in } W^{1,p}(\Omega)\Big\} \\
&\le \limsup_{h\to+\infty} \int_{\Omega_{\varepsilon_h}\times Y} f\Big(y, \mathcal{T}_{\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big) \\
&\qquad + \nabla_y U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big)\Big)dxdy.
\end{aligned}$$

On the other hand, again by the continuity properties of $\nabla_x U$ and of $\nabla_y U$,

$$\begin{aligned}
&\varepsilon_h \nabla_x U\left(\varepsilon_h\left[\frac{\cdot}{\varepsilon_h}\right] + \varepsilon_h \cdot, \cdot\right) \to 0 \text{ and} \\
&\nabla_y U\left(\varepsilon_h\left[\frac{\cdot}{\varepsilon_h}\right] + \varepsilon_h \cdot, \cdot\right) \to \nabla_y U \text{ uniformly in } \Omega.
\end{aligned}$$

One also has $\mathcal{T}_{\varepsilon_h}(\nabla u) \to \nabla u$ in $(L^p(\mathbf{R}^n \times \mathbf{Y}))^n$. So we can pass to the limit and using (H_1), we get

$$\lim_{h\to+\infty} \int_{\Omega_{\varepsilon_h}\times Y} f\Big(y, \mathcal{T}_{\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big)$$
$$+\nabla_y U\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big] + \varepsilon_h y, y\Big)\Big) dxdy$$
$$= \int_{\Omega\times Y} f(y, \nabla u(x) + \nabla_y U(x,y))dxdy.$$

We conclude the proof by a standard density argument when $U \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))$ (by observing that (H_0) and(H_1) imply the continuity on $L^p(\Omega; W^{1,p}_{\text{per}}(Y))$ of the right hand side in the above inequality). □

Proof (of Lemma 3): After the two steps what we know is that the "limit" of the functional $\int_\Omega f(\frac{x}{\varepsilon}, \nabla u_\varepsilon)\, dx$ is

$$\inf\Big\{\int_{\Omega\times Y} f(y, \nabla u(x) + \nabla_y V(x,y))dxdy : V \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))\Big\}.$$

We want to prove that this value is nothing else than $\int_\Omega f^p_{\text{hom}}(\nabla u(x))dx$. From the definition of f^p_{hom}, we have

$$\int_\Omega f^p_{\text{hom}}(\nabla u(x))dx$$
$$= \inf\Big\{\int_Y f(y, \nabla u(x) + \nabla v(y))dy : v \in W^{1,p}_{\text{per}}(Y)\Big\}.$$

One inequality is straightforward, since for u in $W^{1,p}(\Omega)$ and V in $L^p(\Omega; W^{1,p}_{\text{per}}(Y))$, the following inequality holds for a.e. $x \in \Omega$,

$$\int_Y f(y, \nabla u(x) + \nabla_y V(x,y))dy \geq f^p_{\text{hom}}(\nabla u(x)).$$

The reverse inequality is obvious if $\int_\Omega f^p_{\text{hom}}(\nabla u(x))dx = +\infty$. To prove it in the case where $\int_\Omega f^p_{\text{hom}}(\nabla u(x))dx < +\infty$, we make use of Castaing's selection theorem.

Castaing theorem on measurable selections

Let $\mathcal{O}$, X be sets, and $\mathcal{G}$ a multifunction from $\mathcal{O}$ to X. A function $\sigma : \Omega \to X$ will be said to be a selection of $\mathcal{G}$ if $\sigma(x) \in \mathcal{G}(x)$ for every $x \in \mathcal{O}$. The measurable selection result below is known as Castaing's theorem.

Theorem Let X be a separable metric space, $(\mathcal{O}, \mathcal{M})$ a measurable space, and $\mathcal{G}$ a multifunction from $\mathcal{O}$ to X. Assume that for every

$x \in \mathcal{O}$, $\mathcal{G}(x)$ is nonempty and complete in X. Assume moreover, that for every closed subset F of X, $\{x \in \mathcal{O} : \mathcal{G}(x) \cap F \neq \phi\}$ belongs to $\mathcal{M}$. Then $\mathcal{G}$ admits a $\mathcal{M}$-measurable selection.

Note first, that due to (H_0) and (H_1), f^p_{hom} is convex and continuous on $\mathbf{R}^n$. Due to (H_2) and the Poincaré-Wirtinger inequality, the infimum defining $f^p_{\text{hom}}(z)$, is achieved for every $z \in \mathbf{R}^n$.

This, and (H_1) imply that for $z \in \mathbf{R}^n$, $\Gamma(z)$ is nonempty, and strongly closed, where Γ is the multifunction defined by

$$\Gamma(z) \doteq \left\{ v \in W^{1,p}_{\text{per}}(Y) : \int_Y v(y)dy = 0, \right.$$
$$\left. \int_Y f(y, z + \nabla v(y))dy = f^p_{\text{hom}}(z) \right\}.$$

Now, by Castaing Theorem, Γ has a $\mathcal{B}(\mathbf{R}^n)$-measurable selection, where $\mathcal{B}(\mathbf{R}^n)$ denotes the Borel σ-algebra of $\mathbf{R}^n$.

Let σ denotes such a measurable selection. Fix $u \in W^{1,p}(\Omega)$. For a.e. $x \in \Omega$, set $U(x) = \sigma(\nabla u(x))$. Then U is $\mathcal{L}(\Omega)$-measurable, with values in $W^{1,p}_{\text{per}}(Y)$ so that

$$f^p_{\text{hom}}(\nabla u(x)) = \int_Y f(y, \nabla u(x) + \nabla_y U(x)(y))dy \text{ for a.e. } x \in \Omega.$$

Integrating over Ω yields

$$\int_{\Omega \times Y} f(y, \nabla u(x) + \nabla_y U(x)(y))\, dydx < +\infty,$$

so that by (H_2), $\nabla_y U$ is in $(L^p(\Omega \times Y))^n$. By the Poincaré-Wirtinger inequality, U belongs to $L^p(\Omega; W^{1,p}_{\text{per}}(Y))$, hence the claim

$$\int_\Omega f_{\text{hom}}(\nabla u(x))dx = \int_{\Omega \times Y} f(y, \nabla u(x) + \nabla_y U(x)(y))dxdy$$
$$\geq \inf \left\{ \int_{\Omega \times Y} f(y, \nabla u(x) + \nabla_y V(x,y))dxdy : V \in L^p(\Omega; W^{1,p}_{\text{per}}(Y)) \right\}.$$

The ingredients of the proof are therefore:

Lemma 1 Periodic unfolding, lower semicontinuity (convexity).

Lemma 2 Periodic unfolding, growth conditions.

Lemma 3 Castaing selection theorem, lower semicontinuity (convexity), coerciveness.

6 Homogenization of quasiconvex energies

This concerns the homogenization problem in the general case of quasiconvex integral energies, under p-growth and coerciveness assumptions, defined on vector-valued configurations.

For such energies the homogenization result is well established, and goes back, among others, to [B] (1985) and [Mü] (1987).

We show how the use of periodic unfolding simplifies the treatment by reducing it again to a weak convergence problem, see [CDDA2].

The hypotheses are: m, $n \in \mathbf{N}$. f Carathéodory energy density

$$(H_0)\quad \begin{cases} f : (x,z) \in \mathbf{R}^n \times \mathbf{R}^{nm} \mapsto f(x,z) \in [0,+\infty[, \\ f(\cdot,z) \text{ Lebesgue measurable and } Y\text{-periodic for every } z \in \mathbf{R}^{nm}, \\ f(x,\cdot) \text{ continuous for a.e. } x \in \mathbf{R}^n. \end{cases}$$

$$(QC) \qquad f(x,\cdot) \text{ is quasiconvex for a.e. } x \in \mathbf{R}^n.$$

There are $p \in [1,+\infty[$, $M > 0$, $a \in L^1(Y)$ Y-periodic such that

$$(H_1) \qquad f(x,z) \le a(x) + M|z|^p \quad \text{for a.e. } x \in \mathbf{R}^n, \text{ and every } z \in \mathbf{R}^{nm},$$

$$(H_2) \qquad |z|^p \le f(x,z) \quad \text{for a.e. } x \in \mathbf{R}^n, \text{ and every } z \in \mathbf{R}^{nm}.$$

$$f_{\text{hom}} : z \in \mathbf{R}^{nm} \mapsto \lim_{t\to+\infty} \frac{1}{t^n} \inf\left\{ \int_{tY} f(y, z+\nabla v)dy : v \in W_0^{1,p}(tY;\mathbf{R}^m) \right\}.$$

The main result here is:

Theorem Let f satisfy (H_0) and (QC). Let $p \in]1,+\infty[$, and assume that (H_1) and (H_2) hold. Then, for every $\{\varepsilon_h\} \subset]0,+\infty[$ converging to 0, $\Omega \in \mathcal{A}_0$, and u in $W^{1,p}(\Omega;\mathbf{R}^m)$,

$$\begin{aligned} &\inf\left\{ \liminf_{h\to+\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right)dx : \{u_h\}W^{1,p}(\Omega;\mathbf{R}^m), \right. \\ &\qquad \left. u_h \to \text{u } in\ L^p(\Omega;\mathbf{R}^m) \right\} \\ &= \inf\left\{ \limsup_{h\to+\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right)dx : \{u_h\} \subset W^{1,p}(\Omega;\mathbf{R}^m), \right. \\ &\qquad \left. u_h \to u\ in\ L^p(\Omega;\mathbf{R}^m) \right\} \end{aligned}$$

$$= \inf\left\{ \liminf_{h\to+\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right) dx : \{u_h\} \subset u + W_0^{1,p}(\Omega; \mathbf{R}^m), \right.$$
$$\left. u_h \to u \text{ in } L^p(\Omega; \mathbf{R}^m) \right\}$$
$$= \inf\left\{ \limsup_{h\to+\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right) dx : \{u_h\} \subset u + W_0^{1,p}(\Omega; \mathbf{R}^m), \right.$$
$$\left. u_h \to u \text{ in } L^p(\Omega; \mathbf{R}^m) \right\}$$
$$= \int_\Omega f_{\text{hom}}(\nabla u) dx.$$

Remark In the vector-valued case, for $u \in W^{1,p}(\Omega)$, the quantity

$$\lim_{t\to+\infty} \frac{1}{t^n} \inf\left\{ \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y V(x,y)) dx dy : \right.$$
$$\left. V \in L^p(\Omega; W_0^{1,p}(tY; \mathbf{R}^m)) \right\}$$

plays the role of

$$\inf\left\{ \int_{\Omega\times Y} f(y, \nabla u(x) + \nabla_y V(x,y)) dx dy : V \in L^p(\Omega; W_{\text{per}}^{1,p}(Y)) \right\}.$$

Therefore,

Proposition *Assume that f satisfies (H_0), (QC), (H_1), and (H_2) for some $p \in]1, +\infty[$. Then, for every $\Omega \in \mathcal{A}_0$ and $u \in W^{1,p}(\Omega; \mathbf{R}^m)$, the limit below exists, and*

$$\lim_{t\to+\infty} \frac{1}{t^n} \inf\left\{ \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y V(x,y)) dx dy : \right.$$
$$\left. V \in L^p(\Omega; W_0^{1,p}(tY; \mathbf{R}^m)) \right\}$$
$$= \int_\Omega f_{\text{hom}}(\nabla u(x)) dx.$$

Proof: Passage to the limit on t after an application of Castaing selec-

tion theorem.

Various formulation for the limit energy:

$$\lim_{t\to+\infty}\frac{1}{t^n}\inf\Big\{\int_{\Omega\times tY} f(y,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W_0^{1,p}(tY;\mathbf{R}^m))\Big\}$$

$$=\lim_{t\to+\infty}\inf\Big\{\int_{\Omega\times Y} f(ty,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W_0^{1,p}(Y;\mathbf{R}^m))\Big\}$$

$$=\inf_{h\in\mathbf{N}}\inf\Big\{\int_{\Omega\times Y} f(hy,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W_0^{1,p}(Y;\mathbf{R}^m))\Big\}.$$

The proof of the homogenization theorem reduces to the following lemmas, where the presence of the *additional scale parameter* t makes the computations more involved. □

Lemma 1 *Assume that f satisfies (H_0), (QC), (H_1), and (H_2) for some $p\in[1,+\infty[$.Let$\{\varepsilon_h\}\subset]0,+\infty[$ converge to 0, Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega;\mathbf{R}^m)$. Then*

$$\inf\Big\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)dx : \{u_h\}\subset W_0^{1,p}(\Omega;\mathbf{R}^m),\ u_h\to 0 \text{ in } L^p(\Omega;\mathbf{R}^m)\Big\}$$

$$\geq\liminf_{h\to+\infty}\inf\Big\{\int_{\Omega\times Y} f(hy,\nabla u(x)+\nabla_y V(x,y))dxdy : V\in L^p(\Omega;W_0^{1,p}(Y;\mathbf{R}^m))\Big\}.$$

Lemma 2 *Assume that f satisfies (H_0) and (H_1) for some $p\in]1,+\infty[$. Let $\{\varepsilon_h\}\subset]0,+\infty[$ converge to 0, Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega;\mathbf{R}^m)$. Then*

$$\inf\Big\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)dx : \{u_h\}\subset W_0^{1,p}(\Omega;\mathbf{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbf{R}^m)\Big\}$$

$$\leq \inf_{k\in\mathbf{N}} \frac{1}{k^n} \inf\left\{ \int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y V(x,y))dxdy : \right.$$
$$\left. V \in L^p(\Omega; W^{1,p}_{\text{per}}(kY;\mathbf{R}^m)) \right\}.$$

Proof (of Lemma 1): $u \in W^{1,p}(\Omega;\mathbf{R}^m)$.

Parameter doubling to keep into account the asymptotic homogenization formula:

$$\inf\left\{ \liminf_{h\to+\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u + \nabla u_h\right)dx : \{u_h\} \subset W^{1,p}_0(\Omega;\mathbf{R}^m), \right.$$
$$\left. u_h \to 0 \ in \ L^p(\Omega;\mathbf{R}^m) \right\}$$
$$\geq \sup_{\nu\in\mathbf{N}} \inf\left\{ \liminf_{h\to+\infty} \int_\Omega f(\nu hx, \nabla u + \nabla v_h)dx : \{v_h\} \subset W^{1,p}_0(\Omega;\mathbf{R}^m), \right.$$
$$\left. v_h \to 0 \ in \ L^p(\Omega;\mathbf{R}^m) \right\}.$$

Then, we use the periodic unfolding method coupled with the *De Giorgi localization argument* because of the zero boundary datum constraint in the homogenization formula.

$$\sup_{v\in\mathbf{N}} \inf\left\{ \liminf_{h\to+\infty} \int_\Omega f(\nu hx, \nabla u + \nabla v_h)dx : \{v_h\} \subset W^{1,p}_0(\Omega;\mathbf{R}^m), \right.$$
$$\left. v_h \to 0 \text{ in } L^p(\Omega;\mathbf{R}^m) \right\}$$
$$\geq \liminf_{v\to+\infty} \lim_{h\to+\infty} \inf\left\{ \int_{\Omega\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u(x)) + \nabla_y V(x,y))dxdy : \right.$$
$$\left. V \in L^p(\Omega; W^{1,p}_0(Y;\mathbf{R}^m)) \right\}.$$

Finally, the quasiconvexity of f and the properties of the unfolding operator provide

$$\liminf_{\nu\to+\infty} \lim_{h\to+\infty} \inf\left\{ \int_{\Omega\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u(x)) + \nabla_y V(x,y))dxdy : \right.$$
$$\left. V \in L^p(\Omega; W^{1,p}_0(Y;\mathbf{R}^m)) \right\}$$
$$\geq \liminf_{h\to+\infty} \inf\left\{ \int_{\Omega\times Y} f(hy, \nabla u(x) + \nabla_y V(x,y))dxdy : \right.$$
$$\left. V \in L^p(\Omega; W^{1,p}_0(Y;\mathbf{R}^m)) \right\}. \qquad \square$$

Proof (of Lemma 2): $k \in \mathbf{N}$, $U \in C^1(\mathbf{R}^n \times \mathbf{R}^n; \mathbf{R}^m)$ with $U(x, \cdot)$ kY-periodic for every $x \in \Omega$. $u_h(x) = u(x) + \varepsilon_h U(\frac{x}{\varepsilon_h})$.
Parameter doubling.

$$\int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right) dx = \int_{\Omega_{k\varepsilon_h} \times Y} \mathcal{T}_{k\varepsilon_h}\left(f\left(\frac{\cdot}{\varepsilon_h}, \nabla u_h(\cdot)\right)\right)(x, y) dx dy$$
$$= \int_{\Omega_{k\varepsilon_h} \times Y} f\left(ky, \mathcal{T}_{k\varepsilon_h}(\nabla u)(x, y) + \varepsilon_h \nabla_x U\left(k\varepsilon_h\left[\frac{x}{k\varepsilon_h}\right] + k\varepsilon_h y, ky\right)\right.$$
$$\left. + \nabla_y U\left(k\varepsilon_h\left[\frac{x}{k\varepsilon_h}\right] + k\varepsilon_h y, ky\right)\right) dx dy.$$

As h goes to $+\infty$,

$$\inf\left\{\limsup_{h \to +\infty} \int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right) dx : \{v_h\} \subset W^{1,p}(\Omega; \mathbf{R}^m),\right.$$
$$\left. v_h \to u \text{ in } L^p(\Omega; \mathbf{R}^m)\right\}$$
$$\leq \frac{1}{k^n} \int_{\Omega \times kY} f(y, \nabla u(x) + \nabla_y U(x, y)) dx dy,$$

for every $U \in C^1(\mathbf{R}^n \times \mathbf{R}^n; \mathbf{R}^m)$ with $U(x, \cdot)$ kY-periodic for every $x \in \Omega$.

A density argument, and the consideration of the infimum on k, completes the proof. □

7 Homogenization of pointwise gradient constrained convex energies

Homogenization problem in the presence of pointwise oscillating gradient constraints are considered in several papers: first conjectured in [BLP] (1978), results are proved (among others) in [CSa] (1984/85), [CoDA] (1994), [CDA] (2001) via measure-theoretical arguments.

The periodic unfolding method provides a simpler and more direct proof.

The energy density f satisfies

$$(H_0) \quad \begin{cases} f : (x, z) \in \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{f}(\mathbf{x}, \mathbf{z}) \in [\mathbf{0}, +\infty] \\ f\mathcal{L}(\mathbf{R}^n) \times \mathcal{B}(\mathbf{R}^n)\text{-measurable} \\ f(\cdot, z)Y\text{-periodic for every} z \in \mathbf{R}^n, f(x, \cdot)\text{convex fora.e. } x \in \mathbf{R}^n. \end{cases}$$

$p \in [1, +\infty]$, $a \in L^1_{\text{per}}(Y)$, $M \geq 0$, $r > 0$

(H_1) $\quad f(x,z) \leq a(x)+M|z|^p$ for a.e. $x \in \mathbf{R}^n$ and every $z \in \operatorname{dom} f(x,\cdot)$,

(H_2) $\quad |z|^p \leq f(x,z)$ for a.e. $x \in \mathbf{R}^n$ and every $\mathbf{z} \in \mathbf{R}^N$,

(B) $\quad B_r \subset \operatorname{dom} f(x,\cdot)$ for a.e. $x \in \mathbf{R}^n$.

$$f_{\text{hom}} : z \in \mathbf{R}^n \mapsto \inf\left\{\int_Y f(y, z+\nabla v(y))dy : v \in W^{1,p}_{\text{per}}(Y)\right\}.$$

Theorem *Let $p \in]n,+\infty[$, $a \in L^1_{\text{per}}(Y)$, $\mathrm{M} \geqslant 0, \mathrm{r} > 0$, let f satisfies (H_0), (H_1), (H_2), (B). Then f_{hom} is convex, lower semicontinuous, and*

$$|z|^p \leq f_{\text{hom}}(z) \text{ for every } z \in \mathbf{R}^n,$$

$$B_r \subset \operatorname{dom} f_{\text{hom}}.$$

Let Ω be a convex bounded open set, and let $\{\varepsilon_h\} \subset]0,+\infty[$ converge to 0. Then, for every u in $W^{1,p}(\Omega)$

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h(x)\right)dx : \{u_h\} \subset W^{1,p}(\Omega),\right.\\
&\qquad \left. u_h \to u \text{ in } C^0(\overline{\Omega})\right\}\\
&= \inf\left\{\limsup_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h(x)\right)dx : \{u_h\} \subset W^{1,p}(\Omega),\right.\\
&\qquad \left. u_h \to u \text{ in } C^0(\overline{\Omega})\right\}\\
&= \inf\left\{\int_{\Omega\times Y} f(y, \nabla u(x) + \nabla_y V(x,y))dxdy : V \in L^p(\Omega; W^{1,p}_{\text{per}}(Y))\right\}\\
&= \int_\Omega f_{\text{hom}}(\nabla u(x))dx.
\end{aligned}$$

As consequence, for every $u \in W^{1,p}_0(\Omega)$,

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right)dx : \{u_h\} \subset W^{1,p}_0(\Omega),\right.\\
&\qquad \left. u_h \to u \text{in } C^0(\overline{\Omega})\right\}\\
&= \inf\left\{\limsup_{h\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_h}, \nabla u_h\right)dx : \{u_h\} \subset W^{1,p}_0(\Omega),\right.\\
&\qquad \left. u_h \to u \text{ in } C^0(\overline{\Omega})\right\}\\
&= \int_\Omega f_{\text{hom}}(\nabla u)dx.
\end{aligned}$$

As corollaries, homogenization results under the "model case" assumptions

$$f(x,z) \leq a(x) \text{ for a.e. } x \in \mathbf{R}^n, \text{ and every } \mathbf{z} \in \mathrm{dom}\mathbf{f}(\mathbf{x},\cdot),$$
$$\mathrm{dom} f(x,\cdot) \subset \overline{B_R} \text{ for a.e. } x \in \mathbf{R}^n,$$

where a $\in L^1_{\mathrm{per}}(Y)$, and $R > 0$.

Since f may take the value $+\infty$, the above integrals involve pointwise gradient constrains. Indeed, for fixed $h \in \mathbf{N}$, the configurations v that make the integral $\int_\Omega f(\frac{x}{\varepsilon_h}, \nabla v)dx$ finite must satisfy the constraint $\nabla v(x) \in \mathrm{dom} f(\frac{x}{\varepsilon_h}, \cdot)$ for a.e. $x \in \Omega$. This occurs in the "classical" elasto-plastic torsion case when $f(x,z) = |z|^2 + I_{[0,\varphi(\frac{x}{\varepsilon_h})]}(|z|)$, where φ is a nonnegative Y-periodic function in $L^\infty(Y)$, and $I_{[0,\varphi(\frac{x}{\varepsilon_h})]}(|z|)$ is the indicator function of $[0, \varphi(\frac{x}{\varepsilon_h})]$. This corresponds to the gradient constraint $|\nabla v(x)| \leq \varphi(\frac{x}{\varepsilon_h})$ for a.e. $x \in \Omega$.

In summary, here again, the ingredients for the proof are the periodic unfolding method and some abstract inner regularity results (which hold only for convex open sets).

References

[B] Braides A.: *Homogenization of Some Almost Periodic Coercive Functional*; Rend. Accad. Naz. Sci. XL Mem. Mat. **9**, (1985), 313-322

[BLP] Bensoussan A.: Lions J. L., Papanicolaou G. : "*Asymptotic Analysis for Periodic Structures*"; Stud. Math. Appl. **5**, North Holland (1978)

[CDA] Carhone L., De Arcangelis R.: "*Unbounded Functionals in the Calculus of Variations. Representation, Relaxation, and Homogenization*"; Chapman and Hall/CRC Monogr. Surv. Pure Appl. Math. **125**, Chapman and Hall/CRC, Boca Raton, FL (2001)

[CoDA] Corbo Esposito A., De Arcangelis R.: *Homogenization of Dirichlet Problems with Nonnegative Bounded Constraints on the Gradient*; J. Analyse Math. **64**, (1994), 53-96

[CDDA1] Cioranescu D., Damlamian A., De Aracngelis R.: *Homogenization of Nonlinear Integrals via the Periodic Unfolding Method*; C. R. Acad. Sci. Paris Sér. I Math. **339**, (2004), 77-82.

[CDDA2] Cioranescu D., Damlamian A., De Arcangelis R.: *Homogenization of Quasiconvex Integrals via the Periodic Unfolding Method*, SIAM Journal on Mathematical Analysis (SIMA)Vol. 37, No. 5, pp. 14351453 (2006)

[CDDA3] Cioranescu D., Damlamian A., De Arcangelis R.: *Homogenization of Integrals with Pointwise Gradient Constraints via the Periodic Unfolding Method*, Ricerche di Matematica 55: 3153 (2006) (DOI: 10.1007/s11587-006-0003-0)

[CDG] Cioranescu D., Damlamian A., Griso, G.: *Periodic Unfolding and Homogenization.* C. R. Math. Acad. Sci. Paris 335, 99-104 (2002)

[CSa] Caubone L., Salerno S.: *Further Results on a Problem of Homogenization with Constraints on the Gradient*; J. Analyse Math. **44**, (1984/85), 1-20

[CSb] Carbone L., Sbordone C.: *Some Properties of* Γ*-Limits of Integral Functionals*; Ann. Mat. Pura Appl. (4), **122**, (1979), 1-60

[Ma] Marcellini P.: *Periodic Solutions and Homogenization of Non Linear Variational Problems*; Ann. Mat. Pura Appl. (4), **117**, (1978), 139-152

[Mü] Müller S.: *Homogenization of Nonconvex Integral Functionals and Cellular Elastic Materials*; Arch. Rational Mech. Anal. **99**, (1987), 189-212

Sparse System Solution and the HSL Library*

Iain S. Duff
Computational Science and Engineering Department
Rutherford Appleton Laboratory, Oxfordshire, UK.
E-mail: i.s.duff@rl.ac.uk

Abstract

We consider the solution of large sparse systems, sketch their ubiquity, and briefly describe some of the algorithms used to solve these systems.

The HSL mathematical software library started life in 1963 as the Harwell Subroutine Library making it one of the oldest such libraries. The main strengths of the Library lie in packages for large scale system solution. It is particularly strong in direct methods for sparse matrices and optimization. The Library has been used worldwide by a wide range of industries.

We briefly discuss the history of the library and its organization and contents. We discuss the evolution of some of our current packages and the efforts to ensure reliability, robustness, and efficiency.

We describe in some detail the functionality of one of our most popular sparse direct codes.

1 Introduction

Sparse systems of linear equations

$$\mathbf{Ax} = \mathbf{b} \tag{1.1}$$

are simply sets of linear equations where the coefficient matrix $\mathbf{A}$ has sufficient zero entries to cause it to be beneficial to exploit this fact. Thus we are not here concerned with any particular structure or any particular application area. In fact, we emphasize in Section 2 the ubiquity of sparse systems and illustrate the diversity of their structure. In Section 3, we briefly introduce the main elements of the direct solution of sparse equations indicating their complexity on a range of structures. We then discuss the origins, structure, and development of HSL in Section 4 before a brief summary of the paper in Section 5.

*The author is partly supported by the EPSRC grant GR/S42170/01).

2 Sparse matrices

Sparse matrices have arisen naturally in numerical applications since the mid 1960s. Some of the earliest applications involving the solution of sparse systems with a general structure were in the solution of ordinary differential equations using backward difference formulae and in linear programming. The former area was the main driving force in the development of sparse matrix methods at AERE Harwell in the 1970s. We list some of the major numerical application areas stimulating and benefiting from sparse matrix research in Table 1.

Table 1 Some numerical applications

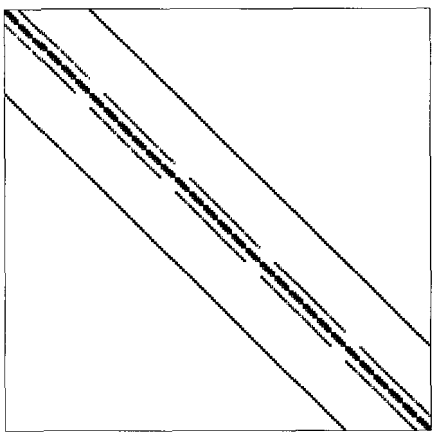
Stiff ODEs ... BDF ... Sparse Jacobian
Linear Programming
..... simplex
..... interior point
Optimization/Nonlinear Equations
Elliptic Partial Differential Equations
Eigensystem Solution
Two Point Boundary Value Problems
Least Squares Calculations

In a more general context, we can look at application areas that often make extensive use of sparse matrices or sparse equation solvers. We show a range of these in Table 2. In this list, standard applications in the hard sciences are listed along with slightly more esoteric applications in the soft sciences.

Table 2 Application areas

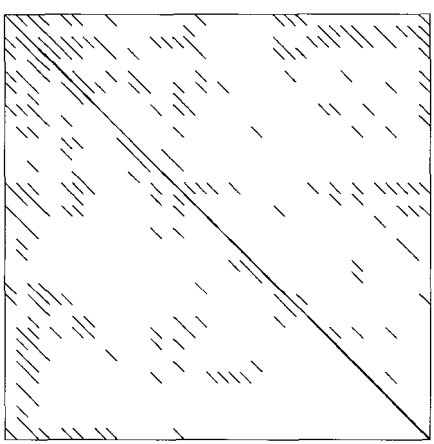
Physics	CFD
	Lattice gauge
	Atomic spectra
Chemistry	Quantum chemistry
	Chemical engineering
Civil engineering	Structural analysis
Electrical engineering	Power systems
	Circuit simulation
	Device simulation
Geography	Geodesy
Demography	Migration
Economics	Economic modelling
Behavioural sciences	Industrial relations
Politics	Trading
Psychology	Social dominance
Business administration	Bureaucracy
Operations research	Linear Programming

We now show some pictures of sparse matrices from various applications in order to illustrate different structures for sparse matrices. The thermal simulation example exhibits a structure which is very structured and familiar to most of you. It is typical of a matrix arising in the finite-difference discretization of a three-dimensional elliptic PDE. In this case, the inclusion of thermal terms gives this matrix, from oil reservoir modelling, interesting properties.



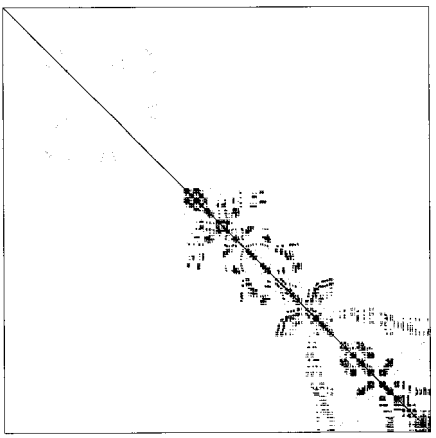
Thermal Simulation; SHERMAN2

Our so-called weather matrix is somewhat more interesting and models the combination of chemical kinetics and atmospheric transport. It is in fact a block matrix where each block is diagonal or tridiagonal and comes from studies in atmospheric pollution, a hot topic in environmental science.



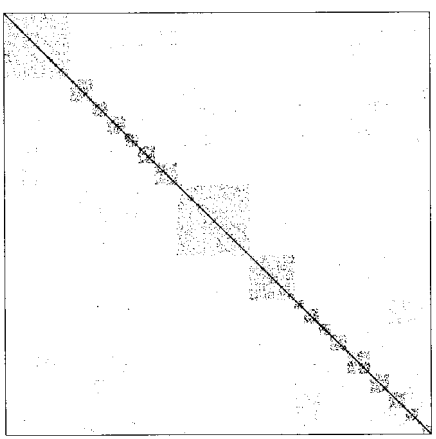
Weather Matrix; FS 760 3

The matrix from dynamic calculations is typical of a matrix arising from a finite-element discretization of a structures problem, in this case in a study of the effect of earthquake vibrations on a building in western USA.



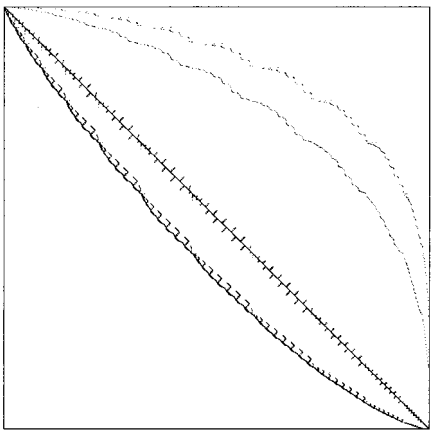
Dynamic Calculation in Structures; BCSSTM13

The power system matrix also comes from western USA. Those with an eagle eye will see that the matrix is not quite block diagonal but there are only few entries outside the diagonal blocks that are themselves sparse. The blocks correspond to the power system network for a single utility and the off-diagonal entries to the much fewer links between the utilities, that often only carry loads when there is a problem in one utilities capacity.



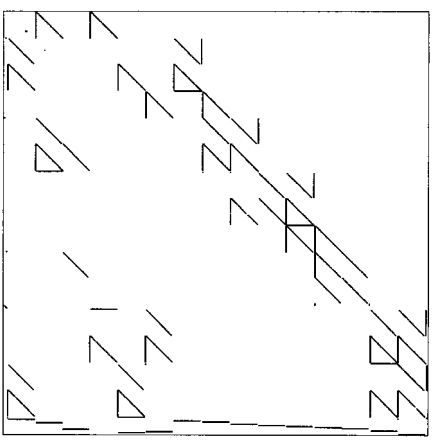
Power Systems; BCSPWR07

The matrix from the simulation of computing system comes from a Markov model of a computing system and has the remarkable anti-symmetric property that if there is an entry a_{ij} then the entry a_{ji} does not exist, for all $i \neq j$.



Simulation of Computing Systems; GRE 1107

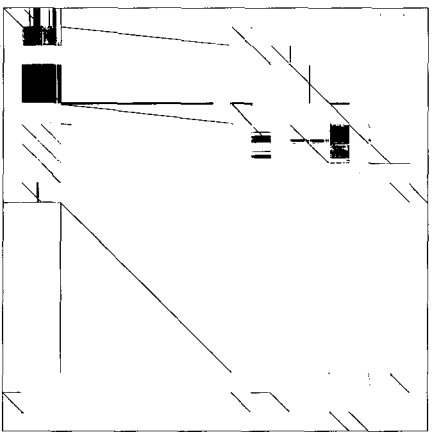
The chemical engineering industry is a rich source of unsymmetric sparse matrices that are particularly challenging for solution by iterative methods. There is certainly a structure to the matrix but somewhat more irregular than the earlier examples. Notice that the diagonal is nearly all zero.



Chemical Engineering; WEST0381

The matrix from an econometric input/output model from South

East Asia also has considerable structure but not one that can be exploited like that of the first matrix we displayed.



Economic Modelling; ORANI678

Duff and Reid distributed a set of sparse matrices from Harwell in the late 1970s but the main test collection for many years was the Harwell-Boeing Sparse Matrix Collection. This is available by anonymous ftp from `ftp.numerical.rl.ac.uk` in directory pub/harwell_boeing or from the Web page `http://www.cse.clrc.ac.uk/nag/hb/hb.shtml`

This set was later developed by Duff, Grimes, and Lewis [7] to include larger matrices in a wider range of application areas and to define more language-friendly formats and auxiliary files for other matrix properties (for example, eigenvalues) and associated information (for example, sparsity orderings). The Rutherford-Boeing Sparse Matrix Collection [8] will be supported by the GRID-TLSE Project `http://www.enseeiht.fr/lima/tlse` and is also available by anonymous FTP to `ftp.cerfacs.fr` in the directory pub/algo/matrices or `http://www.cerfacs.fr/algor/Softs/RB/index.html`

An extended set of test matrices available from Tim Davis at `http://www.cise.ufl.edu/research/sparse/matrices` and Matrix market at `http://math.nist.gov/MatrixMarket`.

3 Direct Methods

Although equation (1.1) nominally has the solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

it must be stressed that this should only be thought of notationally. It is really crucial that one must not use or even think of the inverse of $\mathbf{A}$ in this context.

For sparse $\mathbf{A}$, $\mathbf{A}^{-1}$ is usually dense. Indeed, if $\mathbf{A}$ is irreducible, one can prove [5] that $\mathbf{A}^{-1}$ will always be dense in a structural sense. That is, there exists a set of entries in the original sparsity pattern of $\mathbf{A}$ that make any position in the matrix $\mathbf{A}^{-1}$ nonzero.

Examples of sparse matrices that are very sparse but have dense inverses are tridiagonal and arrowhead matrices, where an arrowhead matrix has entries only in all positions on the diagonal and the last row and column. These examples are particularly interesting since, although their inverses are dense, linear systems involving these as coefficient matrices can be solved with no extra storage, as we shall shortly show.

If we thus dismiss the use of the inverse, we must propose other methods for solving the systems of the form (1.1). In some instances, iterative methods [11, 12] can be used, often based on Krylov sequences, but these are not guaranteed to converge on general systems and usually require very sophisticated preconditioning so we do not consider them further here. Instead we look at direct methods for solution [6] that involve some matrix factorization representation of the inverse. The methods that we consider here are all based on Gaussian Elimination, that generates the factorization:

$$\mathbf{PAQ} \rightarrow \mathbf{LU} \tag{3.1}$$

where permutations $\mathbf{P}$ and $\mathbf{Q}$ are chosen to preserve sparsity and maintain stability, and $\mathbf{L}$ and $\mathbf{U}$ are lower and upper triangular matrices, respectively. When $\mathbf{A}$ is symmetric, the factorization is of the form

$$\mathbf{PAP^T} \rightarrow \mathbf{LDL^T}. \tag{3.2}$$

The solution to equation (1.1) is then easily obtained by solving the lower triangular system

$$\mathbf{Ly} = \mathbf{Pb}$$

followed by the upper triangular system

$$\mathbf{UQ^Tx} = \mathbf{y}.$$

Clearly, as in the case for dense systems, most of the work is usually in the factorization. The work in the forward and back substitution is proportional to the number of entries in the factors. This subdivision of work is reflected in software for sparse direct methods. Although the exact subdivision of tasks for sparse direct solution will depend on the algorithm and software being used, a common subdivision is given by:

Analyse An analysis phase where the matrix structure is analysed to produce a suitable ordering and data structures for efficient factorization.

Factorize A factorization phase where the numerical factorization is performed.

Solve A solve phase where the factors are used to solve the system using forward and backward substitution.

We note the following:

- ANALYSE is sometimes preceded by a preordering phase to exploit structure.
- For general unsymmetric systems, the ANALYSE and FACTORIZE phases are sometimes combined to ensure the ordering does not compromise stability.
- The concept of separate ANALYSE and FACTORIZE phases is not present for dense systems.

It is crucially important to try to ensure sparsity in the factors **L** and **U**. This is done by choosing an ordering for the elimination. For example, if we pivot down the diagonal of the matrix in the left-hand side of Figure 3.1 then the resulting matrix of factors will be dense, as shown on the right-hand side of Figure 3.1.

$$\begin{matrix} \times & \times & \times & \times & \times \\ \times & \times & & & \\ \times & & \times & & \\ \times & & & \times & \\ \times & & & & \times \end{matrix} \quad \xrightarrow{\mathbf{L/U}} \quad \begin{matrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{matrix}$$

Figure 3.1 Factorization of reverse arrowhead matrix

However, if we permute this matrix symmetrically to put the last row and column to the end, obtaining the arrowhead matrix shown on the left-hand side of Figure 3.2, then the factors require no more space than the original matrix as shown on the right-hand side of Figure 3.2.

$$\begin{matrix} \times & & & & \times \\ & \times & & & \times \\ & & \times & & \times \\ & & & \times & \times \\ \times & \times & \times & \times & \times \end{matrix} \quad \xrightarrow{\mathbf{L/U}} \quad \begin{matrix} \times & & & & \times \\ & \times & & & \times \\ & & \times & & \times \\ & & & \times & \times \\ \times & \times & \times & \times & \times \end{matrix}$$

Figure 3.2 Factorization of arrowhead matrix

The complexity of LU factorization on a dense matrix of order n is:

$$\frac{2}{3}n^3 + \mathcal{O}(n^2) \quad \text{floating-point operations (flops)}$$
$$n^2 \quad \text{storage,}$$

while, for a band matrix (order n, semi-bandwidth k), it is:

$$2k^2n \text{ flops}, 2nk \text{ storage.}$$

For a five-diagonal matrix (on a $k \times k$ grid) as would arise in the finite-difference discretization of a two-dimensional Laplacian, the complexity is:

$$\mathcal{O}(k^3) \text{ flops} \quad \text{and} \quad \mathcal{O}(k^2 \log k) \text{ storage}$$

while, for a tridiagonal or arrowhead matrix, the complexity is:

$$\mathcal{O}(n) \text{ work} \quad \text{and} \quad \text{storage.}$$

Indeed our target complexity for sparse matrix computations is $\mathcal{O}(n) + \mathcal{O}(\tau)$ for a sparse matrix of order n with τ entries.

4 HSL

4.1 Mathematical software libraries

The benefits and advantages of using high quality mathematical software libraries include:

- Shorten application development cycle, cutting time-to-market and gaining competitive advantage
- Reduce development costs
- Increase modularity
- More time to focus on specialist aspects of applications
- Improve application accuracy and robustness

4.2 HSL

HSL began life as the Harwell Subroutine Library in 1963 and was originally developed by Mike Powell and Mike Hopper as an internal library for users of the IBM mainframe at AERE Harwell. However, the reputation of the Harwell Subroutine Library spread so quickly that it was being sent out to external users on request as early as 1964. HSL packages are

now used worldwide by academics and commercial organisations, and are incorporated into a large number of commercial products.

HSL is now a collection of portable, fully documented and tested packages in standard Fortran, primarily written and developed by the Numerical Analysis Group at the Rutherford Appleton Laboratory although some routines have been written by visitors, colleagues and collaborators, and students of staff at RAL. The particular strengths are currently:

- Sparse matrix computations
- Optimization
- Large-scale system solution

There are two libraries: HSL 2004 and HSL Archive. HSL Archive consists of superseded routines and public domain software and is free for non-commercial use. All codes need a licence although academic and commercial are differentiated.

The most recent version of HSL is called HSL 2004 and was released in January 2004. HSL is marketed by Hyprotech UK, which was acquired by Aspen Technology in May 2002. For further details see: `www.cse.clrc.ac.uk/nag/hsl`

4.3 Organization of HSL

The HSL is organized into chapters, each identified by two letters, for example:

- MA: matrix routines (solvers)
- MC: matrix routines (manipulation)
- EB: unsymmetric eigensystems

Within each chapter, each package has a 2-digit identifier, generally allocated chronologically, for example:

- `MA48`: package for solving unsymmetric sparse equations
- `MA49`: package for sparse QR factorization and for solving sparse least-squares

Following the Fortran 77 convention limiting the length of character strings, each subroutine has a six character identifier, for example:

- `MA48AD`: double precision analysis subroutine of `MA48` package
- `MA57BD`: double precision factorize subroutine of `MA57` package

More recently, the prefix HSL_ (for example, HSL_MA48) has been used to identify Fortran 90 or 95 packages.

The number of routines in the main chapters in HSL is shown in Table 3.

MA	Matrix solution	26
MC	Matrix manipulation	33
ME	Complex matrices	8
MI	Iterative solvers and preconditioners	9
MP	MPI packages .. all solvers	4
E	Eigensystems	8
V	Optimization	13

Table 3 Number of routines in some major chapters of HSL.

Each package has a specification sheet, a short "demo" test program, and an exhaustive test deck.

Fortran source code is always provided. Version numbers in the form a.b.c have been recently introduced to HSL. Changes to c are very minor, perhaps involving changes to comments in the code. The level b represents minor bug fixes, while at level a we expect more major fixes and perhaps new entries or facilities.

4.4 Development of HSL

HSL is both **revolutionary** and **evolutionary**.

By **revolutionary**, we mean that codes have been introduced that are radically different in technique and algorithm design than anything that has preceded them. Examples of this are:

MA18	First sparse direct code	1971
MA27	First multifrontal code	1982

By **evolutionary**, we mean that some of our codes evolve, sometimes as a result of major changes in programming paradigm and sometimes because of added functionalities. Examples of this morphing are:

$$\begin{aligned} \texttt{MA18} \longrightarrow \texttt{MA28} \longrightarrow \texttt{MA48} \\ \texttt{MA32} \longrightarrow \texttt{MA42} \longrightarrow \\ \texttt{MA37} \longrightarrow \texttt{MA41} \longrightarrow \\ \texttt{MA17} \longrightarrow \texttt{MA27} \longrightarrow \texttt{MA57} \end{aligned}$$

We look in more detail at an example of evolution by giving more details for the last example above, namely our flagship code for symmetric sparse systems. The history of our HSL codes for solving the symmetric system with $\mathbf{A} = \mathbf{A}^{\mathbf{T}}$ is shown in Figure 4.1.

- `MA17` ... 1971 (Curtis and Reid)
  - Sparse symmetric
  - LDL^T .. 1×1 pivots only
  - linked lists
- `MA27` ... 1982 (Duff and Reid)
  - Sparse symmetric indefinite
  - LDL^T .. 1×1 and 2×2 pivots
  - Multifrontal
- `MA47` ... 1995 (Duff and Reid)
  - Sparse symmetric indefinite structured
  - LDL^T .. 1×1 and 2×2 pivots, with $\begin{smallmatrix} \times & \times \\ \times & 0 \end{smallmatrix}$ and $\begin{smallmatrix} \times & \times \\ \times & 0 \end{smallmatrix}$ pivots
  - Multifrontal
- `MA57` ... 2000 (Duff)
  - See Section 4.5

Figure 4.1 Example of evolution

4.5 HSL code `MA57`

To show the sophistication of our recent codes, we show the features of our current flagship code for solving sparse symmetric equations `MA57` in Figure 4.2.

- Analysis (ordering and symbolic factorization)
  - AMD (Approximate Minimum Degree) ordering
- Factorization ($PAP^T \longrightarrow LDL^T$)
  - Factorizes singular matrices
  - Pivoting options (including Schnabel-Eskow)
  - Stop and restart (or discard factors)
  - Option to return or alter pivots
- Solve (Fwd/Bwd substitution)
  - Several entries for error analysis and iterative refinement
  - Multiple rhs (using level 3 BLAS)
  - Partial solve (using L, D, or L^T)

Figure 4.2 `MA57` features

Additions made to Version 3.x.y of MA57 are:

- METIS nested dissection ordering available as an option
- Automatic choice of ordering ... decision made from matrix characteristics
- Built-in option for scaling matrix (transparent to user)
- Static pivoting option

We show the first and second pages of the specification sheet for MA57 in Figures 4.3 and 4.4. The first page shows that the structure of the code follows the subdivision of direct solution methods that we discussed earlier. On the second page, we see details of the call for the analysis entry where the number of parameters are reduced by combing control and information parameters into arrays. In a Fortran 90 code, of course, the work arrays can be made internal and dynamic and derived data types can be used to create more structure and further reduce the parameter list.

4.6 Parallel codes in HSL

In recent years, we have introduced some parallel codes to HSL. The earliest parallel code was an OpenMP version of MA41.

Work on this code [1] was later developed by teams originally at RAL and CERFACS and now also at Lyon, ENSEEIHT-IRIT, and Bordeaux to produce the much downloaded MUMPS package [2]. Note that this package is freely available by request to mumps@cerfacs.fr but is not in HSL.

MPI-based routines that are available in HSL are in the MP chapter:

HSL_MP42	Multiple front method .. equation entry
HSL_MP43	Multiple front method .. element entry
HSL_MP62	Symmetric element entry multiple front
HSL_MP48	General unsymmetric using singly bordered block diagonal form

4.7 HSL summary

It is impossible to discuss in detail the many sparse codes in HSL in an article of this kind but we present a list of HSL sparse codes in Table 4.

aspentech CCLRC

HSL MA57

PACKAGE SPECIFICATION HSL 2004

1 SUMMARY

To **solve a sparse symmetric system of linear equations.** Given a sparse symmetric matrix $\mathbf{A}=\{a_{ij}\}_{n\times n}$ and an n-vector **b** (or an $n\times s$ matrix **B**), this subroutine solves the system **Ax=b** (**AX=B**). The matrix **A** need not be definite.

The multifrontal method is used. It is a direct method based on a sparse variant of Gaussian elimination and is discussed further by Duff and Reid, ACM Trans. Math. Software 9 (1983), 302-325. A detailed discussion on the MA57 strategy and performance is given by Duff, ACM Trans. Math. Software 30 (2004), 118-144. More recent work on pivoting and scaling strategies is given in the Technical Report RAL-TR-2004-020 by Duff and Pralet. This will be published in the SIAM Journal on Matrix Analysis and Applications and can be obtained from the web site

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

The MA57 package has a range of options including several sparsity orderings, multiple right-hand sides, partial solutions, error analysis, scaling, a matrix modification facility, and a stop and restart facility. Although the default settings should work well in general, there are several parameters available to enable the user to tune the code for his or her problem class or computer architecture.

ATTRIBUTES — Version: 3.0.1. **Types:** Real (single, double) **Calls:** FD15, MC64, MC71, MC47, BLAS routines _GEMM, _TPSV, and _GEMV, and (optionally) METIS_NODEND from the MeTiS package. **Remark:** Supersedes MA27. **Original date:** Original code: September 2000, Version 3.0.0: March 2005. **Origin:** I. S. Duff, Rutherford Appleton Laboratory.

2 HOW TO USE THE PACKAGE

2.1 Argument lists and calling sequences

There are five entries:

(a) MA57I/ID sets default values for the components of the arrays that hold control parameters. Normally the user will call MA57I/ID prior to any call to MA57A/AD. If non-default values for any of the control parameters are required, they should be set immediately after the call to MA57I/ID.

(b) MA57A/AD accepts the pattern of **A** and chooses pivots for Gaussian elimination using a selection criterion to preserve sparsity. It subsequently constructs subsidiary information for the actual factorization by MA57B/BD. The user may provide the pivot sequence; in which case only the necessary information for MA57B/BD will be generated.

(c) MA57B/BD factorizes a matrix **A** using the information from a previous call to MA57A/AD. The actual pivot sequence used may differ from that of MA57A/AD if **A** is not definite.

(c) MA57C/CD uses the factors generated by MA57B/BD to solve a system of equations **Ax=b** (**AX=B**).

(c) MA57D/DD uses the factors generated by MA57B/BD to solve a system of equations **Ax=b** (**AX=B**) using iterative refinement and (optionally) returning estimates of the error.

(d) MA57E/ED maps the data into new larger size arrays following a failure of MA57B/BD through insufficient storage so that the numerical factorization can be continued using these larger arrays.

A call to MA57C/CD or MA57D/DD must be preceded by a call to MA57B/BD which in turn must be preceded by a call to MA57A/AD. Since the information passed from one subroutine to the next is not corrupted by the second, several calls to MA57B/BD for matrices with the same sparsity pattern but different values may follow a single call to

All use is subject to licence, see http://hsl.rl.ac.uk/acuk/cou.html
For any commercial application, a separate licence must be signed.

HSL 2004 MA57 Version 3.0.1 Documentation date: 21st May 2006 1

Figure 4.3 First page of MA57 specification sheet

MA57A/AD, and similarly MA57C/CD or MA57D/DD can be used repeatedly to solve for different right-hand sides b (B). Note that it would be possible to use MA57A/AD on several matrices before calling MA57B/BD. When we state that parameters "must be unchanged since the call to *subroutine*", we mean a successful call of the routine on the same matrix.

2.1.1 To set default values of controlling parameters

The single precision version

```
CALL MA57I(CNTL,ICNTL)
```

The double precision version

```
CALL MA57ID(CNTL,ICNTL)
```

CNTL is a REAL (DOUBLE PRECISION in the D version) array of length 5 that need not be set by the user. On return it contains default values. For further information see Section 2.2.

ICNTL is an INTEGER array of length 20 that need not be set by the user. On return it contains default values. For further information see Section 2.2.

2.1.2 To perform symbolic manipulations

The single precision version

```
CALL MA57A(N,NE,IRN,JCN,LKEEP,KEEP,IWORK,ICNTL,INFO,RINFO)
```

The double precision version

```
CALL MA57AD(N,NE,IRN,JCN,LKEEP,KEEP,IWORK,ICNTL,INFO,RINFO)
```

N is an INTEGER variable that must be set by the user to the order n of the matrix **A**. It is not altered by the subroutine. **Restriction:** $N \geq 1$.

NE is an INTEGER variable that must be set by the user to the number of entries input in IRN and JCN. It is not altered by the subroutine. **Restriction:** $NE \geq 0$.

IRN and JCN are INTEGER arrays of length NE. The user must set them so that each diagonal entry a_{ii} is represented by IRN(k)=i and JCN(k)=i and each pair of off-diagonal entries a_{ij} and a_{ji} is represented by IRN(k)=i and JCN(k)=j or by IRN(k)=j and JCN(k)=i. Entries (on or off the diagonal) that are known to be zero can be excluded. Multiple entries are permitted. If IRN(k) or JCN(k) are less than 1 or greater than N the entry is ignored. These arrays are not altered by any of the calls to the MA57 package. They must be preserved by the user between this call and a call to MA57D/DD for the same matrix.

LKEEP is an INTEGER variable that must be set by the user to the length of array KEEP. It might be more efficient to allocate more than the minimum required, say about N to 2*N more space. **Restriction:** LKEEP $\geq$ 5*N+NE+MAX(N,NE)+42.

KEEP is an INTEGER array of length LKEEP. It need not be set by the user and must be preserved between a call to MA57A/AD and subsequent calls to MA57B/BD. If the user wishes to input the pivot sequence, the position of variable i in the pivot order should be placed in KEEP(I), I=1,2,..., N and ICNTL(6) should be set to 1. The subroutine may replace the given order by another that gives the same fill-in pattern and virtually identical numerical results.

IWORK is an INTEGER array of length 5*N that is used as workspace.

ICNTL is an INTEGER array of length 20 that contains control parameters and must be set by the user. Default values for the components may be set by a call to MA57I/ID. Details of the control parameters are given in Section 2.2.

INFO is an INTEGER array of length 40 that need not be set by the user. On return from MA57A/AD, a non-negative value for INFO(1) indicates that the subroutine has performed successfully. For nonzero values, see Section

All use is subject to licence, see http://hsl.rl.ac.uk/acuk/cou.html
For any commercial application, a separate licence must be signed.

Figure 4.4 Second page of MA57 specification sheet

Table 4 Some of the sparse matrix codes in HSL that use direct methods. In many cases there is also a version for complex matrices. There are parallel versions of MA41, using OpenMP, and of MA42, MA43, MA62, and MA48 using MPI. An out-of-core multifrontal code will soon be available.

Package	System solved	Algorithm
MA38	Unsymmetric assembled	Multifrontal
MA41	Unsymmetric assembled	Multifrontal
MA42	Unsymmetric assembled and unassembled	Frontal
MA43	Unsymmetric assembled	Frontal
MA45	Weighted least squares	Normal equations
MA46	Unsymmetric unassembled	Multifrontal
MA48	Unsymmetric assembled	Markowitz-Threshold
MA49	Rectangular assembled	Multifrontal QR
MA55	Symmetric positive definite	Variable band
MA57	Symmetric indefinite assembled	Multifrontal
MA62	Symmetric definite unassembled	Frontal
MA67	Symmetric indefinite structured	Zero-tracking

5 Summary

The twin aims of our talk are to emphasize the ubiquity of sparse matrices and the availability of high quality codes for solving sparse systems with HSL. We should stress that there are several packages available elsewhere, for example, the already mentioned MUMPS, although we do not know of a greater concentration of codes than in HSL.

To sum up:

- Sparse matrices occur in very many application areas.
- Sparse direct methods can be used to robustly solve large sparse problems.
- There are many packages available that implement direct methods.
- There are several packages implementing direct methods in HSL (`www.cse.clrc.ac.uk/nag/hsl`).

References

[1] P. R. Amestoy and I. S. Duff. Vectorization of a multiprocessor multifrontal code. *Int. J. of Supercomputer Applics.*, 3:41–59, 1989.

[2] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Analysis and Applications*, 23(1):15–41, 2001.

[3] A. R. Curtis and J. K. Reid. Fortran subroutines for the solution of sparse sets of linear equations. Technical Report AERE R 6844, Her Majesty's Stationery Office, London, 1971.

[4] I. S. Duff. MA57 – A new code for the solution of sparse symmetric indefinite systems. Technical Report RAL-TR-2002-024, Rutherford Appleton Laboratory, Oxfordshire, England, 2002. To appear in *ACM Trans. Math. Softw.*

[5] I. S. Duff, A. M. Erisman, C. W. Gear, and J. K. Reid. Sparsity structure and Gaussian elimination. *SIGNUM Newsletter*, 23(2):2–8, April 1988.

[6] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct Methods for Sparse Matrices.* Oxford University Press, Oxford, England, 1986.

[7] I. S. Duff, R. G. Grimes, and J. G. Lewis. Sparse matrix test problems. *ACM Trans. Math. Softw.*, 15(1):1–14, March 1989.

[8] I. S. Duff, R. G. Grimes, and J. G. Lewis. The Rutherford-Boeing Sparse Matrix Collection. Technical Report RAL-TR-97-031, Rutherford Appleton Laboratory, Oxfordshire, England, 1997. Also Technical Report ISSTECH-97-017 from Boeing Information & Support Services, Seattle and Report TR/PA/97/36 from CERFACS, Toulouse.

[9] I. S. Duff and J. K. Reid. MA27 – A set of Fortran subroutines for solving sparse symmetric sets of linear equations. Technical Report AERE R10533, Her Majesty's Stationery Office, London, London, 1982.

[10] I. S. Duff and J. K. Reid. MA47, a Fortran code for direct solution of indefinite sparse symmetric linear systems. Technical Report RAL 95-001, Rutherford Appleton Laboratory, Oxfordshire, England, 1995.

[11] Y. Saad. *Iterative Methods for Sparse Linear Systems Second Edition.* Society for Industrial and Applied Mathematics, 2003.

[12] H. A. van der Vorst. *Iterative Krylov Methods for Large Linear systems.* Cambridge University Press, 2003.

Modelling and Numerics for the Transient Simulation of the Blast Furnace Process

Heinz W. Engl
Industrial Mathematics Institute, Johannes Kepler Universität, 4040 Linz, Austria.
E-mail: heinz.engl@jku.at

Gerald Goekler
MathConsult GmbH
4040 Linz, Austria.
E-mail: goekler@mathconsult.co.at

Andrea Schatz
MathConsult GmbH
4040 Linz, Austria.
E-mail: schatz@mathconsult.co.at

Helmut Zeisel
Siemens-VAI / ATI 3
4031 Linz, Austria.
E-mail: helmut.zeisel@siemens.com

Abstract

The blast furnace process represents a way to produce hot metal from iron ore. In a joint work of Industrial Mathematics Competence Center (IMCC) and Siemens VAI, we have developed a transient mathematical model of the blast furnace with the aim to calculate the form and position of the coke / ore layers and of the cohesive zone. Further, the model should be capable of predicting the amount of reducing agents needed to produce hot metal with some prescribed temperature.

In this paper, we present the mathematical model as well as the numerical methods we use and compare the results to the measurements of a real blast furnace.

1 Introduction—a short description of the blast furnace process

A blast furnace is a vertical shaft (see Fig. 1.1) with the principal objective to produce molten iron, which requires a temperature of iron and slag greater than 1400 [°C]. The raw materials in the blast furnace are

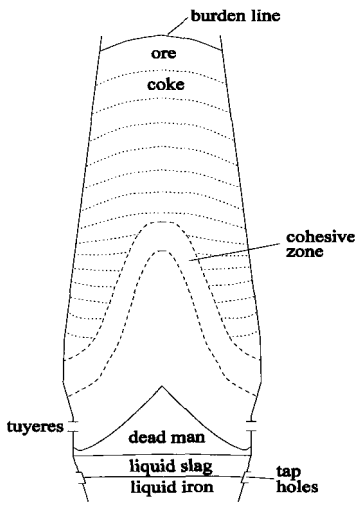


Figure 1.1 A section of a iron blast furnace

solids (ore, coke, flux) charged in layers into the top and blast (preheated air with additions), which is blown in through tuyeres near the bottom of the furnace and used for combustion of coke.

Coke is the universal reducing agent used in the blast furnace which acts both as a reductant as well as a supplier of heat.

Above the cohesive zone, the charged material consisting of alternate layers of coke and ore, travels downwards and is pre-reduced by the ascending gas components CO and H_2 (indirect reduction). Additional chemical reactions in this region include the vaporization of moisture and the decomposition of carbonates.

The cohesive zone, where the softening of the iron ore starts, has the shape of a W or of an inverse V and extends from the middle of the shaft to 3-5 m above tuyere level. The permeable coke layers act as gas distributors and permit the gas to flow horizontally through them. Below the cohesive zone, almost all of solid material consists of coke. Liquid iron and slag percolate through this coke. During this percolation reduction is finalized and the iron becomes saturated with carbon. Beyond the raceway, there is a closely packed central column of coke called dead

man.

The main product, molten iron, is tapped from the furnace at regular intervals through holes near the bottom of the hearth. By-products are the tapped slag and gas which leaves through the gas collection system at the top.

2 The mathematical model

Mathematical models, describing the blast furnace process can be found in the literature (see e.g. [Zei95], [Yag93], [ANY97b], [ANY97a], [ZAJY00a], [ZAJY00b]).

The mathematical model and numerical simulation of the blast furnace presented in this paper started from partially new requirements and modelling goals. It was carried out in cooperation between Siemens VAI, voestalpine, and the Industrial Mathematics Competence Center.

2.1 Modelling goals

Packing important process parts of the blast furnace into certain modules yields:

1. **Flows of different phases**
   granules (ore, coke, additives), liquids, gas
2. **Material composition**
   process specific materials (phase components), chemical reactions
3. **Temperature distribution**
   temperature fields of all phases

In the mathematical model, each of these (coupled) modules is based on the balance of mass, momentum or enthalpy. Simulations done by Yagi et al. (see [dCNY02]) show that the large number of tuyeres in a ring-shaped arrangement at an approximate height of 5 m above the furnace bottom allows an axisymmetric modelling of the process. Important project goals, which should be a result of the simulation are the (mainly) time dependent calculation of

1. gas flow distribution
2. burden flow - form and position of charged coke and ore layers from burden surface to the cohesive zone
3. form of the cohesive zone and the region below, occupied by the liquid phase
4. form of the dead man (a zone of inactive coke at the bottom of the blast furnace)

Therefore, the mathematical model has to be formulated in a transient way.

The model should be used both as an offline-tool as well as an online-tool. Offline-simulation allows e.g. a better understanding of the process, the prediction of the fuel consumption rate, and simulation of the burden distribution to achieve a uniform gas mass flow through the furnace; online-simulation allows control of the blast furnace in a way to achieve a stable operation of the process and a high and uniform quality of the hot metal. In particular for the online control the mathematical model has to be formulated in a transient way. In the following subsections, the mathematical model is presented in some detail. The most important aspect of our model, the calculation of the position of the ore and coke layers, can be found in Sec. 2.5.2.

2.2 Gas flow

The empirical Ergun relation (see [Erg52]), describing the interaction between gas and packed particles, is commonly used for modelling the flow of gas through particle filled reactors ([Yag93], [ANY97a], [ZAJY00b]). Together with the mass balance, a system of equations for velocity and pressure is determined.

The unknowns of the gas flow module are the velocity of the gas $\vec{v}^{(g)}$ and the gas pressure $p^{(g)}$. The underlying equations are

$$\frac{\partial(\varepsilon^{(g)}\rho^{(g)})}{\partial t} + \frac{1}{r}\nabla.\left(r\varepsilon^{(g)}\rho^{(g)}\vec{v}^{(g)}\right) = q^{(g)} \tag{2.1}$$

and the classical Ergun equation for flow through a porous medium (cf. [Erg52]):

$$-\nabla p^{(g)} = \left(f^{(v)} + f^{(i)}\|\varepsilon^{(g)}\rho^{(g)}\left(\vec{v}^{(g)} - \vec{v}^{(s)}\right)\|\right)\varepsilon^{(g)}\rho^{(g)}\left(\vec{v}^{(g)} - \vec{v}^{(s)}\right) \tag{2.2}$$

with ([RJ70])

$$f^{(v)} := 150\frac{(1-\varepsilon^{(g)})^2\mu^{()}}{(\varepsilon^{(g)})^3\rho^{(g)}(\Phi d_p)^2}, f^{(i)} := 1.75\frac{1-\varepsilon^{(g)}}{(\varepsilon^{(g)})^3\rho^{(g)}(\Phi d_p)}, \tag{2.3}$$

$\varepsilon^{(g)}$ denotes the volume fraction, $\rho^{(g)}$ the density and $q^{(g)}$ the mass sources (from chemical reactions). $\vec{v}^{(s)}$ is the solid velocity, $\mu^{(g)}$ is the dynamic gas viscosity and Φd_p the averaged particle diameter multiplied by a weighting factor $\Phi > 0$ for non-spherical particles. The boundary conditions for the gas flow are a prescribed gas pressure at the top and prescribed mass flow at the tuyeres.

For the calculation of the form of the cohesive zone, the softening of the ore layers at temperatures above 1200 [°C] is included by decreasing

the permeability of the ore layers in the model by a modification of the parameters in the Ergun equation (i.e. an increase of the bed resistance).

2.3 Solid flow — a modified Bingham material approach

Modelling the flow of the charged materials takes advantage of the material law based on a constitutive model for Bingham plastics, developed in the PhD thesis [Sch00], which allows the calculation of the form of the dead man with a special frictional viscosity.

The mass balance equation

$$\frac{\partial(\varepsilon^{(s)}\rho^{(s)})}{\partial t} + \frac{1}{r}\nabla.\left(r\varepsilon^{(s)}\rho^{(s)}\vec{v}^{(s)}\right) = q^{(s)} \tag{2.4}$$

and the momentum balance,

$$\begin{aligned}
\varepsilon\rho\frac{\partial v_r}{\partial t} + \varepsilon\rho v\nabla v_r &= \frac{\partial(\varepsilon p)}{\partial r} - \frac{\partial}{\partial r}(\varepsilon(\mu_0 - \frac{1}{3}\kappa)\nabla.v) - \frac{\partial}{\partial r}(\varepsilon\kappa\frac{\partial v_r}{\partial r}) \\
&\quad -\frac{\partial}{\partial z}(\frac{1}{2}\varepsilon\kappa(\frac{\partial v_r}{\partial z} + \frac{\partial v_z}{\partial r})) - \varepsilon\kappa\frac{\partial v_r}{\partial r}\frac{1}{r} + \varepsilon v\frac{v_r}{r^2} \\
\varepsilon\rho\frac{\partial v_z}{\partial t} + \varepsilon\rho v\nabla v_z &= \frac{\partial\varepsilon p}{\partial z} - \frac{\partial}{\partial z}(\varepsilon(\mu_0 - \frac{1}{3}\kappa)\nabla.v) - \frac{\partial}{\partial z}(\varepsilon\kappa\frac{\partial v_z}{\partial z}) \\
&\quad -\frac{\partial}{\partial r}(\frac{1}{2}\varepsilon\kappa(\frac{\partial v_r}{\partial z} + \frac{\partial v_z}{\partial z})) - \varepsilon\rho g
\end{aligned} \tag{2.5}$$

form the system for the calculation of the solid pressure $p^{(s)}(t,r,z)$ and the velocity $\vec{v}^{(s)}(t,r,z)$. The stress tensor σ is given by

$$\sigma = -(p^{(s)} - \mu_0 tr(D))I + \kappa \operatorname{dev} D, \tag{2.6}$$

with

$$\kappa = (\frac{\eta(1 - \exp(-n\frac{\|\operatorname{dev} D\|}{\eta}))}{\|\operatorname{dev} D\|} + 2\mu_1), \tag{2.7}$$

μ_0 and μ_1 denote the bulk and the shear viscosity. D denotes the rate of deformation tensor, $\operatorname{dev} D := D - \frac{\operatorname{tr}(D)}{3}I$ and $\|A\| := \sqrt{\operatorname{tr}(A^T A)/2)}$. η is a variable derived from a Drucker-Prager yield criterion (for details see [Sch00] Sec. 3.4.)

At the furnace top, the solid pressure and the direction of the solid velocity are prescribed. At the walls, the wall friction angle is given. At the tuyeres, a velocity profile and the mass flow are prescribed.

2.4 Modelling of liquid flow

In our model, the cohesive zone is not assumed to be a priori known, but is a result of the calculation. The form and position of the cohesive zone is dependent on softening and melting temperatures as well as the interdependency between the gas flow and temperature distribution.

The source of the liquid phase is given by temperature dependent melting rates. In order to calculate a liquid flow field below the cohesive zone, we use the mass balance equation

$$\frac{1}{r}\nabla.\left(r\rho_{typ}^{(l)}\vec{v}^{(l)}\right) = 0 \tag{2.8}$$

and the Navier Stokes equation with a constant typical density $\rho_{typ}^{(l)}$ and viscosity κ. The boundary conditions are the atmospheric pressure at the furnace top, the normal velocity of the liquid phase at the bottom (pointing into the hearth) and no normal flow through the walls. These equations form a system for the calculation of a pressure field $p^{(l)}(t,r,z)$ and a velocity $\vec{v}^{(l)}(t,r,z)$ of a liquid.$\vec{v}^{(l)}$ is calculated once and serves as a prescribed velocity field for the liquid phase. This is a simplifying approach, but is sufficient for our needs. The region occupied by the liquid phase is a priori unknown, the solution of the liquid mass balance with mass sources (see sec.2.5) yields the unknown liquid volume fraction $\varepsilon^{(l)}$.

2.5 Chemistry and temperature

The model of the chemistry and temperature module consists of balance equations for all components of all phases, enthalpy balance equations for all temperature fields and reaction rate models for 15 (significant) chemical processes.

2.5.1 Gas composition

The gas phase consists of $n^{(g)}$ components $\mathcal{V}^{(g)} := \{\nu_1^{(g)},\ldots,\nu_{n^{(g)}}^{(g)}\}$ with mass fractions $\vec{\xi}^{(g)} := (\xi_1^{(g)},\ldots,\xi_{n^{(g)}}^{(g)})$. The mass balance equation for the component $\nu_i^{(g)}$in the gas phase has the transient axisymmetric form for $\xi_i^{(g)}(t,r,z)$

$$\frac{\partial(\xi_i^{(g)}\,\varepsilon^{(g)}\,\rho^{(g)})}{\partial t} + \frac{1}{r}\nabla.\left(r\xi_i^{(g)}\vec{G}^{(g)}\right) = q_i^{(g)} \tag{2.9}$$

with the effects of mass transport due to convection (the small diffusion is omitted). $\varepsilon^{(g)}(:= 1-\varepsilon^{(s)}-\varepsilon^{(l)})$ and $\rho^{(g)}$ denote the volume fraction and density, $q_i^{(g)}$ is the mass source for the substance $\nu_i^{(g)}$ and $\vec{G}^{(g)} :=$

$\varepsilon^{(g)}\,\rho^{(g)}\,\vec{v}^{(g)}$ denotes the mass flow density. Boundary conditions are prescribed gas composition at the tuyeres.

2.5.2 Solid material composition

Each mixture of raw materials in a layer is combined in the model into one solid phase: an ore layer consists of a mix of sinter, pellets and lump ore, a coke layer of different sorts of coke.

Let $\mathcal{P}$ denote this set of different sub-phases (layers), where each member $p \in \mathcal{P}$ consists of $n^{(p)}$ components $\{\nu_1^{(p)}, \ldots, \nu_{n^{(p)}}^{(p)}\} =: \mathcal{V}^{(p)}$ with the vector of mass fractions $\vec{\xi}^{(p)} := (\xi_1^{(p)}, \ldots, \xi_{n^{(p)}}^{(p)})$.

The mass balance equation for a substance $\nu_i^{(p)}$ with mass fraction $\xi_i^{(p)}$ in a sub-phase p is given by

$$\frac{\partial(\xi_i^{(p)}\,\varepsilon^{(p)}\,\rho^{(p)})}{\partial t} + \frac{1}{r}\nabla\cdot\left(r\xi_i^{(p)}\vec{G}^{(p)}\right) = q_i^{(p)} \tag{2.10}$$

modelling the convective mass transport. $\varepsilon^{(p)}$ and $\rho^{(p)}$ denote the volume fraction and density, $q_i^{(p)}$ are the mass sources for the substance ν_i in the phase p, and $\vec{G}^{(p)} := \varepsilon^{(p)}\,\rho^{(p)}\,\vec{v}^{(s)}$ is the mass flow density, with $\vec{v}^{(s)}$ denoting the solid velocity field (for all sub-phases).

The boundary conditions reflect the layer-by-layer charging of the blast furnace process. The initial distribution of coke and ore layer is an important model input and is the reason for the need for a transient formulation. Figure 3.1 visualizes a typical mix of coke and ore layers. Such a mix is described by a material matrix M consisting of disjoint material layers $\mathrm{l}(n;p)$. A material matrix $\mathcal{M}$ consists of disjoint material

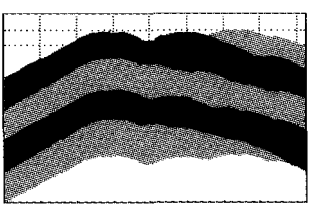


Figure 2.1 Example of a blast furnace material matrix of iron carriers and coke

layers $l_{(n,p)}$, where the number n denotes the charging order and the phase $p \in \mathcal{P}$ denotes the material the respective layer consists of. In each point (r_m, z_m) of $\mathcal{M}$, one phase $p^{\mathcal{M}}(r_m, z_m) \in \mathcal{P}$ is located, with the prescribed volume fraction $\varepsilon^{(p^{\mathcal{M}})}$ and composition $\vec{\xi}^{(p^{\mathcal{M}})}$. The next layer of material is charged when the height of the material stack falls below a specified value H_{charge}.

The transient formulation requires initial conditions i.e. each point $(r, z) \in \Omega$ is filled with some phase p_{start}.

2.5.3 Liquid composition

The mass balance equation for the liquid phase

$$\frac{\partial(\varepsilon^{(l)}\,\rho^{(l)})}{\partial t} + \frac{1}{r}\nabla.\left(r\vec{G}^{(l)}\right) = q^{(l)} \tag{2.11}$$

is used for the calculation of $\varepsilon^{(l)}$. $q^{(l)}$ denotes the liquid mass sources which result from melting. The melting is formulated as an (artificial) chemical reaction (see Table. 2.1), dependent on the melting point of the raw materials. $\vec{G}^{(l)}(:= \varepsilon^{(l)}\,\rho^{(l)}\,\vec{v}^{(l)})$ denotes the liquid mass flow density and $\vec{v}^{(l)}$ denotes the liquid velocity field (see Sec.2.4).

The mass balance equation for a substance $\nu_i^{(l)}$ with mass fraction $\xi_i^{(l)}$ in the liquid phase for $\xi_i^{(l)}(t, r, z)$ is given by

$$\frac{\partial(\xi_i^{(l)}\,\varepsilon^{(l)}\,\rho^{(l)})}{\partial t} + \frac{1}{r}\nabla.\left(r\xi_i^{(l)}\vec{G}^{(l)}\right) = q_i^{(l)}, \tag{2.12}$$

$q_i^{(l)}$ are the mass sources for the substance ν_i. The domain occupied by the liquid phase is time dependent. The boundary of this region is defined by the points of the domain, where melting starts. In these points, $\varepsilon^{(l)}$ and $\xi_i^{(l)}$ are prescribed to be zero.

2.5.4 Temperature distribution of gas, solids and liquid

For the calculation of the gas, solid and liquid temperature fields $T^{(.)}$((.) :=(g),(s) or (l)) the enthalpy balances are formulated taking into account heat diffusion, convection, the strong coupling due to interphase heat exchange between the gaseous and the solid phase and the heat sources due to chemical reactions.

The enthalpy balances have the transient axisymmetric form

$$\begin{aligned}\frac{\partial\varepsilon^{(.)}\rho^{(.)}h^{(.)}}{\partial t} &+ \varepsilon^{(.)}\rho^{(.)}\vec{v}^{(.)}c_p^{(.)}\nabla T^{(.)} - \nabla.\left(\varepsilon^{(.)}\lambda^{(.)}\nabla T^{(.)}\right)\\ &= \sum_{(p)=(s),(g),(l)} H^{(.,p)}\left(T^{(p)} - T^{(.)}\right) + S^{(.)}\end{aligned} \tag{2.13}$$

with$\nabla.(\lambda\nabla T) = \frac{\partial}{\partial r}(\lambda\frac{\partial T}{\partial r}) + \frac{\lambda}{r}\frac{\partial T}{\partial r} + \frac{\partial}{\partial z}(\lambda\frac{\partial T}{\partial z})$.

$h^{(.)}$ denotes the specific enthalpy of the corresponding phase, $\lambda^{(.)}$ the effective heat conduction coefficient, $H^{(.,p)}$ the interphase heat exchange coefficient and $S^{(.)}$ the heat sources from chemical reactions. The gas temperature at the gas inflow boundaries is prescribed.

When a new layer is charged, it has a prescribed initial temperature $T_{in}^{(s)}$. The heat loss through the walls is modelled with a Robin boundary condition $\lambda \frac{\partial T^{(s)}}{\partial n} = H_{\Gamma_i}^{out}(T^{(s)} - T^{out})$where $H_{\Gamma_i}^{out}$ denotes the known heat flow through the wall. The boundary condition for the liquid phase is given by the melting point of the ore layers. The transient formulation requires an initial temperature fields $T^{(\cdot)}(0,x) = T_{start}^{(\cdot)}$ for all $x \in \Omega$.

2.5.5 Chemical reactions — Mass and heat sources

Table 2.1 is a listing of important heterogenous reactions in the reduction shaft and the blast furnace. Heterogenous reactions between gas and

Table 2.1 Collection of considered reactions

Type	Reaction stoichiometry
Reduction	$3\ Fe_2O_3 + CO\ /\ H_2 \rightarrow 2\ Fe_3O_4 + CO_2\ /\ H_2O$ $Fe_3O_4 + CO\ /\ H_2 \rightarrow 3\ FeO + CO_2\ /\ H_2O$ $FeO + CO\ /\ H_2 \rightarrow Fe + CO_2\ /\ H_2O$
Drying	$H_2O(l) \rightarrow H_2O$
Boudouard	$CO_2 + C \leftrightarrow 2\ CO$
Watergas	$C + H_2O \leftrightarrow CO + H_2$
Melting	$.(s) \rightarrow .\ (l)$
Dir. Reduction	$FeO\ (l) + C \rightarrow Fe\ (l) + CO$
Carburization	$3\ Fe\ (l) + C \rightarrow Fe_3C(l)$

solid particles can be modelled by the "shrinking core" model where the reaction rate is determined by the diffusion of the gaseous reactant through the gas boundary layer, the diffusion through the product layer and the chemical reaction with the solid reactant (for details see [FFE95], [oIR87]).

The general reaction rate model for homogenous reactions uses the Arrhenius rate law, which is an empirical expression for describing the temperature dependence of rate constants $k(T)$(see e.g. [FFE95]). The empirical melting rates are formulated so as to cause an immediate phase transition when the melting point is reached.

All reaction rate models yield nonlinear source terms for the component mass balances for all components $\nu_i^{(\cdot)}$ of a general form

$$q_i^{(\cdot)} = \lambda(T^{(g)}, T^{(s)}, T^{(l)}, \vec{\xi}^{(g)}, \vec{\xi}^{(s)}, \vec{\xi}^{(l)}) f(\xi_i). \tag{2.14}$$

The mass sources for all phases and all phase components as well as all heat sources for the enthalpy balances, are determined from these reaction rate models.

3 Numerics and results

3.1 Overview

The modular structure of the mathematical model forms the basis of the numerical solution.

A schematic overview of the modules and the numerical algorithms and methods which are used for the modules is given by:

module	math. model	numerics	variables
gas flow	Sec.2.2	Sec. 3.2	$p^{(g)}, v^{(g)}$
solid flow	Sec.2.3	[Sch00]	$p^{(s)}, v^{(s)}$
liquid flow	Sec.2.4	[Sch00]	$v^{(l)}$
chemistry, temperature	Sec. 2.5	Sec. 3.3	$\vec{\xi}^{(g)}, \vec{\xi}^{(p)}, \vec{\xi}^{(l)}, \varepsilon^{(l)}, T^{(g)}, T^{(s)}, T^{(l)}$

The numerical algorithm for the solid and liquid flows is a finite element method where artificial diffusion is added in streamline direction to avoid non-physical oscillations which may occur in case of dominant convection. A description of this numerical method can be found in [Sch00], Sec. 4.2.

3.2 Numerics of the gas flow

The blast furnace model is time dependent, but due to the fact that the solid layers are moving slowly in the gas time scale, the time dependence can be omitted and a steady state solution is calculated for each time step in the solid time scale.
Substituting (2.2) into (2.1) yields a non-linear second order partial differential equation for the gas pressure $p^{(g)}$.

$$-\nabla.(\frac{\nabla p^{(g)}}{f^{(v)} + f^{(i)}\|G^{(g)}\|}) = q^{(g)} \tag{3.1}$$

with $G^{(g)} := \varepsilon^{(g)}\rho^{(g)}\vec{v}^{(g)}$. This equation is solved by successive substitution of the gas mass flow $G_k^{(g)}$ of the preceding step. With $c_k := f_k^{(v)} + f_k^{(i)}\|G_k^{(g)}\|$, the system for the $(k+1)$-th (inner) iteration then is the linear equation

$$-\nabla.\left(\frac{1}{c_k}\nabla \bar{p}_{k+1}^{(g)}\right) = q_K^{(g)}. \tag{3.2}$$

Equation (3.2) is solved using a finite element method with bilinear test and basis functions for the weak formulation. For the calculation of c_k, $G_k^{(g)}$ has to be determined:

The Ergun equation $-\nabla p^{(g)} = (f^{(v)} + f^{(i)}\|\vec{G}\|)\vec{G}$ (which holds for the finally converged solution) motivates to assume that for the k-th iteration

$$\|\nabla p_k^{(g)}\| = f_k^{(v)}\|G_k^{(g)}\| + f_k^{(i)}\|G_k^{(g)}\|^2 \tag{3.3}$$

holds. Solving this quadratic equation for $\|G_k^{(g)}\|$ leads to

$$G_k^{(g)} = -\frac{1}{\omega(\|\nabla p_k^{(g)}\|)}\nabla p_k^{(g)}$$

with

$$\omega(\|\nabla p_k^{(g)}\|) = \|G_k^{(g)}\| = f_k^{(v)}/2 + \sqrt{(f_k^{(v)}/2)^2 + f_k^{(i)}\|\nabla p_k^{(g)}\|}.$$

3.3 Chemistry / temperature–an extended method of characteristics

In the blast furnace, hundreds of material (coke / ore) layers are transported by a smooth velocity field, not causing any major changes of the form of the interfaces (before softening and melting starts). The discontinuities (in space) in the composition and temperature, caused by these layers, have to be handled by a numerical algorithm for the solution of the transient balance equations.

3.3.1 Method of characteristics

The component mass balance equations are convection-reaction equations. The method of characteristics is suited for solution of such equations. It requires the solution of the equation for the characteristics and the integration of ODEs along these characteristics, which are streamlines of the velocity field.

Starting from a gas (omitting diffusion), solid (material layer consisting of a phase $p \in \mathcal{P}$) or liquid component mass balance with $\xi := \xi_\nu^{(.)}$, $\rho := \varepsilon^{(.)}\rho^{(.)}$, $q_\nu^{(.)} = q$ and $\vec{v} := (v_r, v_z)^{(.)}$,$((.) := (g), (p)_{p\in\mathcal{P}}, (l))$

$$\frac{\partial \xi\rho}{\partial t} + \frac{\partial \xi\rho v_r}{\partial r} + \frac{\xi\rho v_r}{r} + \frac{\partial \xi\rho v_z}{\partial z} = q \tag{3.4}$$

i.e.

$$\rho\frac{\partial \xi}{\partial t} + \rho v_r\frac{\partial \xi}{\partial r} + \rho v_z\frac{\partial \xi}{\partial z} + \xi\underbrace{(\frac{\partial \rho}{\partial t} + \frac{\partial \rho v_r}{\partial r} + \frac{\rho v_r}{r} + \frac{\partial \rho v_z}{\partial z})}_{Q} = q \tag{3.5}$$

yields the ODE

$$\frac{D\xi}{Dt} = \frac{q - \xi Q}{\rho} \tag{3.6}$$

with $\frac{D.}{Dt} := \frac{\partial.}{\partial t} + v_r \frac{\partial.}{\partial r} + v_z \frac{\partial.}{\partial z}$ denoting the total derivative.

This equation expresses the change of the mass fraction ξ for each component ν, while moving along a streamline.

Due to the very strong coupling between all balance equations (via the non-linear reaction rates and large heat exchange of phases), it is the aim to approximate all mass and enthalpy balance equations in all crossing points $(\bar{r}, \bar{z}) \in \Omega$ of the gas, solid and liquid streamlines in a coupled way. For this purpose, all (gas, solid, liquid) streamlines (or characteristic curves) $c^{(.)}$ crossing a point $(\bar{r}, \bar{z})$ at some time t_n are calculated. The characteristic curve $c^{(.)}(t) := (r^{(.)}(t), z^{(.)}(t))$, departing

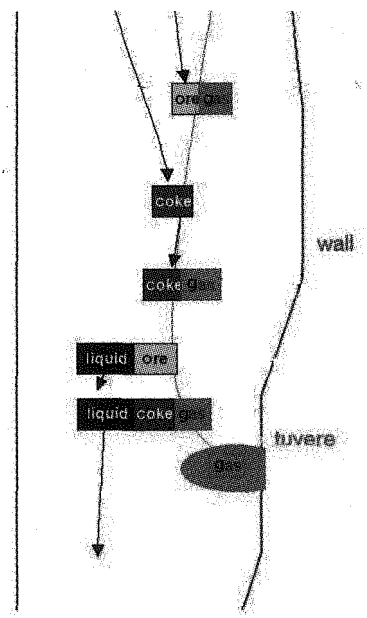


Figure 3.1 The simulation yields crossing of gas/ore, gas/coke, gas/ore/liquid and gas/coke/liquid characteristics

at t_{n-1} or arriving at t_n at a certain point $(\bar{r}, \bar{z})$ is given by the solution of the ODEs

$$\frac{\partial r^{(.)}}{\partial t}(r^{(.)}(t), z^{(.)}(t)) = v_r^{(.)}(r^{(.)}, z^{(.)}(t)), \tag{3.7}$$

$$\frac{\partial z^{(.)}}{\partial t}(r^{(.)}(t), z^{(.)}(t)) = v_z^{(.)}(r^{(.)}(t), z^{(.)}(t)) \quad t \in [t_{n-1}, t_n] \tag{3.8}$$

with final value $(\bar{r}, \bar{z})$ (back tracking!).

The method of characteristics, formulated for all component mass balances and enthalpy balances now consists of three steps:

1. The streamlines $c^{(g)}, c^{(s)}, c^{(l)}$ ending in $(\bar{r}, \bar{z}) \in \Omega$ at t_n are calculated backwards with

$$c^{(\cdot)}(t) := \begin{pmatrix} r^{(\cdot)}(t) \\ z^{(\cdot)}(t) \end{pmatrix} = \begin{matrix} \bar{r} - \int_t^{t_n} v_r^{(\cdot)}(r^{(\cdot)}(\tau), z^{(\cdot)}(\tau))d\tau \\ \bar{z} - \int_t^{t_n} v_z^{(\cdot)}(r^{(\cdot)}(\tau), z^{(\cdot)}(\tau))d\tau \end{matrix} \tag{3.9}$$

   yielding the starting positions $(\bar{r}_0^{(g)}, \bar{z}_0^{(g)}), (\bar{r}_0^{(s)}, \bar{z}_0^{(s)}), (\bar{r}_0^{(l)}, \bar{z}_0^{(l)})$ at time t_{n-1}(i.e. $c^{(\cdot)}(t_{n-1}) = (\bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$).

2. Determination of the composition and temperature of the gas, solid and liquid phase at the starting points $\vec{\xi}^{(\cdot)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$, $T^{(\cdot)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$ and $\varepsilon^{(l)}(t_{n-1}, \bar{r}_0^{(l)}, \bar{z}_0^{(l)})$. Here, for the solid phase, the composition $\vec{\xi}^{(p)}(t_{n-1}, \bar{r}_0^{(s)}, \bar{z}_0^{(s)})$ of the layer, consisting of solid phase p located in $(\bar{r}_0^{(s)}, \bar{z}_0^{(s)})$ at t_{n-1}, is determined.

3. Forward integration of the system of ODEs defined by all $n^{(g)}$ gas, $n^{(p)}$ solid, $n^{(l)}$ liquid component mass balances along the streamlines $c^{(g)}, c^{(s)}, c^{(l)}$,

$$\begin{aligned} \xi_1^{(\cdot)}(t_n, \bar{r}, \bar{z}) &= \xi_1^{(\cdot)}(t_{n-1}, \bar{r}_0, \bar{z}_0) \\ &\quad + \int_{t_{n-1}}^{t_n} \frac{q_1^{(\cdot)}(c^{(\cdot)}(t)) - \xi_1^{(\cdot)}(c^{(\cdot)}(t))q^{(\cdot)}(c^{(\cdot)}(t))}{\rho^{(\cdot)}(c^{(\cdot)}(t))} dt \\ &\vdots \\ \xi_{n^{(\cdot)}}^{(\cdot)}(t_n, \bar{r}, \bar{z}) &= \xi_{n^{(\cdot)}}^{(\cdot)}(t_{n-1}, \bar{r}_0, \bar{z}_0) \\ &\quad + \int_{t_{n-1}}^{t_n} \frac{q_{n^{(\cdot)}}^{(\cdot)}(c^{(\cdot)}(t)) - \xi_{n^{(\cdot)}}^{(\cdot)}(c^{(\cdot)}(t))q^{(\cdot)}(c^{(\cdot)}(t))}{\rho^{(\cdot)}(c^{(\cdot)}(t))} dt. \end{aligned} \tag{3.10}$$

The formulation of the (gas, solid, liquid) enthalpy balance equations (omitting the heat conduction) as an ODE yields

$$\frac{DT^{(\cdot)}}{Dt} = \frac{S^{(\cdot)} + \sum_{(p)=(g),(s),(l)} H^{(\cdot,p)}(T^{(\cdot)} - T^{(p)})}{\varepsilon^{(\cdot)}\rho^{(\cdot)}c_p^{(\cdot)}}.$$

The temperatures $T^{(g)}, T^{(s)}, T^{(l)}$ at time t_n in $(\bar{r}, \bar{z})$ are then calculated by

$$T^{(\cdot)}(t_n, \bar{r}, \bar{z}) = T^{(\cdot)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)}) + \int_{t_{n-1}}^{t_n} \tag{3.11}$$

$$\frac{S^{(\cdot)}(t, c^{(\cdot)}(t)) + \sum_{(p)=(g),(s),(l)} H^{(\cdot,p)}(t_n, c^{((\cdot)}(t))(T^{(\cdot)}(t, c(t)) - T^{(p)}(t_n, c^{(\cdot)}(t))}{\varepsilon^{(\cdot)}(t, c^{(\cdot)}(t))\rho^{(\cdot)}(t, c(t))c_p^{(\cdot)}(t, c^{(\cdot)}(t))} dt$$

Under the assumption of a constant liquid density, the reformulation of the liquid mass balance yields $\frac{D\varepsilon^{(l)}}{Dt} = \frac{q^{(l)}}{\rho^{(l)}} - \varepsilon^{(l)}\mathrm{div}(\vec{v^{(l)}})$. Due to $\mathrm{div}(\vec{v^{(l)}}) = 0$ (see sec. 2.4), $\varepsilon^{(l)}(t_n, \bar{r}, \bar{z})$ is calculated by

$$\varepsilon^{(l)}(t_n, \bar{r}, \bar{z}) = \varepsilon^{(l)}(t_{n-1}, \bar{r}_0^{(l)}, \bar{z}_0^{(l)}) + \int_{t_{n-1}}^{t_n} \frac{q^{(l)}(t, c^{(l)}(t))}{\rho^{(l)}(t, c^{(l)}(t))} dt. \quad (3.12)$$

3.3.2 Incorporation of different time scales

In the previous section, the method of characteristics formulated for the solution of all balance equations of all characteristics crossing all points of Ω has been formulated. The different time scales of the various balance equations need a more detailed examination.

The movement speed of burden is 3-5 [m/h], the typical gas velocity is 3-6 [m/s]. The coupled solution of systems in such different time scales can be approximated by the calculation of the steady state solution of the fast changing (gas) system for each time step of the slowly changing (solid, liquid) system.

This approach especially requires the knowledge of the starting point of a gas particle passing an arbitrary point $(\bar{r}, \bar{z})$ of the domain. The resulting system of ODEs for the steady state gas composition and temperature as an approximation at (burden) time t_n at a point $(\bar{r}, \bar{z})$ is then determined as follows

1. Let $c^{(g)}$ denotes the gas characteristic, starting at some point of the gas inlet $(r_0^{(g)}, z_0^{(g)})$ and passing $(\bar{r}, \bar{z})$ after time $t^{(g)}$

$$\begin{aligned} \bar{r} &= r_0^{(g)} + \int_0^{t^{(g)}} v_r^{(g)}(t_n, c^{(g)}(t))dt, \\ \bar{z} &= z_0^{(g)} + \int_0^{t^{(g)}} v_z^{(g)}(t_n, c^{(g)}(t))dt. \end{aligned} \quad (3.13)$$

2. The initial conditions to be set at the tuyere are the composition $\xi_{in}^{(g)}$ and temperature $T_{in}^{(g)}$ of gas.

3. The mass fraction $\xi_i^{(g)}$ of component $\nu_i^{(g)}$ at time t_n in $(\bar{r}, \bar{z})$ is calculated by

$$\xi_i^{(g)}(t_n, \bar{r}, \bar{z}) = \xi_{i,in}^{(g)} + \int_0^{t^{(g)}} \frac{q_i^{(g)}(t_n, c^{(g)}(t)) - \xi_i^{(g)}(t_n, c^{(g)}(t))q^{(g)}(t_n, c^{(g)}(t))}{\rho^{(g)}(t_n, c^{(g)}(t))} dt. \quad (3.14)$$

The equation for the gas temperature is reformulated in the same manner.

3.3.3 Fast and accurate approximation of balance equations — discretization of domain and equations

For the overall discretization Ω_h of the blast furnace geometry, we use a general quadrilateral grid. Then, we construct a structured sub-grid Ω_h^s, which has to be linked to the grid Ω_h, as follows (see Fig. 3.2):

The cartesian sub-grid denoted by Ω_h^s consists of $n_r \times n_z$ squares b_{ij} with edge length l_s. With $i \in [1, n_r]$ and $j \in [1, n_z]$, the barycenter of b_{ij} has coordinates $(l^s(i-1/2), l^s(j-1/2))$. All squares with barycenter outside Ω are excluded. Ω_h^s has the advantage of a fast determination of the square b_{ij}, containing an arbitrary point $\in \Omega_h$.

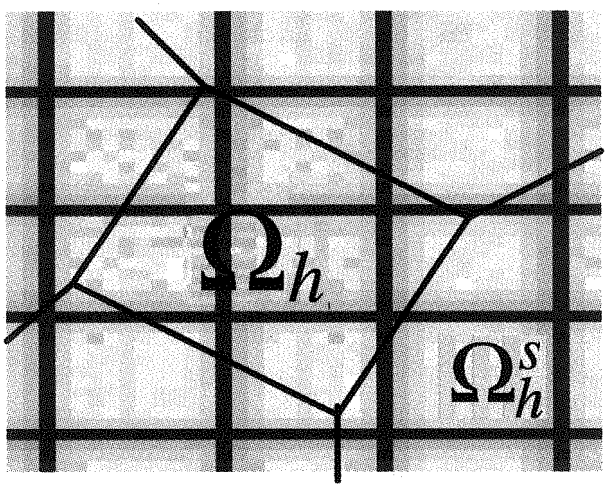


Figure 3.2 The discretized domain Ω_h and the sub-grid Ω_h^s

The discretized balance equations of sections 3.3.1 and 3.3.2 are now solved on Ω_h^s.

1. The characteristics equations in a sub-grid point b_{ij} with coordinates (r_{ij}, z_{ij}) in a given solid and liquid time interval $\Delta t := t_n - t_{n-1}$ (typically 100-500 [s]) or gas interval $\Delta t := \Delta t^{(g)} < t^{(g)}$ (typically < 0.2[s], see Sec. 3.3.4) is solved by the explicit Euler scheme:

$$\begin{pmatrix} r_0^{(.)} \\ z_0^{(.)} \end{pmatrix} = \begin{pmatrix} r_{ij} \\ z_{ij} \end{pmatrix} - \Delta t \vec{v}^{(.)}(r_{ij}, z_{ij}). \tag{3.15}$$

In a non-constant velocity field with $\vec{v}^{(.)}(t_n, r_{ij}, z_{ij}) \neq \vec{v}^{(.)}(t_{n-1}, r_0^{(.)}, z_0^{(.)})$, an averaged velocity $\bar{v}^{(.)}$ "between" $\vec{v}^{(.)}(r_{ij}, z_{ij})$ and $\vec{v}^{(.)}$

$(r_0^{(\cdot)}, z_0^{(\cdot)})$ can be calculated by

$$\bar{v}_{ij}^{(\cdot)} = \begin{cases} v^{(\cdot)}(r_{ij}, z_{ij}), & v^{(\cdot)}(\widehat{r}, \widehat{z}) = v^{(\cdot)}(r_{ij}, z_{ij}) \\ \frac{v^{(\cdot)}(\widehat{r},\widehat{z}) - v^{(\cdot)}(r_{ij},z_{ij})}{\ln(v^{(\cdot)}(\widehat{r},\widehat{z}) - v^{(\cdot)}(r_{ij},z_{ij}))} & \text{else,} \end{cases} \tag{3.16}$$

where $(\widehat{r}, \widehat{z})$ denotes the coordinates of the nearest sub-cell of $(r_0^{(\cdot)}, z_0^{(\cdot)})$ (see Theorem 5.1 of [KP00]).

2. From the coordinates of the starting points $(\bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$, in the structured sub-grid, the initial compositions and temperatures can now be calculated fast by a bilinear interpolation in the square of 4 subcells in which the point $(\bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$ lies yielding $\vec{\xi}_0^{(\cdot)} := \vec{\xi}^{(\cdot)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)}), T_0^{(\cdot)} := T^{(\cdot)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$ and $\varepsilon_0^{(l)} := \varepsilon^{(l)}(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$.
Remark: Only the fraction of the 4 subcells, which have the same solid sub-phase $p(t_{n-1}, \bar{r}_0^{(\cdot)}, \bar{z}_0^{(\cdot)})$ are included in the bilinear interpolation. This approach preserves the layered structure of the coke and ore layers in the simulation.

3. Due to the large source terms, an implicit 1-step BDF method ([HG87], Sec. V.1 p. 247) is chosen for the approximation of the coupled systems of ODEs (3.14) and (3.10),(3.11),(3.12) for the calculation of $(\vec{\xi}^{(g)}, \vec{\xi}^{(p)}, \vec{\xi}^{(l)}, T^{(g)}, T^{(s)}, T^{(l)}, \varepsilon^{(l)})\ (t_n, b_{ij})$: $\dot{y}(t) = f(t, y)$ is approximated by $y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y_{n+1})$. Applying this formula yields for all gas, solid and liquid components and temperatures

$$\begin{aligned} \xi_{i,k+1}^{(\cdot)} &= \xi_{i,0}^{(\cdot)} + \Delta t \frac{q_{i,k+1}^{(\cdot)} - \xi_{i,k+1}^{(\cdot)} q_k^{(\cdot)}}{\varepsilon_k^{(\cdot)} \rho_k^{(\cdot)}} \\ T_{k+1}^{(\cdot)} &= T_0^{(\cdot)} + \Delta t \frac{S_{k+1}^{(\cdot)} + \sum_{p=g,l,s} H_k^{(\cdot,p)}(T_{k+1}^{(\cdot)} - T_{k+1}^{(p)})}{\varepsilon_k^{(\cdot)} \rho_k^{(\cdot)} c_{p_k}^{(\cdot)}} \\ \varepsilon_{k+1}^{(l)} &= \varepsilon_0^{(l)} + \Delta t \frac{q_k^{(l)}}{\rho_k^{(l)}}. \end{aligned} \tag{3.17}$$

The mass and heat sources, resulting from heterogenous and homogenous reactions are approximated by a first order Taylor approximation with numerical derivatives; e.g., this linearization of a component mass source $q_i^{(\cdot)}$ yields

$$\begin{aligned} q_{i,k+1}^{(\cdot)} = q_{i,k}^{(\cdot)} &+ \sum_{i=1}^{n^{(\cdot)}} \nabla_{\nu_i^{(\cdot)}} q_{i,k}^{(\cdot)} (\xi_{i,k+1}^{(\cdot)} - \xi_{i,k}^{(\cdot)}) \\ &+ \nabla_{T^{(g)}} q_{i,k}^{(\cdot)} (T_{k+1}^{(g)} - T_k^{(g)}) + \nabla_{T^{(s)}} q_{i,k}^{(\cdot)} (T_{k+1}^{(s)} - T_k^{(s)}). \end{aligned} \tag{3.18}$$

With the vector of unknowns $x_{k+1} := (\vec{\xi}^{(g)}_{k+1}, \vec{\xi}^{(p)}_{k+1}, \vec{\xi}^{(l)}_{k+1}, T^{(g)}_{k+1}, T^{(s)}_{k+1}, T^{(s)}_{k+1}, \varepsilon^{(l)}_{k+1})$ in a sub-grid point b_{ij} of Ω^h_s, the matrix A_k and right side b_k, including the source terms and derivatives from the above equations are assembled. Typical dimensions of A_k are in the range of 20-40. The resulting linear system

$$A_k x_{k+1} = b_k \tag{3.19}$$

is solved by Gaussian elimination. The iteration is stopped when for two subsequent iterations x_k and x_{k+1}, $\|x_{k+1} - x_k\|_\infty$ is below a given tolerance.

3.3.4 Determination of the gas steady state with global subgrid-loops

In this section, the algorithm for the calculation of the steady state solution in the fast changing gas time scale for each time step of the slowly changing solid-liquid system (see Sec. 3.3.2) is outlined.

We do not know the starting point (at the gas inlet) of any gas flow characteristic passing a sub-grid point. But with repeated (bottom to top = typical gas flow direction) loops over all sub-grid points, the steady state of the gas flow is assumed to be reached, when the maximum difference of gas composition and temperature of two subsequent global sub-grid loops is below a given tolerance:

set $\vec{X}_1 := (\vec{X}^{(g)}_1 / \vec{X}^{(s)}_1 / \vec{X}^{(l)}_1) := (\xi^{(g)}, T^{(g)} / \xi^{(p)}, T^{(s)} / \xi^{(l)}, T^{(s)}, \varepsilon^{(l)})_1 = (\xi^{(g)}, T^{(g)} / \xi^{(p)}, T^{(s)} / \xi^{(l)}, T^{(s)} \varepsilon^{(l)})(t_{n-1}) \forall\ b_{ij}$

while $(\|\vec{X}^{(g)}_{k+1} - \vec{X}^{(g)}_k\|_\infty > \epsilon)$ // no steady state reached

{

k = k+1

$for\ i = 1, \ldots n_z$

$for j = n_r, \ldots, 1$

$\Delta t^{(s)} = \Delta t^{(l)} = t_n - t_{n-1}$, $\Delta t^{(g)}(b_{ij}) = \frac{l_s}{\|\vec{v}^{(g)}\|}$

approximation of balance equations in b_{ij} (Sec.3.3.3)

update $X_{k+1}(b_{ij})$

}

set $(\xi^{(g)}, \xi^{(p)}, \xi^{(l)}, T^{(g)}, T^{(s)}, T^{(l)}, \varepsilon^{()l})(t_n) = \vec{X}_k$

The choice of $\Delta t^{(g)}(b_{ij}) = \frac{l_s}{\|\vec{v}^{(g)}\|}$ yields a backtracking distance of exactly one cell length l_s for each b_{ij}. This algorithm typically does not need more than 1-4 (bottom to top) loops over all b_{ij} to reach convergence.

3.3.5 Boundary conditions

For the setting of chemistry and temperature boundary conditions (= initial conditions in the method of characteristics) for the gas phase, all subcells b_{ij} of the sub-grid Ω_h^s with a distance smaller than l^s to the inlet boundary are marked as gas inlet cells. When the backtracked point $(r_0^{(g)}, z_0^{(g)})$ is contained in such an inlet cell, $(\vec{\xi}_0^{(g)}, T_0^{(g)})$ is defined by the boundary condition $(\vec{\xi}_{in}^{(g)}, T_{in}^{(g)})$.

For the setting of solid boundary conditions, the Siemens VAI Burden Distribution Model provides the initial states for the form, position and width of the coke and ore layers, using the knowledge of the charging characteristics of the blast furnace and the distribution of the charged materials (i.e. the material matrix of Sec. 2.5.2). The numerical procedure of the charging of connected material layers is characterized as follows:

Each layer of the material matrix $\mathcal{M}$ is projected onto the sub-grid Ω_h^s. The resulting discrete layers $l_{(n,p)}^h$ are a connected set of subcells $b_{ij} \in \Omega_h^s$ (Fig. 3.3 visualizes the layers $l_{(n,p)}$ and the projected layers $l_{(n,p)}^h$ projected onto Ω_h^s). During the simulation, when the height of

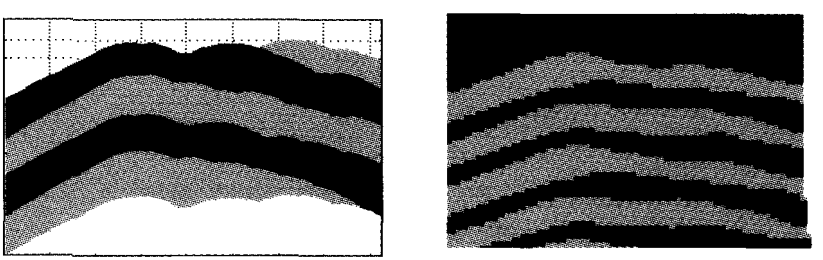


Figure 3.3 Material layers $l_{(n;p)}$ and resulting charged layers $l_{(h,p)}^h$ projected onto sub-grid computed by Burden Distribution Model

material stack falls below a prescribed value, the next layer $l_{(n,p)}^h$ is charged, i.e. the connected set of subcells defined by $l_{(n,p)}^h$ is filled with material p (and with initial values $\varepsilon_{in}^p, \xi_{in}^{(p)}, T_{in}^{(s)}$).

3.3.6 Incorporation of diffusion and wall heat loss

For the inclusion of wall heat loss, the operator splitting technique, presented in [KP00] is applied to the enthalpy balance of the solid phase:

First, starting from the temperature field at time t_{n-1}, $T^{(s)}(t_{n-1}, b_{ij})$

$\forall b_{ij} \in \Omega_h^s$, the convection-reaction part of the enthalpy balance (2.13)

$$\frac{\partial \varepsilon^{(s)} \rho^{(s)} h^{(s)}}{\partial t} + \varepsilon^{(s)} \rho^{(s)} \vec{v}^{(s)} c_p^{(s)} \nabla T^{(s)} = H^{(g,s)} \left(T^{(g)} - T^{(s)} \right) + S^{(s)} \tag{3.20}$$

is solved for $T^{(s)}$ in a loop over all subcells as described in Sec. 3.3.4, yielding an interim approximation $\bar{T}_k^{(s)}(b_{ij}) \forall b_{ij}$.

Then, the field $\bar{T}_k^{(s)}(b_{ij})$ is projected onto the grid Ω_h by interpolation, yielding a field $\bar{T}_k^{(s)}(P_i) \forall i \in \{1, \bar{N}\}$. In the second step, the diffusion equation

$$\frac{\partial \varepsilon^{(s)} \rho^{(s)} h^{(s)}}{\partial t} - \nabla . \left(\varepsilon^{(s)} \lambda^{(s)} \nabla T^{(s)} \right) = 0 \tag{3.21}$$

is solved on Ω_h with initial conditions $T^{(s)}(P_i) = \bar{T}^{(s)}(P_i) \forall i \in \{1, ..., \bar{N}\}$ and boundary conditions $T^{(s)} = T_{in}^{(s)}$ at the furnace top, prescribed heat loss $\lambda \frac{\partial T^{(s)}}{\partial n} = H_{\Gamma_i}^{out}(T^{(s)} - T^{out})$ through the wall and zero normal derivative through the outlets. The numerical solution of the second-order equation (3.21) is calculated with an FEM with bilinear elements. The result $T_k^{(s)}(P_i) \forall i \{1, ..., \bar{N}\}$ is projected onto the grid Ω_h^s by interpolation, yielding $T_{k+1}^{(s)}(b_{ij}) \forall b_{ij}$.

3.4 The global iteration

In the sections 3.2 and 3.3 (and in more detail in [Goe05]) and in [Sch00]), the numerical methods for the solution of the various modules of the blast furnace model have been presented. Due to the coupling of all processes and hence modules, a global iteration over all modules for solving the complete mathematical model at time $t_n = t_{n-1} + \Delta t$ is required (given a solution at time t_{n-1}). For each time step, we have to iterate over the gas flow, solid flow and chemistry / temperature model (K denotes the iteration counter within each time step):

- initial values from t_{n-1} are given by $(p^{(g)}, v^{(g)}, p^{(s)}, v^{(s)}, v^{(l)}, \vec{\xi}^{(g)}, T^{(g)}, p, \vec{\xi}^{(p)}, T^{(s)}, \varepsilon^{(l)}, \vec{\xi}^{(l)}, T^{(l)})(t_{n-1})$ and are chosen to be starting values for t_n (global iteration K=0).
- solve module **Liquid flow** (see [Sch00]), update $v^{(l)}$(only at t_0).
- *WHILE CONVERGENCE IS NOT REACHED*
  - $K = K+1$
  - solve module **Gas flow**, update $(v_K^{(g)}, p_K^{(g)})$ (sec. 3.2)

- solve module **Solid flow**, update $(v_K^{(s)}, p_K^{(s)})$ ([Sch00])
- solve module **Chemistry / temperatures** (sec. 3.3)
- update $(\vec{\xi}^{(g)}, T^{(g)}, p, \vec{\xi}^{(p)}, T^{(s)}, \varepsilon^{(l)}, \vec{\xi}^{(l)}, T^{(l)})_K$

- set values at t_n
  $(p^{(g)}, v^{(g)}, p^{(s)}, v^{(s)}, \vec{\xi}^{(g)}, T^{(g)}, p, \vec{\xi}^{(p)}, T^{(s)}, \varepsilon^{(l)}, \vec{\xi}^{(l)}, T^{(l)})(t_n) =$
  $(p^{(g)}, v^{(g)}, p^{(s)}, v^{(s)}, \vec{\xi}^{(g)}, T^{(g)}, p, \vec{\xi}^{(p)}, T^{(s)}, \varepsilon^{(l)}, \vec{\xi}^{(l)}, T^{(l)})_{K+1}$.

When convergence is reached, we have a new approximation for all values at t_n. In the following Sec. 3.5, results of this algorithm are presented, checked for accuracy and compared to measurements.

3.5 Results

The results of the following example are calculated with the algorithm roughly described in the previous sections. The input quantities, e.g. geometry, raw materials, reaction and process parameters are typical values of a real blast furnace, although we do not give any detailed information about the exact input data and results for confidentiality reasons; the iron ore is a mixture of Sinter, lump ore and pellets with main components $Fe_2O_3(> 80\%), CaO, SiO_2$. Coke mainly consists of $C(> 85\%), Al_2O_3, SiO_2$. The distribution of the coke / ore layers is given by a material matrix (see Fig. 3.5).

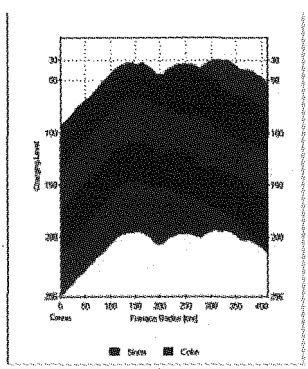


Figure 3.4 Material layers from Burden Distribution Model

Starting from a prescribed daily hot metal production, the fuel consumption is the main result of the simulation (the fuel consumption is defined as the consumed carbon from the coke/oil/tar - gasification at the tuyere and the carbon consumption from carburization, Boudouard

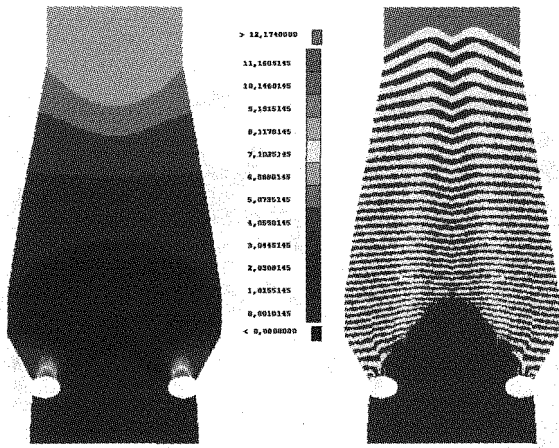


Figure 3.5 Solid velocity [m/h] and coke—ore layer distribution

reaction and direct reduction). Other results are the amount, composition (CO, CO_2, H_2O, H_2, N_2) and temperature of the top gas and the liquid phase (metal and slag). All pictures presented in this section are snapshots at some fixed time of the 2 day time interval, for which this calculation has been carried out.

In Fig. 3.5, the burden velocity and the distribution of coke an ore layers can be seen. In the lower part, all (yellow) ore layers have been molten. Fig. 3.6 visualizes the burden temperature and the reduction degree of the solid iron ore. In Fig. 3.7, the higher pressure loss and slower gas velocity in the cohesive zone can be seen. Fig. 3.8 visualizes the concentration of CO_2 in the blast furnace and Fig. 3.9 shows the distribution of the temperature of liquid iron and slag.

We have tested the accuracy of the numerical algorithm in two ways. The transient mass balances are not enforced in our model and hence can be used as a test for the accuracy of the simulation by comparing the amount of in- and out-flowing components over longer periods (which should be equal). The table below shows a comparison of the (8 h-) floating average of the in- and out-flowing carbon and oxygen, which are contained in the solid, gaseous and liquid phase (in [kg / t hot metal]):

C		O	
in (coke, tar, oil)	397	in (Fe_2O_3,H_2O,blast)	783
out (hot metal, CO,CO_2)	399	out (CO,CO_2,H_2O)	771

Another important test is given by comparing the results to measurements on a real blast furnace, e.g., the composition and temperature of

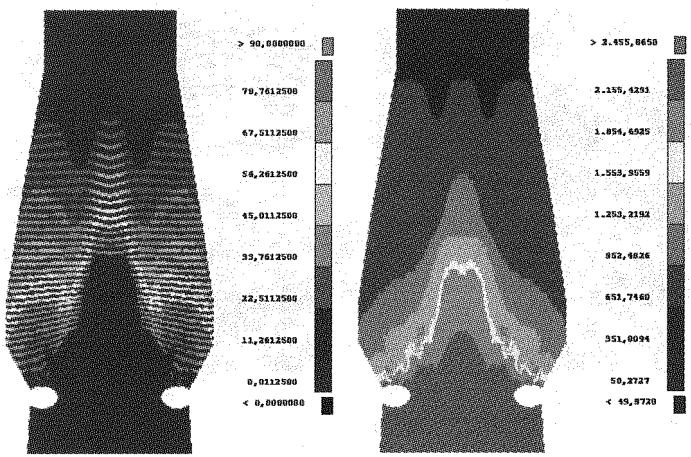


Figure 3.6 Reduction degree of ore [%] and solid temperature [°C]

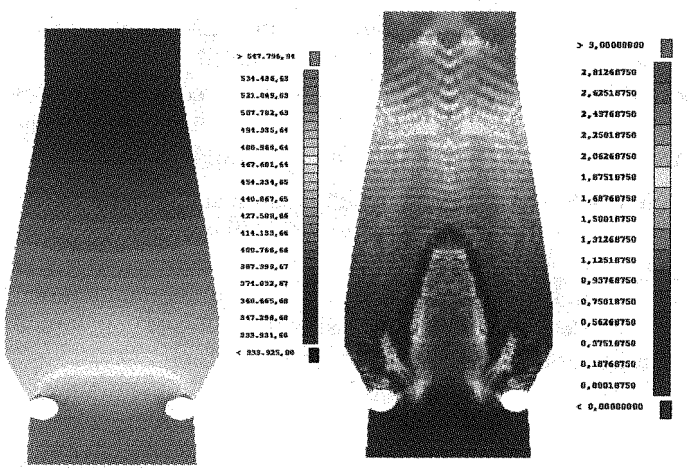


Figure 3.7 Gas pressure [Pa] and mass flow density [kg / m^2s]

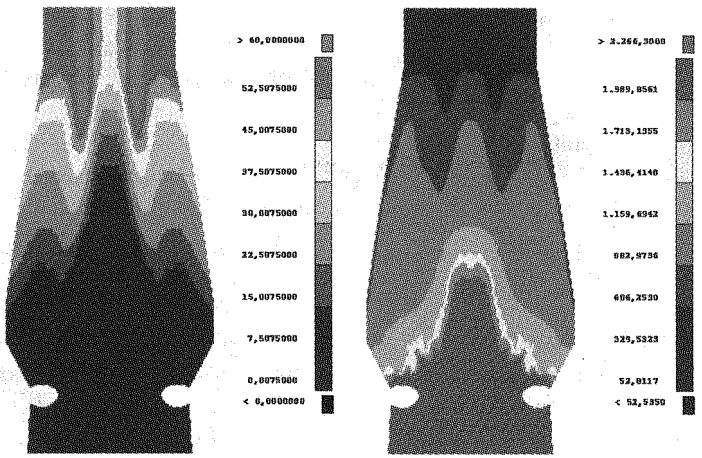


Figure 3.8 Gas utilization (= molar fraction of CO_2 : molar fraction of $CO_2 + CO$) [%]) and temperature [°C]

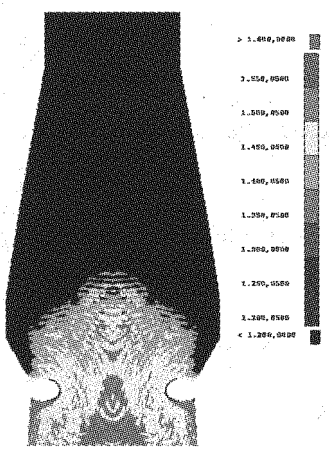


Figure 3.9 Liquid temperature [°C]

top gas and hot metal/slag, the gas utilization and temperature at a burden probe in the upper part of the blast furnace and the fuel consumption. We have compared the results to measured values on 6 different days of a real blast furnace with qualitatively and quantitatively very good results. E.g., the predicted amount of coke required for producing hot metal at some prescribed temperature always was within a range of 5 [kg/t Hot Metal] of the measured values, the total amounts being in the range of 400 [kg / t Hot Metal]). Thus, the differences between computed and measured values are comparable to measurement errors.

These tests show the practicability of the mathematical model and of the numerical algorithm.

Acknowledgements

This work was partially supported by the Austrian Ministry for Economy and Labour and by the Government of Upper Austria within the framework 'Industrial Competence Centers'. The authors would like to thank Steffen Wiedner (voestalpine) for providing important fundamental blast furnace knowledge and process data. The authors are grateful for the valuable modelling contributions of Georg Aichinger (Siemens VAI), for the programming work of Jan Mener (MathConsult) and Manuela Widi (Siemens VAI) for providing important data.

References

[ANY97a] P. R. Austin, H. Nogami, and J. Yagi. A mathematical model for blast furnace reaction analysis base on the four fluid model. *ISIJ International*, 37(8):748–755, 1997.

[ANY97b] P. R. Austin, H. Nogami, and J. Yagi. A mathematical model of four phase motion and heat transfer in the blast furnace. *ISIJ International*, 37(5):458–467, 1997.

[dCNY02] J. A. de CASTRO, H. Nogami, and J. Yagi. Three-dimensional multiphase mathematical modeling of the blast furnace based on the multifluid model. *ISIJ International*, 42(1):44–52, 2002.

[Erg52] S. Ergun. Fluid flow through packed columns. *Chemical Engineering Progress*, 48(2):89–94, 1952.

[FFE95] E. Fitzer, W. Fritz, and G. Emig. *Technische Chemie, Einführung in die Reaktionstechnik.* Springer, 1995.

[Goe05] G. Goekler. *Moving Layers - Tracking Characteristics. A Mathematical Model and the Numerical Simulation of two Ironmaking Processes.* PhD thesis, Johannes Kepler Universität Linz, 2005.

[HG87] E. Hairer and Wanner G. *Solving Ordinary Differential Equations II.* Springer, 1987.

[KP00] S. Kim and Chock D. P. Dual-mesh characteristics for particle- mesh methods for convection-dominated diffusion problems. *University of Kentucky Mathematics Department Research Report*, 2000.

[oIR87] Joint Society on Iron and Steel Basic Research. *Blast Furnace Phenomena and Modelling.* Elsevier, London, New York, 1987.

[RJ70] J. Radestock and R. Jeschar. Über die Strömung durch die Hochofenschüttung. *Stahl und Eisen*, 90(22):1249–1255, 1970.

[Sch00] A. Schatz. *Lump Ore ,Pellets and Dead Man. Mathematical Modelling and Numerical Simulation of the COREX Reduction Shaft.* PhD thesis, Johannes Kepler Universität Linz, 2000.

[Yag93] J. Yagi. Mathematical modeling of the flow of four fluids in a packed bed. *ISIJ International*, 33(6):619–639, 1993.

[ZAJY00a] S. A. Zaimi, T. Akiyama, Guillot J., and J. Yagi. Sophisticated multi-phase multi-flow modeling of the blast furnace. *ISIJ International*, 40(4):322–331, 2000.

[ZAJY00b] S. A. Zaimi, T. Akiyama, Guillot J., and J. Yagi. Validation of a blast furnace solid flow model using reliable 3-d experimental results. *ISIJ International*, 40(4):332–341, 2000.

[Zei95] H. Zeisel. *Mathematische Modellierung und numerische Simulation der Vorgänge im Hochofen.* PhD thesis, Johannes Kepler Universität Linz, 1995.

Gelification of Hydrocarbons: a Model Problem

Loredana Faienza Antonio Fasano
Mario Primicerio
Dipartimento di Matematica "U. Dini"
Viale Morgagni 67/a, 50134 Firenze.

Abstract

We consider the mixture of a light monocomponent oil and a paraffin, assuming that the latter can be dissolved in the oil or can be present in form of a crystalline phase.

We study the case where a prescribed temperature gradient drives the diffusive migration of the dissolved paraffin accompained by the process of crystallization/dissolution.

We consider the case in which a gel can be formed and present a mathematical model to take this phenomenon into account. The paper also contains some qualitative results and numerical simulation.

1 Introduction

A conceptual model for the behaviour of mixtures of light and heavy hydrocarbons in the presence of temperature gradients can be constructed considering an ideal solution of a selected heavy hydrocarbon (in the sequel generically referred to as paraffin) in a light monocomponent oil. For a given concentration of paraffin two temperatures can be defined, T_c and T_G, such that for $T > T_c$, the paraffin is completely dissolved in the oil , for $T \in (T_G, T_c)$ paraffin exists both dissolved in oil and in form of segregated crystals and for $T < T_G$ *gelification* occurs i.e. the mixture has the form of a gel (see e.g. [1], [7], [3], [4], [2]). Conversely, we can say that two increasing functions of temperature are given, $c_1(T)$ and $c_2(T)$, $c_1 < c_2$, such that (assuming thermodynamical equilibrium) paraffin is completely dissolved if its concentration is below $c_1(T)$, gel is formed if the concentration of paraffin is above $c_2(T)$, and for $c \in (c_1(T), c_2(T))$ paraffin may be both dissolved in oil and in form of segregated crystals. We will call c_1 and c_2 the *saturation* and the *gelification* concentration respectively. If the system is not in thermodynamical equilibrium, crystals can persist also below saturation.

In order to illustrate the main features of a mathematical model describing the system we will consider a simplified situation in which:

i. the mixture occupies the slab $x \in (0, L)$, and all relevant quantities depend on the space coordinate x (and possibly on time t);

ii. temperature is costant with respect to time t and is a linear function of x[1];

iii. c_1 and c_2 are linear functions of T (and thus of x, because of (ii));

$$c_1(x) = a_1 + b_1 x, \tag{1.1}$$

$$c_2(x) = a_2 + b_2 x, \tag{1.2}$$

where a_i and b_i (i=1,2) are positive costants such that

$$c_1(x) < c_2(x), \quad x \in [0, L].$$

Upon renormalization, we can take $L = 1$.

One further assumption concerns the kinetics of transformation between dissolved and segregated paraffin. We denote by u and v the concentrations of dissolved and segregated paraffin, respectively. Still in the spirit of making the simplest assumptions that preserve the essential features of the problem, we assume that the rate of crystallization (or of dissolution) of paraffin is proportional to the difference between u and the saturation concentration c_1.

Therefore $\beta > 0$ is given such that:

$$v_t = \beta(u - c_1), \quad \text{unless } v \equiv 0 \text{ and } u \leq c_1. \tag{1.3}$$

Introducing the Heaviside function

$$H(z) = \begin{cases} 0\,, z \leq 0, \\ 1\,, z > 0, \end{cases}$$

we can write

$$v_t = H(v + (u - c_1)_+)\beta(u - c_1), \tag{1.4}$$

where $(z)_+ = \max(z, 0)$.

Paraffin has essentially the same density of the solvent, so that sedimentation is irrelevant. On the contrary, diffusion can not be neglected. Diffusivity is much larger for the dissolved paraffin then for the segregated phase. Moreover, the solution gelifies (i.e. when $u + v \geq c_2$) all species become practically immobile. We will consider the case in

[1]This assumption is justified if thermal diffusivity does not depend on paraffin concentration, if latent heat of crystallization and gelification can be neglected, and if the characteristic heat diffusion time is much shorter than the time scale of the phenomenon under consideration.

which diffusivity is non vanishing only for the dissolved phase, so that the equations to be satisfied (in some suitable weak sense) are:

$$u_t - [DH(c_2 - u - v)u_{xx}] = -v_t,\ x \in (0,1),\quad t > 0, \tag{1.5}$$

$$v_t = H(v + (u - c_1)_+)\beta(u - c_1),\ x \in (0,1),\quad t > 0 \tag{1.6}$$

(where D is the diffusivity of the dissolved wax), with the initial and boundary conditions:

$$u_x(0,t) = 0,\quad t > 0, \tag{1.7}$$

$$u_x(1,t) = 0,\quad t > 0, \tag{1.8}$$

$$u(x,0) = c_1(x),\ x \in (0,1), \tag{1.9}$$

$$v(x,0) = c^\star - c_1(x) > 0,\ x \in (0,1). \tag{1.10}$$

In (1.7)-(1.10) we have assumed that:

- walls $x = 0$ and $x = 1$ are impermeable;
- at time $t = 0$ the total concentration of paraffin is $c^\star$ and there is thermodynamical equilibrium between the segregated and dissolved phases.

According to the scheme above, once gel is formed, the quantity $u+v$ remains constant in time and equal to the value of c_2 at the location in which gelification has occurred. Actually gel undergoes a process ("ageing") consisting in progressive thickening with oil expulsion ([5], [6]). Since the time scale of ageing is usually much longer then the one associated with diffusion of dissolved paraffine, in this model, we disregard it.

2 The stage preceding the appearance of free boundaries

Assume $c^\star$ is such that:

$$c^\star - c_1(1) > 0, \tag{2.1}$$

$$c^\star - c_2(0) < 0, \tag{2.2}$$

so that the initial situation is always above saturation and below gelification. This means that if (1.5)-(1.10) admits a (continuous) solution, then a non-void set $\mathbb{Q}_T \equiv (0,1) \times (0,T)$ will exists such that:

$$v > 0 \text{ and } u + v < c_2(x) \text{ in } \mathbb{Q}_T.$$

As we shall see, both inequalities will be violated (at different finite times) giving origin to the sets

$$\begin{array}{ll} \{v = 0\}\ , & \text{the crystal-free region,} \\ \{u + v = c_2(x)\}, & \text{the gel region.} \end{array}$$

Each part of the boundary of these sets lying in $(0,1) \times \mathbb{R}^+$ of the (x,t) plane is a free boundary of the problem.

In this section we analyze the problem in $\mathbb{Q}_T$, i.e. before the appearance of any free boundary. It is clear that if (2.1) and/or (2.2) are violated then $\mathbb{Q}_T = \emptyset$, at least one free boundary starts from $t = 0$ and the analysis of Section 3 below applies.

In $\mathbb{Q}_T$ the problem reduces to the classical parabolic problem for u

$$\begin{cases} u_t - Du_{xx} + \beta(u - c_1) = 0, \text{ in } \mathbb{Q}_T, \\ u(x,0) = c_1(x), & x \in (0,1), \\ u_x(0,t) = u_x(1,t) = 0, & t \in (0,T), \end{cases} \tag{2.3}$$

and to the first order problem for v

$$\begin{cases} v_t = \beta(u - c_1), & \text{in } \mathbb{Q}_T, \\ v(x,0) = c^\star - c_1(x), & x \in (0,1), \end{cases} \tag{2.4}$$

which can be integrated immediately once (2.3) has been solved. Of course (2.3) has a unique solution that is positive and continuous in $\overline{\mathbb{Q}}_T$, infinitely differentiable in any closed rectangle $[0,1] \times [\epsilon, T]$, $\epsilon > 0$. Moreover:

Proposition 2.1. *We have*

$$0 < u_x < b_1, \qquad \textit{in } \mathbb{Q}_T. \tag{2.5}$$

Proof: The function

$$w(x,t) = (u_x - b_1)e^{\beta t} \tag{2.6}$$

satisfies

$$\begin{cases} w_t - Dw_{xx} = 0, & \text{in } \mathbb{Q}_T, \\ w(x,0) = 0, & x \in (0,1), \\ w(0,t) = w(1,t) = -b_1 e^{\beta t}, & t > 0. \end{cases} \tag{2.7}$$

Then the strong maximum principle yields (2.5). □

Moreover we have

Corollary 2.2. *The segregated wax concentration v is decreasing* w.r.t. *x in $\mathbb{Q}_T$.*

Proof: Indeed $v_x(x,0) = -b_1 < 0$ and, by (2.4), (2.5), it is

$$v_{xt} < 0, \text{in } \mathbb{Q}_T \tag{2.8}$$

□

Next we use some additional imformation on $w(x,t)$. As it is well known, (2.7) can be solved explicitly and $w(x,t)$ is symmetric w.r.t. $x = 1/2$ with $w_x > 0$ for $x < 1/2$ and $w_x < 0$ for $x > 1/2$. Hence:

Proposition 2.3. *We have*

$$u_{xx} > 0 \text{ , in } (0,1/2) \times (0,T), \tag{2.9}$$
$$u_{xx} < 0 \text{ , in } (1/2,1) \times (0,T). \tag{2.10}$$

Next we prove

Proposition 2.4. *The difference $u(x,t) - c_1(x)$ is negative (positive) for $x > 1/2$ (for $x < 1/2$).*

Proof: Since $u_{xx}(1/2,t) = 0$, from (2.3) we see that, setting $z(t) = u(1/2,t) - c_1(1/2)$, it solves

$$\dot{z}(t) + \beta z(t) = 0, \quad z(0) = 0,$$

implying $z(t) \equiv 0$. Using (2.5) the proof is concluded. □

Since the total wax concentration (dissolved plus segregated)

$$C(x,t) = u(x,t) + v(x,t)$$

satisfies

$$\begin{cases} C_t = Du_{xx}, & \text{in } \mathbb{Q}_T, \\ C(x,0) = c^\star, & x \in (0,1). \end{cases} \tag{2.11}$$

(2.8) implies

Proposition 2.5. *The total wax concentration $C(x,t)$ is increasing* w.r.t. *time in $(0,1/2) \times (0,T)$ and decreasing in $(1/2,1) \times (0,T)$.*

On the other hand, since w_{xx} is negative in $\mathbb{Q}_T$, because of the maximum principle, we also have $C_{xt} < 0$. As $C_x(x,0) = 0$ we conclude

Proposition 2.6. *The total wax concentration is decreasing* w.r.t. *x in $\mathbb{Q}_T$ (and its slope is increasing with time).*

Let

$$T_0 = \sup\{T : v > 0 \text{ and } u + v < c_2 \text{ in } \mathbb{Q}_T\}.$$

According to Corollary 2.2 and Proposition 2.5, if $T_0 < +\infty$ at least one of the following two facts has to occur:

$$C(0,T_0) = c_2(0), \quad v(1,T_0) = 0. \tag{2.12}$$

We have

Proposition 2.7. *T_0 has to be finite.*

Proof: Suppose $T_0 = +\infty$. Proposition 2.4 and 2.5 ensure that v, C and u have a limit for $t \to +\infty$. Then, necessarily

$$\lim_{t \to +\infty} u(x,t) = c_1(x) = a_1 + b_1 x \tag{2.13}$$

which is incompatible with the no flux boundary conditions in (2.3). □

Now we are motivated to study the problem with free boundaries.

3 Qualitative analysis of the problem with free boundaries

Proposition 2.7 says that either a gelification front or a crystal disappearance front has to appear at a finite time T_0. Actually, we will prove that both of them have to be present after some time.

In this a-priori analysis we assume that a classical solution $\{s, \sigma, u, v\}$ to problem (1.5)-(1.10) exists in the following sense.

1. There exists two Lipschitz continuous functions $s(t)$, $\sigma(t)$ so that the half strip $(0,1) \times (0,+\infty)$ is partitioned into three regions:
   - the gel region $\mathcal{G}$: $0 < x < s(t)$, $t > T_g$, $s(T_g) = 0$, $x = s(t)$ being a free boundary;
   - the region with the crystalline phase $\mathcal{C}$: $\hat{s}(t) \le x \le \hat{\sigma}(t)$, $t > 0$, where $\hat{s}(t) = 0$, for $0 \le t \le T_g$, $\hat{s}(t) = s(t)$ for $t > T_g$, and $\hat{\sigma}(t) = 1$ for $0 \le t \le T_d$, $\hat{\sigma}(t) = \sigma(t)$, for $t > T_d$, $x = \sigma(t)$ being a second free boundary;
   - the region depleted of the crystalline phase $\mathcal{D}$: $\sigma(t) < x < 1$, $t > T_d$.
2. In region $\mathcal{G}$ (no mass transfer)
$$C \equiv c_2(x). \tag{3.1}$$
3. In region $\mathcal{C}$ $u, v \in C^{2,1}(\mathcal{C})$, $u, u_x, v \in C(\overline{\mathcal{C}})$ and satisfy the differential equations and the initial conditions in (2.3), (2.4); moreover
$$u_x(0,t) = 0, \quad t < T_g \tag{3.2}$$
and
$$u_x(1,t) = 0, \quad t < T_d. \tag{3.3}$$
4. In region $\mathcal{D}$ $u \in C^{2,1}(\mathcal{D})$, $u, u_x \in C(\overline{\mathcal{D}})$ and
$$u_t - D u_{xx} = 0, \quad (x,t) \in \mathcal{D} \tag{3.4}$$
with the boundary condition
$$u_x(1,t) = 0, \quad t > T_d. \tag{3.5}$$

5. The following conditions on the interfaces are satisfied

   a) on the crystal depletion front $x = \sigma(t)$:

$$v = 0 \ , \ t > T_d, \tag{3.6}$$

$$[u] = [u_x] = 0 \ , \ t > T_d; \tag{3.7}$$

   b) on the gelification front $x = s(t)$,

$$C = c_2 \ , \ t > T_g, \tag{3.8}$$

$$u_x = 0 \ , \ t > T_g. \tag{3.9}$$

(The latter conditions express mass conservations and continuity of C on $x = s(t)$: apply Rankine-Hugoniot condition to the equation for C obtained by adding (1.5) and (1.6)).

Proposition 3.1. *There exists a finite time T_d such that $v(1, T_d) = 0$,* i.e. *the crystal depletion front $x = \sigma(t)$ always starts.*

Proof: Suppose that the claim is false. By Proposition 2.7 a first time T_g exists such that $C(0, T_g) = c_2(0)$, i.e. the gelification front $x = s(t)$ starts, $s(T_g) = 0$, on which $C \equiv c_2$. In the region $\widehat{\mathcal{Q}} \equiv \{(x,t) : \widehat{s}(t) < x < 1, t > 0\}$, where

$$\widehat{s}(t) = \begin{cases} 0, & t \leq T_g, \\ s(t), & t > T_g, \end{cases} \tag{3.10}$$

the wax concentration $u(x,t)$ satisfies problem (2.3), thanks to (3.9). Therefore the same results proved for u in $\mathbb{Q}_T$ can be duplicated for $\widehat{\mathcal{Q}}$ with obvious modifications. Particularly we have:

- v is a decreasing function w.r.t. x in $\widehat{\mathcal{Q}}$;
- $u_{xx}(s(t), t) > 0$, because on $x = s(t)$ u_x assumes the minimum value.

Hence, differentiating condition (3.8) with respect to time t we obtain:

$$(v_x(s(t), t) - b_2)\dot{s} + Du_{xx}(s(t), t) = 0, \tag{3.11}$$

thus $\dot{s} > 0$.

Now we note that $x = s(t)$ cannot reach the boundary $x = 1$ at any time $\widetilde{t}$, since this would imply:

$$c^\star = \int_0^1 (u(x, \widetilde{t}) + v(x, \widetilde{t}))dx = \int_0^1 c_2(x)dx, \tag{3.12}$$

that contradicts mass conservation because of (2.2). Therefore, the asymptotic limit of s, s_∞, is such that $s_\infty < 1$. If the crystal depletion front never starts , i.e. $T_d = +\infty$, then

$$\lim_{t\to+\infty} u(x,t) = c_1(x), \qquad x \in (s_\infty, 1). \tag{3.13}$$

Once again, this limit does not match the no flux boundary conditions.□

Proposition 3.2. *Until the boundary $x = \sigma(t)$ is separated from the level curve $u = c_1$, it is strictly decreasing. If $\dot\sigma$ becomes positive in some interval, then σ must coincide with the curve $u = c_1$ in that interval.*

Proof: When it first appears the region $v = 0$ is necessarily embedded in an unsaturated solution. Clearly, as long as $u < c_1$ once crystals have disappeared they cannot be produced again and the boundary $x = \sigma(t)$ keeps being decreasing. Therefore a possible increase of $x = \sigma(t)$ is marked by its intersection with the curve $u = c_1$. If we want $\sigma(t)$ to become increasing then the necessity that v becomes positive for $x < \sigma(t)$ requires that $u > c_1$ in the same region. Thus $x = \sigma(t)$ and $u = c_1$ must coincide. □

Proposition 3.3. *The function v is decreasing* w.r.t. *in its support.*

Proof: The proof relies on the following inequality

$$u_x - b_1 \le 0. \tag{3.14}$$

This can be proved easily as long as the crystal depleted region has not appeared (same technique as in Prop. 2.1). In the presence of the free boundary $x = \sigma(t)$ we may consider the equation

$$u_t - Du_{xx} + \beta H(v)(u - c_1) = 0 \quad \text{in } \mathcal{C} \cup \mathcal{D}$$

and approximate $H(v)$ by means of a sequence mollified monotonic functions $H_n(v)$. Differentiating w.r.t. x and setting $\omega = \omega_x - b_1$, we find

$$\omega_t - D\omega_{xx} + \beta H_n \omega = -\beta H_n' v_x (u - c_1) \tag{3.15}$$

with

$$\omega(x,0) = 0,\ \omega(\hat s, t) = -b,\ \omega(1,t) = -b.$$

Since the right hand side of (3.15) is different from zero only in proximity of $x = \sigma(t)$, where $v_x \le 0$ and $u - e_1 \le 0$, and therefore is nonnegative. The maximum principle implies $\omega < 0$ in $\mathcal{C} \cup \mathcal{D}$ for all approximations, thus proving (3.14). □

At this point we recall that $v_{xt} = -\beta(u_x - b_1)$ in and that $v_x = -b_1$ for $t = 0$ and $v_x \leq 0$ along $x = \sigma(t)$. Hence, as a consequence of (3.14), $v_x \leq 0$, which in term implies that region is connected. □

Remark 3.4. We note that at all points (x, t) such that $x < \inf_{\tau<t} \sigma(\tau)$ actually we have $v_x < -b_1$ and consequently also $C_x = u_x + v_x < 0$.

Let us consider the possible asymptotic behaviour of the solution.

Proposition 3.4. *If the boundaries $x = s(t), x = \sigma(t)$ have the respective limits s_∞, σ_∞ for $t \to \infty$, then they both must coincide with the only solution in $(0, 1)$ of the quadratic equation*

$$(\frac{b_2}{2} - b_1)y^2 + (a_2 + b_1 - a_1)y + a_1 - c^\star = 0. \tag{3.16}$$

Proof: The equality $s_\infty = \sigma_\infty$ follows from the fact that supposing $s_\infty > s_\infty$ implies $u \to c_1(x)$ as $t \to \infty$ in (s_∞, s_∞), contradicting the no flux conditions (note that $u \to c_1(\sigma_\infty)$ in $(\sigma_\infty, 1)$). Equation (3.16) is obtained as the limit for $t \to +\infty$ of the mass balance equation:

$$\oint_{\partial(\mathcal{C}_t \cup \mathcal{D}_t)} (Cdx + Du_x d\tau) = 0, \tag{3.17}$$

where $\mathcal{C}_t$, $\mathcal{D}_t$ are the positions of the domains $\mathcal{C}$, $\mathcal{D}$ where the time coordinate is less than t. We note that $\int_0^1 (u + v)dx = c^*$ up to the time T_G of appearance of the gel.

Thus we can deduce

$$c^* = \int_{s(t)}^{\sigma(t)} (u + v)dx + \int_{s(t)}^{1} udx + \int_{T_G}^{t} c_2(s(\tau))\dot{s}(\tau)d\tau. \tag{3.18}$$

The last term is equal to $a_2 s(t) + \frac{1}{2} b_2 s^2(t)$. Taking the limit $t \to +\infty$ the first term vanishes, the second term tends to$c_1(s_\infty)(1 - s_\infty) = (a_1 + b_1 s_\infty)(1 - s_\infty)$, and we get (3.18). Call $D_2(y)$ the quadratic polynomial in (3.16). We know that $D_2(0) = a_1 - c^\star < 0$, because of (2.1) and that $D_2(1) = \frac{b_2}{2} + a_2 - c^\star > 0$ because of (2.2). Thus there exists a solution in $(0, 1)$ of (3.16). Moreover, irrespectively of the sign of $D_2'' = b_2 - 2b$, supposing that there can be two solutions in $(0, 1)$ contradicts the fact $D_2(0)D_2)(1) < 0$. □

Remark 3.6. The total concentration C is asymptotically discontinuous. The jump is due to v which tends to $c_2(s_\infty) - c_1(s_\infty)$ on the gel surface and to zero in the solution.

We conclude our analysis considering, as a limit case, the situation in which β vanishes. Hence the segregated wax concentration is constant w.r.t. time and equal to $v_0(x)$. In our example

$$v(x,t) = c^* - c_1(x) = c^* - a_1 - b_1 x. \tag{3.19}$$

Of course, in this case $\mathcal{D} = 0$, i.e. the crystalline phase is present everywhere. On the other hand, if the problem

$$\begin{cases} u_t - Du_{xx} = 0, & 0 < x < 1, \quad t > 0, \\ u(x,0) = c_1(x), & 0 < x < 1, \\ u(x,t) = u_x(1,t) = 0, & t > 0 \end{cases} \tag{3.20}$$

is such that

$$C(x,t) < c_2(x), \quad 0 < x < 1, \quad t > 0, \tag{3.21}$$

i.e. if

$$u(x,t) < c_2(x) - c^* + c_1(x) \tag{3.22}$$

gelification does not start.

Since $z(x,t) = u(x,t) - C_1(x)$ is such that $0 > z_x > b_1$ then a gelification starts if and only if, for some finite t^*, $C(0,t) = c_2(0)$ i.e. if

$$u(0,t) = c_2(0) + c_1(0) - c^*. \tag{3.23}$$

Since $u(0,t)$ is monotonically increasing from $c_1(0)$ to its asymptotic value coinciding with the average of $u(x,0)$, a necessary condition for the existence of a finite t^* satisfying (3.23)

$$c^* > a_2 - b_1/2. \tag{3.24}$$

Assume (3.24) holds. Solve (3.20) till time t^* defined in (3.23) and put

$$u^*(x) = u(x,t^*), \quad 0 < x < 1.$$

Then the gelification front in the free boundary $x = s(t)$ of the following problem

$$\begin{cases} u_t - Du_{xx} = 0, & s(t) < x < 1, \; t > t^*, \\ u(x,t^*) = u^*(x), & 0 < x < 1, \\ u_x(s(t),t) = u_x(1,t) = 0, & t > t^*, \\ u(s(t),t) = c_2(s(t)) - v(s(t)). \end{cases} \tag{3.25}$$

We write the last condition explicitly

$$u(s(t),t) = a_2 + b_2 s(t) - c^* + a_1 + b_1 s(t) = A + Bs(t) \tag{3.26}$$

where A and B are positive constants. It is easily seem that the problem can be reduced to a one-phase Stefan problem. Indeed, differentiate (3.26) w.r.t. t and use the condition $u_x(s(t),t) = 0$. We obtain

$$u_t = B\dot{s}. \tag{3.27}$$

Hence the function

$$W(x,t) = u_x(x,t) \tag{3.28}$$

solves the problem

$$\begin{cases} W_t - Du_{xx} = 0, & s(t) < x < 1,\ t > t^*, \\ W(x,t^*) = u_x^*(x), & 0 < x < 1, \\ W(1,t) = 0, & t > t^*, \end{cases} \tag{3.29}$$

with the free boundary condition

$$\begin{cases} W(s(t),t) = 0, \\ W_x(s(t),t) = \frac{B}{D}\dot{s}. \end{cases} \tag{3.30}$$

Since $u_x^*(t)$ is non negative we have a Stefan problem with a monotonically increasing free boundary $s(t)$ for which well-posedness is a classical result.

4 Numerical simulations

First of all we have considered the problem in the normalized variables:

$$\begin{cases} u_\tau - H(c_2 - u - v)c_{\xi\xi} = -v_\tau, & 0 < \xi < 1, \tau > 0, \\ u(\xi,0) = c_1(\xi), & 0 < \xi < 1, \\ u_\xi(0,\tau) = u_\xi(1,\tau) = 0, & \tau > 0, \end{cases} \tag{4.1}$$

while v solves:

$$\begin{cases} v_\tau = H(v + (u - c_1)_+)\widehat{\beta}(u - c_1), & \tau > 0, \\ v(\xi,0) = c^\star - c_1(\xi), & 0 < \xi < 1, \end{cases} \tag{4.2}$$

where all concentrations are divided by $c^\star$ and

$$\xi = x/L, \quad \tau = \frac{D}{L^2}t, \quad \widehat{\beta} = \frac{\beta}{D}L^2;$$

in which L represents the length of a generic slab. The problem (4.1)-(4.2) has been solved numerically considering an oil with the following properties:

- $T_c = 299°\mathrm{K}$;

- $T_G = 289°\text{K}$;
- $c^\star = 20.25\ \text{kg/m}^3$;
- $a_1 = 20.12\ \text{kg/m}^3$, $a_2 = 20.32\ \text{kg/m}^3$, $b_1 = b_2 = 0.09\ \text{kg/m}^2$.

Moreover in our simulations we have assumed that:

$$\beta = 10^{-5}\text{s}^{-1}, \qquad D = 10^{-7}\text{m}^2/\text{s}.$$

We consider an initial situation in which the solution is completely saturated but not gelified (see Figs.4.1 – 4.2).

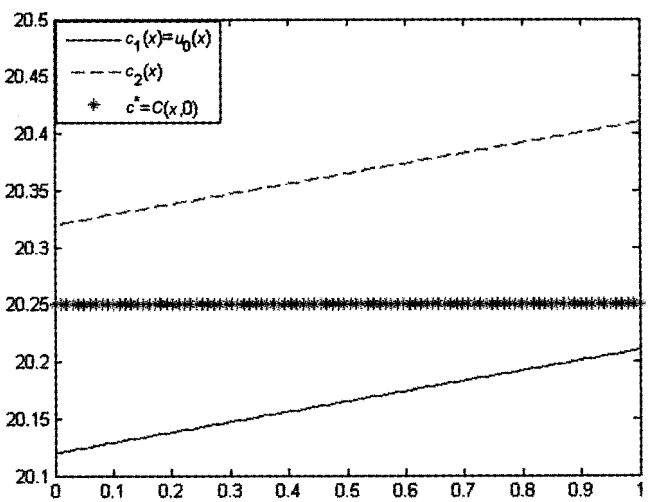


Figure 4.1 Saturation and gelification curves and the total wax concentration

The numerical results of the model problem (obtained through a finite difference approximation of the problems (4.1), (4.2)) produce the concentrations profiles shown in Fig.4.3.

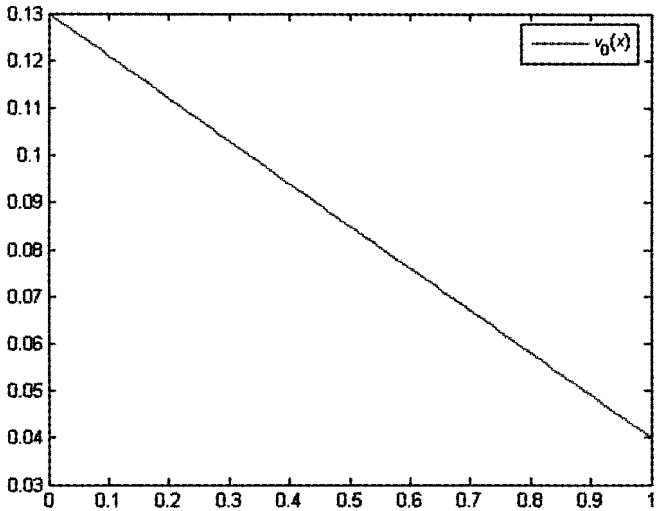


Figure 4.2 The initial segregated wax profile

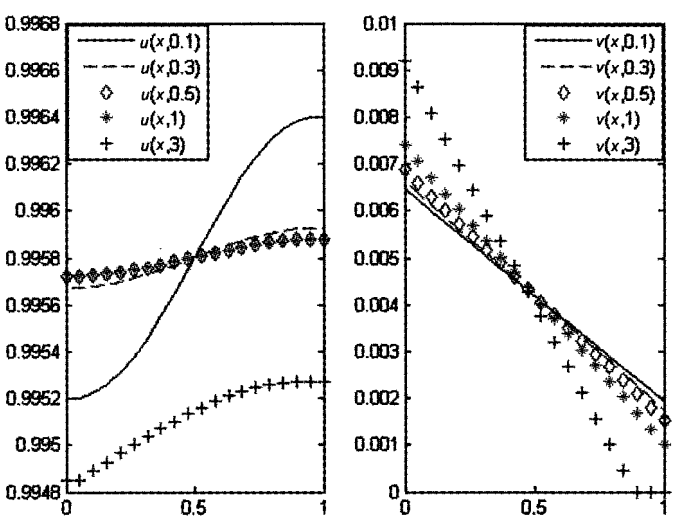


Figure 4.3 Profiles of dissolved (left) and segregated (right) wax at different times

References

[1] L.F.A. Azevedo, A.M. Teixeira, A critical review of the modeling of wax deposition mechanism, Petroleum Science and Technology, 2003, 21 (3&4):393-408.

[2] J.H.Hausen, A. Frechenslund, K.J. Pedersen, H.P.Romminger, A thermodynamic model for predicting wax formation in crude oils , AIChE J. 34 1937-1942.

[3] K.G.Paso, H.S.Fogler, Bulk Stabilization in Wax Deposition Systems, Energy Fuels 18(4):1005-1013.

[4] P.Singh, A.Youyen, H.S.Fogler, AIChE Journal, 2001, 47(9):2111-2124.

[5] P.Singh, R.Verkartesan, S.Fogler, Morphological evolution of thr tick wax deposit during ageing, AIChE J. 2001,47: 6-18.

[6] P.Singh, R.Verkartesan, S.Fogler, Formation and ageing of incipient thin film wax-oil gels, AIChE J. 46 1059-1074.

[7] C.H.Wu, J.L.Creek, K.S.Wang, R.M.Carlson, S.Cheung, P.J.Shuler, Y.Tang, Measurement of Wax Deposition in Paraffin Solutions, prepared for presentation at the 2002 Spring National Meeting, New Orleans, March 10-14, "Wax Thermodynamic and Deposition",pp 90-95.

On the Martingale Representation Theorem and on Approximate Hedging a Contingent Claim in the Minimum Deviation Square Criterion

Nguyen Van Huu

Vietnam National University, Hanoi

Vuong Quan Hoang

ULB Belgium.

Abstract

In this work we consider the problem of the approximate hedging of a contingent claim in the minimum mean square deviation criterion. A theorem on martingale representation in case of discrete time and an application of the result for semi-continuous market model are also given.

Keywords Hedging, contingent claim, risk neutral martingale measure, martingale representation

1. Introduction

The activity of a stock market takes place usually in discrete time. Unfortunately such markets with discrete time are in general incomplete and so super-hedging a contingent claim is requires usually an initial price two great, which is not acceptable in practice.

The purpose of this work is to propose a simple method for approximate hedging a contingent claim or an option in minimum mean square deviation criterion.

Financial market model with discrete time:

Without loss of generality let us consider a market model described by a sequence of random vectors $\{S_n,\ n = 0, 1, \cdots, N\}$, $S_n \in \mathbb{R}^d$, which are discounted stock prices defined on the same probability space $\{\Omega,\ \Im,\ P\}$ with $\{F_n,\ n = 0, 1, \cdots, N\}$ being a sequence of sigma-algebras of information available up to the time n, whereas "risk free" asset chosen as a numeraire $S_n^0 = 1$.

A F_N-measurable random variable H is called a contingent claim (in the case of a standard call option $H = \max(S_n - K, 0)$, K *is strike price*).

Trading strategy:

A sequence of random vectors of d-dimension $\gamma = (\gamma_n,\ n = 1, 2, \cdots, N)$ with $\gamma_n = (\gamma_n^1, \gamma_n^2, \cdots, \gamma_n^d)^T$ (A^T denotes the transpose of matrix A), where γ_n^j is the number of securities of type j kept by the investor in the interval $[n-1,\ n)$ and γ_n is F_{n-1} -measurable (based on the information available up to the time $n-1$), then $\{\gamma_n\}$ is said to be predictable and is called *Portfolio or trading strategy* .

Assumptions:

Suppose that the following conditions are satisfied:

(i) $\Delta S_n = S_n - S_{n-1},\ H \in L_2(P),\ n = 0, 1, \cdots, N.$

(ii) Trading strategy γ is self-financing, i.e. $S_{n-1}^T \gamma_{n-1} = S_{n-1}^T \gamma_n$ or equivalently $S_{n-1}^T \Delta\gamma_n = 0$ for all $n = 1, 2, \cdots, N$.
Intuitively, this means that the portfolio is always rearranged in such a way its present value is preserved.

(iii) The market is of *free arbitrage*, that means there is no trading strategy γ such that $\gamma_1^T S_0 := \gamma_1 S_0 \leq 0,\ \gamma_N S_N \geq 0,\ P\{\gamma_N S_N > 0\} > 0$.

This means that with such trading strategy one need not an initial capital, but can get some profit and this occurs usually as the asset $\{S_n\}$ is not rationally priced.

Let us consider

$$G_N(\gamma) = \sum_{k=1}^{N} \gamma_k \Delta S_k \text{ with } \gamma_k \Delta S_k = \sum_{j=1}^{d} \gamma_k^j \Delta S_k^j.$$

This quantity is called the *gain* of the strategy γ.

The problem is to find a constant c and $\gamma = (\gamma_n,\ n = 1, 2, \cdots, N)$ so that

$$E_P(H - c - G_N(\gamma))^2 \to \min. \tag{1.1}$$

Problem (1.1) have been investigated by several authors such as H. Folmer, M. Schweiser, M. Schal, M. L. Nechaev with $d = 1$. However, the solution of problem (1.1) is very complicated and difficult for application if $\{S_n\}$ is not a $\{\mathbf{F}_n\}$-martingale under P, even for $d = 1$.

By the fundamental theorem of financial mathematics, since the market is of free arbitrage, there exists a probability measure $Q \sim P$ such that under Q, $\{S_n\}$ is an $\{\mathbf{F}_n\}$-martingale, i.e. $E_Q(S_n|F_n) = S_{n-1}$ and the measure Q is called *risk neutral martingale probability measure.*

We try to find c and γ so that

$$E_Q(H - c - G_N(\gamma))^2 \to \min \text{ over } \gamma. \tag{1.2}$$

Definition 1.1 $(\gamma_n^*) = (\gamma_n^*(c))$ minimizing the expectation in (1.2) is called Q- optimal strategy in the minimum mean square deviation (MMSD) criterion corresponding to the initial capital c.

The solution of this problem is very simple and the construction of the Q-optimal strategy is easy to implement in practice.

Notice that if $L_N = dQ/dP$ then

$$E_Q(H - c - G_N(\gamma))^2 = E_P[(H - c - G_N)^2 L_N]$$

can be considered as an weighted expectation under P of $(H - c - G_N)^2$ with the weight L_N. This is similar to the pricing asset based on a risk neutral martingale measure Q.

In this work we give a solution of the problem (1.2) and a theorem on martingale representation in the case of discrete time.

It is worth to notice that the authors M.Schweiser, M.Schal, M.L. Nechaev considered only the problem (1.1) with S_n of one-dimension and M.Schweiser need the additional assumptions that $\{S_n\}$ satisfies non-degeneracy condition in the sense that there exists a constant δ in

$(0,1)$ such that

$$(E[\Delta S_n|F_{n-1}])^2 \leq \sigma E[(\Delta S_n)^2|F_{n-1}] \qquad \text{P-a.s. for all } n = 1, 2, \cdots, N.$$

and the trading strategies γ_n's satisfy:

$$E[\gamma_n \Delta S_n]^2 < \infty,$$

while in this article $\{S_n\}$ is of d-dimension and we need not the preceding assumptions.

The organization of this article is as follows:

The solution of the problem (1.2) is fulfilled in paragraph 2. (Theorem 2.1) and a theorem on the representation of a martingale in terms of the differences ΔS_n (Theorem 3.1) will be also given (the representation is similar to the one of a martingale adapted to a Wiener filter in the case of continuous time).

Some examples are given in paragraph 3.

The semi-continuous model, a type of discretization of diffusion model, is investigated in paragraph 4.

2. Finding the optimal portfolio

Notation. Let Q be a probability measure such that Q is equivalent to P and under Q $\{S_n,\ n = 1, 2, \cdots, N\}$ is an integrable square martingale and let us denote $E_n(X) = E_Q(X|F_n)$, $H_N = H$, $H_n = E_Q(H|F_n) = E_n(H)$; $\mathrm{Var}_{n-1}(X) = [\mathrm{Cov}_{n-1}(X_i, X_j)]$ denotes the conditional variance matrix of random vector X when F_{n-1} is given, Γ is the family of all predictable strategies γ.

Theorem 2.1 If $\{S_n\}$ *is an* $\{F_n\}$*-martingale under* Q *then*

$$E_Q(H - H_0 - G_N(\gamma^*))^2 = \min\{E_Q(H - c - G_N(\gamma))^2 : \gamma \in \Gamma\}, \quad (2.1)$$

where γ_n^* *is a solution of the following equation system:*

$$[\mathrm{Var}_{n-1}(\Delta S_n)]\gamma_n^* = E_{n-1}((\Delta H_n \Delta S_n) \qquad \text{P- a.s.,} \qquad (2.2)$$

Proof: At first let us notice that the right side of (2.1) is finite. In fact, with $\gamma_n = 1$ for all n, we have

$$E_Q(H - c - G_N(\gamma))^2 = E_Q\left(H - c - \sum_{n=1}^{N}\sum_{j=1}^{d}\delta S_n^j\right)^2 < \infty.$$

Furthermore, we shall prove that $\gamma^*\Delta S_n$ is integrable square under Q.

Recall that (see [Appendix A]) if $Y, X_1, X_2, \cdots, X_d$ are $d+1$ integrable square random variables with $E(Y) = E(X_1) = \cdots = E(X_d) = 0$ and if $\widehat{Y} = b_1X_1 + b_2X_2 + \cdots + b_dX_d$ is the optimal linear predictor of Y on the basis of $X_1, X_2, \cdots, X_d$ then the vector $b = (b_1, b_2, \cdots, b_d)^T$ is the solution of the following equations system:

$$\mathrm{Var}(X)b = E(YX), \tag{2.3}$$

and as $\mathrm{Var}(X)$ is non-degenerated b is defined by

$$b = [\mathrm{Var}(X)]^{-1}E(YX), \tag{2.4}$$

and in all cases

$$bE(YX) \leq E(Y^2), \tag{2.5}$$

where $X = (X_1, X_2, \cdots, X_k)^T$.

Furthermore,

$$Y - \widehat{Y}\perp, \text{ i.e. } E[X_i(Y - \widehat{Y})] = 0,\ i = 1, \cdots, k. \tag{2.6}$$

Applying the above results to the problem of conditional linear prediction of ΔH_n on the basis of $\Delta S_n^1, \Delta S_n^2, \cdots, \Delta S_n^d$ as F_n is given we obtain from (2.3) the formula (2.2) defining the regression coefficient vector γ^*. On the other hand we have from (2.3) and (2.5):

$$\begin{aligned} E(\gamma_n^*\Delta S_n)^2 &= EE_{n-1}(\gamma_n^*\Delta S_n^T\Delta S_n\gamma_n^{*T}) = E(\gamma_n^*\mathrm{Var}_{n-1}(\Delta S_n)\gamma_n^{*T}) \\ &= E(\gamma_n^*E_{n-1}(\Delta H_n\Delta S_n^T)) \leq E(\Delta H_n)^2 < \infty. \end{aligned}$$

With the above remarks we can consider only, with no loss of generality, trading strategies γ_n such that

$$E_{n-1}(\gamma_n\Delta S_n)^2 < \infty.$$

We have:

$$H_N = H_0 + \Delta H_1 + \cdots + \Delta H_N$$

and

$$E_{n-1}(\Delta H_n - \gamma_n^T \Delta S_n)^2 = E_{n-1}(\Delta H_n)^2 - 2\gamma_n^T E_{n-1}((\Delta H_n \Delta S_n) \\ + \gamma_n^T E_{n-1}(\Delta S_n \Delta S_n^T)\gamma_n.$$

This expression takes the minimum value when $\gamma_n = \gamma_n^*$.

Furthermore, since $\{H_n - c - G_n(\gamma))$ is an $\{F_n\}$- integrable square martingale under Q,

$$E_Q(H_N - c - G_N(\gamma))^2 = E_Q\left[H_0 - c - \sum_{n=1}^{N}(\Delta H_n - \gamma_n \Delta S_n)\right]^2$$

$$= (H_0 - c)^2 + E_Q\left[\sum_{n=1}^{N}(\Delta H_n - \gamma_n \Delta S_n)\right]^2$$

$$= (H_0 - c)^2 + \sum_{n=1}^{N} E_Q(\Delta H_n - \gamma_n \Delta S_n)^2$$

(for $\Delta H_n - \gamma_n \Delta S_n$ is martingale difference)

$$= (H_0 - c)^2 + E_Q \sum_{n=1}^{N} E_{n-1}(\Delta H_n - \gamma_n \Delta S_n)^2$$

$$\geq (H_0 - c)^2 + E_Q \sum_{n=1}^{N} E_{n-1}(\Delta H_n - \gamma_n^* \Delta S_n)^2$$

$$= (H_0 - c)^2 + E_Q \sum_{n=1}^{N}(\Delta H_n - \gamma_n^* \Delta S_n)^2$$

$$= (H_0 - c)^2 + E_Q\left[\sum_{n=1}^{N}(\Delta H_n - \gamma_n^* \Delta S_n)\right]^2$$

$$\geq E_Q(H_N - H_0 - G_n(\gamma^*))^2.$$

So $E_Q(H_N - c - G_N(\gamma))^2 \geq E_Q(H_N - H_0 - G_n(\gamma^*))^2$ and the inequality become the equality if $c = H_0$ and $\gamma = \gamma^*$. □

3. Martingale representation theorem

Theorem 3.1. *Let* $\{H_n,\ n = 0,1,2,\cdots\}$, $\{S_n,\ n = 0,1,2,\cdots\}$ *be arbitrary integrable square random variable defined on the same probability space* $\{\Omega, \Im, P\}$, $F_n^S = \sigma(S_0,\cdots,S_n)$. *Denoting by* $\Pi(S,P)$ *the set of probability measures* Q *such that* $Q \sim P$ *and that* $\{S_n\}$ *is* $\{F_n^S\}$ *integrable square martingale under* Q, *then if* $F = \bigvee_{n=0}^{\infty} F_n^S$, $H_n,\ S_n \in L_2(Q)$ *and if* $\{H_n\}$ *is also a martingale under* Q *we have:*

$$1.\ H_n = H_0 + \sum_{k=1}^{n} \gamma_k \Delta S_k + C_n \qquad \text{a.s.,} \tag{3.1}$$

where $\{C_n\}$ *is a* $\{F_n^S\}-Q$*-martingale orthogonal to the martingale* $\{S_n\}$, i.e. $E_{n-1}((\Delta C_n \Delta S_n) = 0$, *for all* $n = 0,1,2,\cdots$, *whereas* $\{\gamma_n\}$ is $\{F_{n-1}^S\}$*- predictable.*

$$2.\ H_n = H_0 + \sum_{k=1}^{n} \gamma_k \Delta S_k := H_0 + G_n(\gamma) \qquad \text{P-a.s..} \tag{3.2}$$

for all n *finite iff the set* $\Pi(S,P)$ *consists of only one element.*

Proof: According to the proof of Theorem 2.1, putting

$$\Delta C_k = \Delta H_k - \gamma_k^* \Delta S_k,\ C_n = \sum_{k=1}^{n} \Delta C_k,\ C_0 = 0, \tag{3.3}$$

then $\Delta C_k \perp \Delta S_k$, by (2.6).

Taking summation of (3.3) we obtain (3.1).

The conclusion 2 follows from the fundamental theorem of financial mathematics. □

Notice 1. By the fundamental theorem of financial mathematics a security market has no arbitrage opportunity and is complete iff $\Pi(S,P)$ consists of the only element and in this case we have (3.2) with γ defined by (2.2). Furthermore, in this case the conditional probability distribution of S_n given F_{n-1}^S concentrates at most $d+1$ points of R^d (see [2], [3]), in particular for $d = 1$, with exception of binomial or generalized binomial market models (see [2], [7]), other models are incomplete.

Notice 2. We can choose the risk neutral martingale probability measure Q so that Q has minimum entropy in $\Pi(S,P)$ as in [2] or Q is near P as much as possible.

Example 3.1. Let us consider a stock with the discounted price S_0 at $t = 0$, S_1 at $t = 1$, where

$$S_1 = \begin{cases} 4S_0/3 & \text{with prob. } p_1 \\ S_0 & \text{with prob. } p_2, \quad p_1, p_2, p_3 > 0, \quad p_1 + p_2 + p_3 = 1 \\ 5S_0/6 & \text{with prob. } p_3 \end{cases}$$

Suppose that there is an option on the above stock with the maturity at $t = 1$ and with strike price $K = S_0$. We shall show that there are several probability measures $Q \sim P$ such that $\{S_0, S_1\}$ is, under Q, a martingale or equivalently $E_Q(\Delta S_1) = 0$.

In fact, suppose that Q is a probability measure such that under Q S_1 takes the values $4S_0/3$, S_0, $2S_0/3$ with positive probability q_1, q_2, q_3, respectively. Then $E_Q(\Delta S_1) = 0 \Leftrightarrow S_0(q_1/3 - q_3/6) = 0 \Leftrightarrow 2q_1 = q_3$, so Q is defined by $(q_1,\ 1 - 3q_1,\ 2q_1)$, $0 < q_1 < 1/3$.

In the above market, the payoff of the option is

$$H = (S_1 - K)_+ = (\Delta S_1)_+ = \max(\Delta S_1, 0).$$

It is easy to get a Q-optimal portfolio

$$\gamma^* = E_Q[H\Delta S_1]/E_Q(\Delta S_1)^2 = 2/3, \ E_Q(H) = q_1 S_0/3,$$

$$E_Q[H - E_Q(H) - \gamma^* \Delta S_1]^2 = q_1 S_0^2 (1 - 3q_1)/9 \to 0 \text{ as } q_1 \to 1/3.$$

However we can not choose $q_1 = 1/3$, because $q = (1/3,\ 0,\ 2/3)$ is not equivalent to P. It is better to choose $q_1 \cong 1/3$ and $0 < q_1 < 1/3$.

Example 3.2 Let us consider a market with one risky asset defined by:

$$S_n = S_0 \Pi Z_i, \text{ or } S_n = S_{n-1} Z_n, n = 1, 2, \cdots, N,$$

where $Z_1, Z_2, \cdots, Z_N$ are the sequence of i.i.d. random variables taking the values in the set $\Omega = \{d_1, d_2, \cdots, d_M)$ and $P(Z_i = d_k) = p_k > 0$, $k = 1, 2, \cdots, M$. It is obvious that a probability measure Q is equivalent to P and under Q $\{S_n\}$ is a martingale if and only if $Q\{Z_i = d_k) = q_k > 0$, $k = 1, 2, \cdots, M$ and $E_Q(Z_i) = 1$, i.e.

$$q_1 d_1 + q_2 d_2 + \cdots + q_M d_M = 1.$$

Let us recall the integral Hellinger of two measure Q and P defined on some measurable space $\{\Omega^*, F\}$:

$$H(P,Q) = \int_{\Omega^*} (dP \cdot dQ)^{1/2}.$$

In our case we have

$$\begin{aligned} H(P,Q) &= \sum \{P(Z_1 = d_{i1},\ Z_2 = d_{i2}, \cdots, Z_N = d_{iN}) \cdot \\ &\qquad Q(Z_1 = d_{i1},\ Z_2 = d_{i2}, \cdots, Z_N = d_{iN})\}^{1/2} \\ &= \sum \{p_{i1}q_{i1}\ p_{i2}q_{i2} \cdots p_{iN}q_{iN}\}^{1/2}, \end{aligned}$$

where the summation is extended over all $d_{i1}, d_{i2}, \cdots, d_{iN}$ in Ω or over all $i_1, i_2, \cdots, i_N$ in $\{1, 2, \cdots, M\}$. Therefore

$$H(P,Q) = \{\sum_{i=1}^{M} (p_i q_i)^{1/2}\}^N.$$

We can define a distance between P and Q by

$$||Q - P||^2 = 2(1 - H(P,Q)).$$

Then we want to choose Q^* in $\Pi(S,P)$ so that $||Q^* - P|| = \inf\{||Q - P|| : QP(S,P)\}$ by solving the following programming problem:

$$\sum_{i=1}^{M} p_i^{1/2} q_i^{1/2} \to \max$$

with the constraints :

a. $q_1 d_1 + q_2 d_2 + \cdots + q_M d_M = 1.$

b. $q_1 + q_2 + \cdots + q_M = 1.$

c. $q_1,\ q_2, \cdots,\ q_M > 0.$

Giving $p_1,\ p_2, \cdots,\ p_M$ we can obtain a numerical solution of the above programming problem. It is possible that the above problem has not a solution. However, we can replace the condition c) by the condition

c'. $q_1,\ q_2, \cdots,\ q_d \geq 0,$

then the problem has always the solution $q^* = (q_1^*,\ q_2^*, \cdots,\ q_M^*)$ and we can choose the probabilities $q_1, q_2, \cdots, q_M > 0$ are sufficiently near to $q_1^*, q_2^*, \cdots, q_M^*$.

4. Semi-continuous market model (discrete in time continuous in state)

Let us consider a financial market with two assets:

+ Free risk asset $\{B_n,\ n = 0, 1, \cdots, N\}$ with dynamics

$$B_n = \exp\left(\sum_{k=1}^{n} r_k\right),\ 0 < r_n < 1. \tag{4.1}$$

+ Risky asset $\{S_n,\ n = 0, 1, \cdots, N\}$ with dynamics

$$S_n = S_0 \exp\left(\sum_{k=1}^{n} [\mu(S_{k-1}) + \sigma(S_{k-1}) g_k]\right), \tag{4.2}$$

where $\{g_n,\ n = 0, 1, \cdots, N\}$ is a sequence of i.i.d. normal random variable $N(0, 1)$. It follows from (4.2) that

$$S_n = S_{n-1} \exp(\mu(S_{n-1}) + \sigma(S_{n-1}) g_n), \tag{4.3}$$

where S_0 is given and $\mu(S_{n-1}) := a(S_{n-1}) - 2(S_{n-1})/2$, with $a(x)$, $\sigma(x)$ being some functions defined on $[0, \infty)$.

The discounted price of risky asset $\tilde{S}_n = S_n/B_n$ is equal to

$$\tilde{S}_n = S_0 \exp\left(\sum_{k=1}^{n} [\mu(S_{k-1}) - r_k + \sigma(S_{k-1}) g_k]\right). \tag{4.4}$$

We try to find a martingale measure Q for this model.

It is easy to see that $E_P(\exp(\lambda g_k)) = \exp(\lambda^2/2)$, for $g_k \sim N(0, 1)$, hence

$$E \exp\left(\sum_{k=1}^{n} [\beta_k(S_{k-1}) g_k - \beta_k(S_{k-1})^2/2]\right) = 1, \tag{4.5}$$

for all random variable $\beta_k(S_{k-1})$.

Therefore, putting

$$L_n = \exp\left(\sum_{k=1}^{n} [\beta_k(S_{k-1}) g_k - \beta_k(S_{k-1})^2/2]\right),\ n = 1, \cdots, N \tag{4.6}$$

and if Q is a measure such that $dQ = L_N dP$ then Q is also a probability measure. Furthermore,

$$\frac{\tilde{S}_n}{\tilde{S}_{n-1}} = \exp(\mu(S_{n-1}) - r_n + \sigma(S_{n-1}) g_n). \tag{4.7}$$

Denoting by E^0, E expectation operations corresponding to P, Q, $E_n(.) = E[(.)|F_n^S]$ and choosing

$$\beta_n = -(a(S_{n-1}) - r_n)/\sigma(S_{n-1}) \tag{4.8}$$

then it is easy to see that

$$E_{n-1}[\tilde{S}_n/\tilde{S_{n-1}}] = E^0[L_n\tilde{S}_n/\tilde{S_{n-1}}|F_n^S]/L_{n-1} = 1$$

which implies that $\{\tilde{S}_n\}$ is a martingale under Q.

Furthermore, under Q, S_n can be represented in the form

$$S_n = S_{n-1}\exp((\mu^*(S_{n-1}) + \sigma(S_{n-1})g_n^*). \tag{4.9}$$

Where $\mu^*(S_{n-1}) = r_n - \sigma^2(S_{n-1})/2$, $g_n^* = -\beta_n + g_n$ is Gaussian $N(0,1)$. It is not easy to show the structure of $\Pi(S,P)$ for this model.

We can choose a such probability measure Q or the weight function L_N to find a Q- optimal portfolio.

Notice 3. The model (4.1), (4.2) is a type of discretization of the following diffusion model:

Let us consider a financial market with continuous time consisting of two assets:

+Free risk asset:

$$B_t = \exp\left(\int_0^t r(u)du\right). \tag{4.10}$$

+Risky asset : $dS_t = S_t[a(S_t)dt + \sigma(S_t)dWt]$, S_0 is given, where $a(.), \sigma(.) : (0,\infty) \to R$ such that $xa(x)$, $x\sigma(x)$ are Lipschitz. It is obvious that

$$S_t = \exp\left\{\int_0^t [a(S_u) - \sigma^2(S_u)/2]du + \int_0^t \sigma(S_u)dW_u\right\},\ 0 \le t \le T. \tag{4.11}$$

Putting

$$\mu(S) = a(S) - \sigma^2(S)/2, \tag{4.12}$$

and dividing $[0,T]$ into N intervals by the equidistant dividing points $0, \Delta,\ 2\Delta, \cdots,\ N\Delta$ with $N = T/\Delta$ sufficiently great, it follows from (4.10), (4.11) that

$$\begin{aligned} S_{n\Delta} &= S_{(n-1)\Delta} \exp\left\{ \int_{(n-1)\Delta}^{n\Delta} \mu(S_u)du + \int_{(n-1)\Delta}^{n\Delta} \sigma(S_u)dW_u \right\} \\ &\cong S_{(n-1)\Delta} \exp\{\mu(S_{(n-1)\Delta})\Delta + (S_{(n-1)\Delta})[W_{n\Delta} - W_{(n-1)\Delta}]\} \\ &\cong S_{(n-1)\Delta} \exp\{(S_{(n-1)\Delta})\Delta + \sigma(S_{(n-1)\Delta})\Delta^{1/2} g_n\} \end{aligned}$$

with $g_n = [W_{n\Delta} - W_{(n-1)\Delta}]/\Delta^{1/2},\ n = 1, \cdots,\ N$, being a sequence of the i.i.d. normal random variables of the law $N(0,1)$, so we obtain the model:

$$S^*_{n\Delta} = S^*_{(n-1)\Delta} \exp\{(S_{(n-1)\Delta})\Delta + (S_{(n-1)\Delta})\Delta^{1/2} g_n\}. \tag{4.13}$$

Similarly we have

$$B^*_{n\Delta} \cong B^*_{(n-1)\Delta} \exp(r_{(n-1)\Delta}\Delta). \tag{4.14}$$

According to (4.10), the discounted price of the stock S_t is

$$\tilde{S}_t = \frac{S_t}{B_t} = S_0 \exp\left\{ \int_0^t [\mu(S_u) - r_u]d_u + \int_0^t \sigma(S_u)dW_u \right\}. \tag{4.15}$$

By Girsanov Theorem, the unique probability measure Q under which $\{\tilde{S}_t,\ F_t^S,\ Q\}$ is a martingale is defined by

$$(dQ/dP)|F_T^S = \exp\left(\int_0^T \beta_u dW_u \ - \frac{1}{2} \int_0^T \beta_u^2 d_u \right) := L_T(\omega), \tag{4.16}$$

where

$$\beta_s = -\frac{((a(S_s) - r_s)}{\sigma(S_s)},$$

and $(dQ/dP)|F_T^S$ denotes the Radon-Nikodym derivative of Q w.r.t. P limited on F_T^S. Furthermore, under Q

$$W_t^* = W_t + \int_0^t \beta_u du$$

is a Wiener process. It is obvious that L_T can be approximated by

$$L_N := \exp\left(\sum_{k=1}^{N} \beta_k \Delta^{1/2} g_k - \Delta\beta_k^2/2\right), \tag{4.17}$$

where

$$\beta_n = -\frac{[a(S_{(n-1)\Delta}) - r_{n\Delta}]}{\sigma(S_{(n-1)\Delta})}. \tag{4.18}$$

Therefore the weight function (4.14) is approximate to Radon-Nikodym derivative of the risk unique neutral martingale measure Q w.r.t. P and Q is used to price derivatives of the market.

Notice 4. In the market model Black- Scholes we have $L_N = L_T$. We want to show now that for the weight function (4.17)

$$E_Q(H - H_0 - G_N(\gamma^*))^2 \to 0 \text{ as } N \to \infty \text{ or } \Delta \to 0.$$

Proposition Suppose that $H = K(S_T)/B_N$ is a integrable square discounted contingent claim. Then

$$E_Q(H - H_0 - G_N(\gamma^*))^2 \to 0 \text{ as } N \to \infty \text{ or } \Delta \to 0, \tag{4.19}$$

provided a and σ are constant (in this case the model (4.10), (4.11) is the model Black-Scholes).

Proof: It is well known (see[4],[5]) that for the model of complete market (4.10), (4.11) there exists a trading strategy $\varphi = (\varphi_t = \varphi(t, S(t)),\ 0 = t = T)$, hedging H, where $\varphi : [0,T] \times (0,\infty) \to R$ is continuously derivable in t and S, such that

$$H(S_T) = H_0 + \int_0^T \varphi_t d\tilde{S}(t) \qquad \text{a.s.}$$

On the other hand we have

$$
\begin{aligned}
&E_{Q_N}\left(H - H_0 - \sum_{k=1}^{N} \gamma^*_{(k-1)\Delta}\Delta\tilde{S}_{n\Delta}\right)^2 \\
&\leq E_{Q_N}\left(H - H_0 - \sum_{k=1}^{N} \phi_{(k-1)\Delta}\Delta\tilde{S}_{n\Delta}\right)^2 \\
&= E_Q\left(\int_0^T \phi_t d\tilde{S}(t) - \sum_{k=1}^{N} \phi_{(n-1)\Delta}\Delta\tilde{S}_{(n-1)\Delta}\right)^2 L_N/L_T \\
&= E_Q\left(\int_0^T \phi_t d\tilde{S}(t) - \sum_{k=1}^{N} \phi_{(k-1)\Delta}\Delta\tilde{S}_{(n-1)\Delta}\right)^2 \to 0 \text{ as } \Delta \to 0.
\end{aligned}
$$

(Since $L_N = L_T$ and by the definition of the stochastic integral Ito as a and σ are constant.) □

A Appendix A

Let $Y, X_1, X_2, \cdots, X_d$ be integrable square random variables defined on the same probability space $\{\Omega, F, P\}$ such that $EX_1 = \cdots = EX_d = EY = 0$.

We try to find a coefficient vector $b = (b_1, \cdots, b_d)^T$ so that

$$E(Y - b_1X_1 - \cdots - b_dX_d)^2 = E(Y - b^TX)^2 = \min_{a \in R^d}(Y - a^TX)^2. \quad \text{(A.1)}$$

Let us denote $EX = (EX_1, \cdots, EX_d)$, $\mathrm{Var}(X) = [\mathrm{Cov}(X_i, X_j),\ i, j = 1, 2, \cdots, d] = EXX^T$.

Proposition The vector b minimizing $E(Y - a^TX)^2$ is a solution of the following equation system:

$$\mathrm{Var}(X)b = E(XY). \quad \text{(A.2)}$$

Putting $U = Y - b^T X = Y - \hat{Y}$, with $\hat{Y} = b^T X$, then

$$E(U^2) = EY^2 - b^T E(XY) \geq 0. \tag{A.3}$$

$$E(UX_i) = 0 \text{ for all } i = 1, \cdots, d. \tag{A.4}$$

$$EY^2 = EU^2 + E\hat{Y}^2. \tag{A.5}$$

$$\rho = \frac{E\hat{Y}^2}{(EY^2 E\hat{Y}^2)^{1/2}} = \left(\frac{E\hat{Y}^2}{EY^2}\right)^{1/2}. \tag{A.6}$$

(ρ is called multiple correlation coefficient of Y relative to X.)

Proof: Suppose at first that $\mathrm{Var}(X)$ is a positively definite matrix. For each $a \in R^d$ We have

$$F(a) = E(Y - a^T X)^2 = EY^2 - 2a^T E(XY) + a^T EXX^T a \tag{A.7}$$

$$F(a) = -2E(XY) + 2\mathrm{Var}(X)a.$$

$$\left[\frac{\partial F(a)}{\partial a_i \partial a_j},\ i, j = 1, 2, \cdots, d\right] = 2\mathrm{Var(X)}.$$

It is obvious that the vector b minimizing $F(a)$ is the unique solution of the following equation:

$$F(a) = 0 \text{ or (A.2)}$$

and in this case (A.2) has the unique solution:

$$b = [\mathrm{Var}(X)]^{-1} E(XY).$$

We assume now that $1 \leq \mathrm{Rank}(\mathrm{Var}(X)) = r < d$.

We denote by $e_1, e_2, \cdots, e_d$ the ortho-normal eigenvectors w.r.t. the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_d$ of $\mathrm{Var}(X)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_d$ and P is a orthogonal matrix with the columns being the eigenvectors $e_1, e_2, \cdots, e_d$, then we obtain:

$$\mathrm{Var}(X) = P\Lambda P^T, \text{ with } \Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \cdots, \lambda_d).$$

Putting

$$Z = P^T X = [e_1^T X, e_2^T X, \cdots, e_d^T X]^T,$$

Z is the principle component vector of X, we have

$$\mathrm{Var}(Z) = P^T \mathrm{Var}(X) P = \Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \cdots, \lambda_r, 0, \cdots, 0).$$

Therefore

$$EZ_{r+1}^2 = \cdots = EZ_d^2 = 0, \text{ so } Z_{r+1} = \cdots = Z_d = 0 \ \text{ P- a.s.}$$

Then

$$\begin{aligned} F(a) &= E(Y - a^T X)^2 = E(Y - (a^T P)Z)^2 \\ &= E(Y - a_1^* Z_1 - \cdots - a_d^* Z_d)^2 \\ &= E(Y - a_1^* Z_1 - \cdots - a_r^* Z_r)^2, \end{aligned}$$

where

$$a^{*T} = (a_1^*, \cdots, a_d^*) = a^T P, \quad \mathrm{Var}(Z_1, \cdots, Z_r) = \mathrm{Diag}(\lambda_1, \lambda_2, \cdots, \lambda_r) > 0.$$

According to the above result $(b_1^*, \cdots, b_r^*)^T$ minimizing $E(Y - a_1^* Z_1 - \cdots - a_r^* Z_r)^2$ is the solution of

$$\begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_r \end{pmatrix} \begin{pmatrix} b_1^* \\ \dots \\ b_r^* \end{pmatrix} = \begin{pmatrix} EZ_1 Y \\ \dots \\ EX_r Y \end{pmatrix} \tag{A.8}$$

or

$$\begin{pmatrix} \lambda_1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \lambda_r & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} b_1^* \\ \dots \\ b_r^* \\ b_{r+1}^* \\ \dots \\ b_d^* \end{pmatrix} = \begin{pmatrix} EZ_1 Y \\ \dots \\ EZ_r Y \\ 0 \\ \dots \\ 0 \end{pmatrix} = \begin{pmatrix} EZ_1 Y \\ \dots \\ EZ_r Y \\ EZ_{r+1} Y \\ \dots \\ EZ_d Y \end{pmatrix} \tag{A.9}$$

with $b_{r+1}^*, \cdots, b_d^*$ arbitrary.

Let $b = (b_1, \cdots, b_d)^T$ be the solution of $b^T P = b^{*T}$, hence $b = Pb^*$ with b^* being a solution of (A.9). Then it is follows from (A.9) that

$$\mathrm{Var}(Z) P^T b = E(ZY) = P^T E(XY)$$

or

$$P^T\mathrm{Var}(X)PP^Tb = P^TE(XY) \text{ (since } \mathrm{Var}(Z) = P^T\mathrm{Var}(X)P)$$

or

$$\mathrm{Var}(X)b = E(XY)$$

which is (A.2). Thus we have proved that (A.2) has always a solution, which solves the problem (A.1).

By (A.7), we have

$$\begin{aligned} F(b) &= \min_a E(Y - a_TX)^2 \\ &= EY^2 - 2b^TE(XY) + b^T\mathrm{Var}(X)b \\ &= EY^2 - 2b^TE(XY) + b^TE(XY) \\ &= EY^2 - b^TE(XY) \geq 0. \end{aligned}$$

On the other hand

$$EUX_i = E(X_iY) - E(X_ib^TX) = 0, \tag{A.10}$$

since b is a solution of (A.2) and (A.10) is the ith equation of the system (A.2).

It follows from (A.10) that

$$E(U\hat{Y}) = 0 \text{ and } EY^2 = E(U+\hat{Y})^2 = EU^2+E\hat{Y}^2+2E(U\hat{Y}) = EU^2+E\hat{Y}^2.$$

□

Example We can use Hilbert space method to prove the above proposition. In fact, let H be the set of all random variables ξ's such that $E\xi = 0$, $E\xi^2 < \infty$, then H becomes a Hilbert space with the scalar product $(\xi, \zeta) = E\xi\zeta$, and with the norm $||\xi|| = (E\xi^2)^{1/2}$. Suppose that $X_1, X_2, \cdots, X_d$, $Y \in H$, L is the linear manifold generated by $X_1, X_2, \cdots, X_d$. We want to find a $\hat{Y} \in H$ so that $||Y - \hat{Y}||$ minimizes, that means $\hat{Y} = b^TX$ solves the problem (A.1). It is obvious that $\hat{Y}$ is defined by

$$\hat{Y} = \mathrm{Proj}_L Y = b^TX \text{ and } U = \hat{Y} - Y \in L.$$

Therefore $(Y - b^TX, X_i) = 0$ or $E(b^TXX_i) = E(X_iY)$ for all $i = 1, \cdots, d$ or $b^TE(X^TX) = E(XY)$ which is the equation (A.2). The rest of the above proposition is proved similarly.

References

[1] Follmer H.,Schweiser M. *Hedging of contingent claim under incomplete information.* App.Stochastic Analysis, Edited by M.Davisand, R.Elliot, London, Gordan&Breach, 1999, pp.389-414.

[2] Follmer H., Schied A. *Stochastic Finance. An introduction in discrete time.* Walter de Gruyter, Berlin- New York, 2004.

[3] Jacod J., Shiryaev A.N. *Local martingales and the fundamental asset pricing theorem in the discrete case.* Finance Stochastic **2**, pp. 259-272.

[4] Harrison, M.J. and Kreps, D.M.,*Martingales and arbitrage in multiperiod securities markets*, J. of Economic Theory, **29** (1979), pp. 381-408.

[5] Harrison, M.J.and Pliska S.R., *Martingales and stochastic integrals in theory of continuous trading, Stochastic Processes and their Applications*, **11** (1981), pp.216-260.

[6] Lamberton, D. and Lapayes,B. *Introduction to Stochastic Calculus Applied in Finance.* Chapman&Hall/CRC, 1996

[7] Nechaev M.L. *On mean -Variance hedging. Proceeding of Workshop on Math.* Finance, May 18-19, 1998. Institute Franco-Russe Liapunov, Ed. by A.Shiryaev, A. Sulem.

[8] Nguyen Van Huu, Tran Trong Nguyen.*On a generalized Cox-Ross-Rubinstein option market model.* Acta Math. Vietnamica, **26** (2001), No**2**,187-204.

[9] Schweiser M. *Variance -optimal hedging in discrete time.* Mathematics of Operation Research, **20** (1995), pp.1-32.

[10] Schweiser M.*Approximation pricing and the variance-optimal martingale measure.* The Annals of Prob. **24**(1966), 1, pp.206-236.

[11] Schal M.*On quadratic cost criteria for option hedging.* Mathematics of Operation Research, **19**(1994), pp. 131-141.

Flexible Stability Domains for Explicit Runge-Kutta Methods

Rolf Jeltsch

Manuel Torrilhon

ETH Zurich,

Seminar for Applied Mathematics, 8092 Zurich, Switzerland.

E-mail: {jeltsch,matorril}@math.ethz.ch

Abstract

Stabilized explicit Runge-Kutta methods use more stages which do not increase the order, but, instead, produce a bigger stability domain for the method. In that way stiff problems can be integrated by the use of simple explicit evaluations for which usually implicit methods had to be used. Ideally, the stability domain is adapted precisely to the spectrum of the problem at the current integration time in an optimal way, i.e., with minimal number of additional stages. This idea demands for constructing Runge-Kutta methods from a given family of flexible stability domain.

In this paper we discuss typical families of flexible stability domains, like a disk, real interval, imaginary interval, spectral gap and thin regions, and present corresponding essentially optimal stability polynomials from which a Runge-Kutta method can be constructed. We present numerical results for the thin region case.

1 Introduction

Explicit Runge-Kutta-methods are popular for the solution of ordinary and partial differential equations as they are easy to implement, accurate and cheap. However, in many cases like stiff equations they suffer from time step conditions, which may become so restrictive, that they render explicit methods useless. The common answer to this problem is the usage of implicit methods which often show unconditional stability. The trade-off for implicit methods is the requirement to solve a large, possibly non-linear, system of equations in each time step. See the text books of Hairer and Wanner [3] and [4] for an extensive introduction to explicit and implicit methods.

An interesting approach to combine both implicit and explicit methods stabilizes the explicit methods by increasing the number of internal

explicit stages. These stages are chosen such that the stability condition of the explicit method is improved. As a result, these methods are very easy to implement. The additional function evaluations that are necessary in the additional stages can be viewed as iteration process yielding a larger time step. In that sense an iterative method which is needed to solve the non-linear system in an implicit method can be compared to the higher number of internal stages in a stabilized explicit method. Note, however, that the stabilized explicit method has no direct analog in terms of iterative solvers for non-linear equations. Hence, they form a new approach.

Typically, an implicit method is A-stable and the stability domain includes the entire negative complex plane. But in applications the spectrum often covers only a specific fraction of the negative complex plane. Clearly, an A-stable implicit method would integrate also such a problem, but an explicit method with a stability domain specialized to this specific fraction might do it in an easier and more efficient way. This is the paradigm of stabilized explicit Runge-Kutta methods.

In an ideal future method the spectrum of the problem is analyzed every few time steps and an explicit Runge-Kutta method would be constructed in real time such that the current spectrum is optimally included with a minimal number of stages. This opens the question how to find a Runge-Kutta method for a given shape of the stability domain. This question can not be answered for general shapes of the stability domain. Instead, we have to restrict ourselves to classes or families of shapes.

This paper discusses various classes of flexible shapes and the resulting optimal Runge-Kutta stability polynomials. For simplicity the shapes may change only according to a real parameter. A classical case is the length of a maximal real interval. Runge-Kutta methods that include a maximal interval of the negative real line have been constructed in many works starting with van der Houwen and Sommeijer [5] and later Lebedev [10]. For detailed references see the text books [4] by Hairer and Wanner and [6] by Hundsdorfer and Verwer. In this paper we also discuss the case of a maximal disk touching the origin with a first, second and third order method. Furthermore, a maximal symmetric interval on the imaginary axis, a spectral gap with maximal width and distance, and a maximal thin region. For each family of shapes we investigate optimal or essentially optimal stability polynomials. These polynomials are the starting point from which an corresponding explicit Runge-Kutta method can be constructed relatively easily. Furthermore, we briefly formulate possible applications in which the respective shapes of spectra occur.

The case of maximal thin regions have been introduced and investigated in [15] by the authors of the present paper. Fully grown essen-

tially optimal Runge-Kutta methods have been constructed and applied to hyperbolic-parabolic partial differential equations. In Sec. 7 and 8 of this paper we review the results and some numerical experiments for the maximal thin region case. The example code for an advection-diffusion equation together with the data of the optimized stability polynomials are available online through [14].

2 Explicit Runge-Kutta methods

We will consider explicit Runge-Kutta methods for the numerical solution of an ordinary differential equation

$$y'(t) = F(y(t)) \tag{1}$$

with $y : \mathbb{R}^+ \to V \subset \mathbb{R}^N$ with $y(0) = y_0$. An extensive presentation and investigation of Runge-Kutta methods can be found in the textbooks [3] and [4]. The stability function of a p-th order, s-stage explicit Runge-Kutta method is a polynomial in the form

$$f_s(z) = 1 + \sum_{k=1}^{p} \frac{z^k}{k!} + \sum_{k=p+1}^{s} \alpha_k \, z^k \tag{2}$$

with $p \leq s$. We call p the order of $f_s(z)$. The stability domain of the method is given by

$$S(f_s) = \{z \in \mathbb{C} \mid |f_s(z)| \leq 1\}. \tag{3}$$

If the method is applied to the ordinary differential equation (1) with a certain time step Δt, the set of the scaled eigenvalues of the Jacobian of F with negative real part

$$G^{(\Delta t)} = \{\Delta t \, \lambda \in \mathbb{C} \mid \lambda \text{ eigenvalue of } DF(y),\ \operatorname{Re}\lambda \leq 0,\ y \in V\} \tag{4}$$

has to be included in the stability domain of the method in order to assure stability. This is referred to as linear stability of the method.

2.1 Optimal stability

Suppose the order p of the method is fixed, then for $s > p$ the remaining coefficients of the stability polynomial (2) can be viewed as parameters which control the shape of the stability domain. For a given equation (1) and time step Δt the problem of an optimally stable explicit method can be formulated as:

$$\text{Find } \{\alpha_k\}_{k=p+1}^{s} \text{ for minimal } s, \text{ such that } G^{(\Delta t)} \subset S(f_s). \tag{5}$$

The coefficients are used to adapt the stability domain to a fixed set of eigenvalues of DF. In many cases the set of eigenvalues changes shape according to a real parameter $r \in \mathbb{R}$ which is not necessarily the time step. For example, the value r could be the length of a real interval or the radius of a disk. This paper considers the families of eigenvalue sets given by $G_r \subset \mathbb{C}$, $r \in \mathbb{R}$. We consider the following optimization problem:

Problem 1. *For fixed s and p find $\{\alpha_k\}_{k=p+1}^{s}$ for the largest r such that*

$$G_r \subset S(f_s) \tag{6}$$

and $f_s(z)$ given by (2).

Here, the number of stages as well as the order is fixed and both the shape of G_r and the coefficients of the stability polynomial are adapted to each other in order to obtain the maximal value of r. The maximal r is called $r_p^{(opt)}(s)$, that is

$$r_p^{(opt)}(s) = \max\left\{r \in \mathbb{R} \mid G_r \subset S\left(f_s\right),\ p \text{ order of } f_s\right\}. \tag{7}$$

In all families of G_r which we considered there existed an optimal f_s. It is clear that the result of this optimization of r is related to the optimization (5). The inversion of the relation $r_p^{(opt)}(s)$ which gives the maximal value of r for a number of stages s can be used to find the minimal number of stages for a given value of r.

2.2 Adaptive method construction

The stability polynomial is not equivalent to a single Runge-Kutta method. In general many different Runge-Kutta methods can be based on the same stability polynomial. All these methods would show the same fundamental overall stability properties.

The construction of actual Runge-Kutta methods from the stability polynomial is not the primary focus of this paper. Indeed, the problem to find optimal stability domains as in (6) affects only the polynomial. The method can be constructed afterwards. Once Runge-Kutta methods are found for the family of optimized stability polynomial, the relation (7) can be used to set up a spectrum-adaptive Runge-Kutta method.

In our setting, the spectrum G_r may change during the computation and this change is represented in different values of r. For adaptivity to the spectrum the relation (7) is inverted to give

$$s_p^{(opt)}(r) = \min\left\{s \in \mathbb{N} \mid r_p^{(opt)}(s) > r\right\}, \tag{8}$$

i.e. a optimal s for given spectrum G_r. In such a spectrum-adaptive calculation the time step may stay constant and instead the number

of stages would vary according to different situations of the spectrum. Each time step the value of r is examined and the number of stages $s = s_p^{(opt)}(r)$ fix an optimal polynomial f_s and corresponding Runge-Kutta method. This method will perform s stages, that are a minimal number of stages required for the respective spectrum G_r. In that sense, the original question (5) can be answered with the solution of (6).

3 Maximal real interval

The case in which G_r is a real interval

$$G_r = [-r, 0] \tag{9}$$

is considered in various papers, for instance in [1], [10], [5], etc, see also the discussion in the book [4], pp.31-36.

3.1 Application: diffusion equations

The case of a real interval is of particular interest when solving parabolic partial differential equations, like the diffusion equation

$$\partial_t u = D\,\partial_{xx} u, \quad x \in [a, b], \quad t > 0$$

where D is the diffusion coefficient. It is usually discretized in a semi-discrete fashion

$$\partial_t u_i = D\frac{u_{i-1} - 2u_i + u_{i+1}}{\Delta x^2}, \quad i = 1, 2, \ldots$$

In the periodic case, the discretization of the Laplacian yields negative eigenvalues in the interval $[-\frac{4D}{\Delta x^2}, 0]$. On fine grids with small grid size Δx this interval becomes very large. The result of the optimal stability polynomials depends on the required order p of the method. We will consider $p = 1, 2$.

3.2 1st order

Since the zeros of the stability polynomial $f_s(z)$ are included in the stability domain, it is obvious that an appropriate distribution of real zeros inside the interval $[-r, 0]$ will provide a maximal value of r. In between the real zeros the value of $|f_s(z)|$ should not exceed unity. On the interval $[-1, 1]$ the Chebyshev polynomials T_s are known to realize such a optimal distribution of zeros for a given degree s. Rescaling and shifting gives the stability polynomial

$$f_s(z) = T_s\left(\frac{z}{s^2} + 1\right) \tag{10}$$

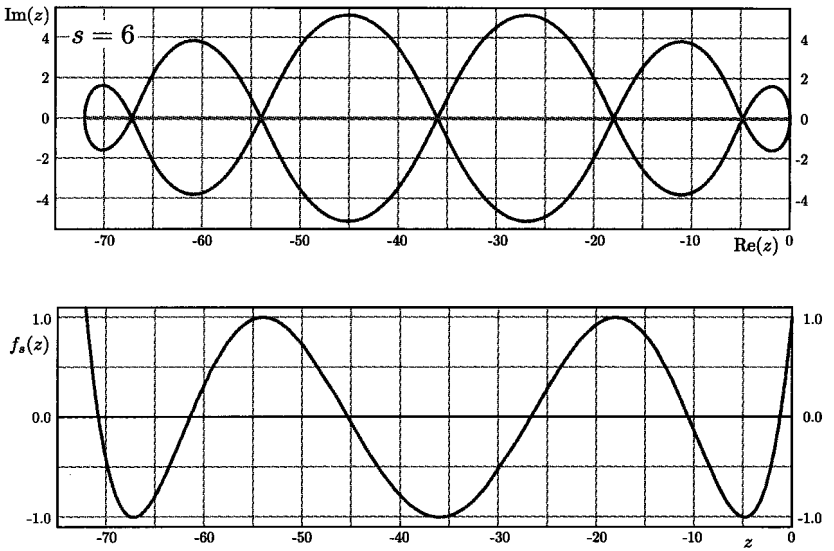


Figure 1: First order stability polynomial of degree $s = 6$ containing maximal real interval. In the first order case these polynomials are given by the Chebyshev polynomials T_s. Top: Boundary of the stability region. Bottom: $f_s(z)$ along the real axis.

and the optimal property

$$G_r \subset S(f_s) \quad \text{with } r = 2s^2. \tag{11}$$

Since we have $f_s'(0) = 1$, $f_s''(0) < 1$ the resulting Runge-Kutta method will be first order, $p = 1$. The rescaling of the Chebyshev polynomials T_s by s^2 essentially follows from the requirement of a method with at least first order of accuracy. The scaling value $T_s'(1) = s^2$ is the largest possible first derivative at $z = 1$ among all polynomials with $|f_s(z)| \leq 1$. This shows the optimality of the Chebyshev polynomials for a maximal real interval. However, higher order can not be obtained based on Chebyshev polynomials.

The stability domain $S(f_s)$ as well as the function $f_s(z)$ for real z are shown in Fig. 1 for the case $s = 6$. The shapes are perfectly symmetric and the interval $[-72, 0]$ is included in the stability domain. The points of the extrema of f_s along the real axis are boundary points of $S(f_s)$.

3.3 2nd order

In the second order case the stability polynomial has the form

$$f_s(z) = 1 + z + \frac{1}{2}z^2 + \sum_{k=3}^{s} \alpha_k\, z^k \tag{12}$$

$$= (1 + \beta_1\, z + \beta_2\, z^2) R_s\,(z) \tag{13}$$

which is written with parameters $\beta_{1,2}$ and $s-2$ zeros z_k in the factor

$$R_s(z) = \prod_{k=1}^{s-2} \left(1 - \frac{z}{z_k}\right). \tag{14}$$

The parameters $\beta_{1,2}$ are considered not to be free but to follow from the order conditions $f_s'(0) = f_s''(0) = 1$. If all the zeros in R_s are real, $z_k \in \mathbb{R}$, $k = 1, 2, \cdots, s-2$, it follows from the order conditions that the quadratic factor in (13) has always two complex conjugated zeros z_{s-1} and $z_s = \bar{z}_{s-1}$. Indeed, the discriminant reads

$$\beta_1^2 - 4\beta_2 = -(1 - \sum\nolimits_{k=1}^{s-2} \frac{1}{z_k})^2 - 2\sum\nolimits_{k=1}^{s-2} \frac{1}{z_k^2} \tag{15}$$

which is negative for real z_k and produces complex roots. Furthermore, for negative real zeros $z_k < 0$, $k = 1, 2, \cdots, s-2$, we have

$$|z_s| = \sqrt{\frac{2}{(1 - \sum_{k=1}^{s-2} \frac{1}{z_k})^2 + \sum_{k=1}^{s-2} \frac{1}{z_k^2}}} < \sqrt{2} \quad \text{and} \quad -1 < \mathrm{Re}\,(z_s) < 0, \tag{16}$$

hence, the two complex roots stay in the vicinity of the origin. Similar results can also be found in [2].

As in the first order case, the question is how to distribute real zeros z_k, $k = 1, 2, \cdots, s-2$, along the interval $[-r, 0]$ such that a maximal value of r is obtained. In this case, however, an analytical result is very involved, see [10] by Lebedev. Usually, the optimization which finds the polynomials has to be conducted numerically. Precise algorithms how to obtain the stability polynomials numerically are, for instance, given in the work [1] by Abdulle and Medovikov. The resulting stability domain satisfy

$$G_r \subset S(f_s) \quad \text{with } r \approx s^2. \tag{17}$$

Hence, the requirement of second order still allows a quadratic dependence of r on the number of stages s. However, the length is halved in comparison to the first order case.

The stability domain and polynomial for the case $s = 9$ are displayed in Fig. 2. The plots use the same axes range as in Fig. 1 for the first

order which showed $s = 6$. Comparison of $f_s(z)$ along the real line with the first order case shows that the current polynomial has a much denser zero distribution. The second order condition leads to a first minimum with positive function value in the interval $[-5, 0]$. This minimum corresponds to the two complex conjugated zeros. All the other extrema points correspond to points where the boundary of the stability domain touches the real axis.

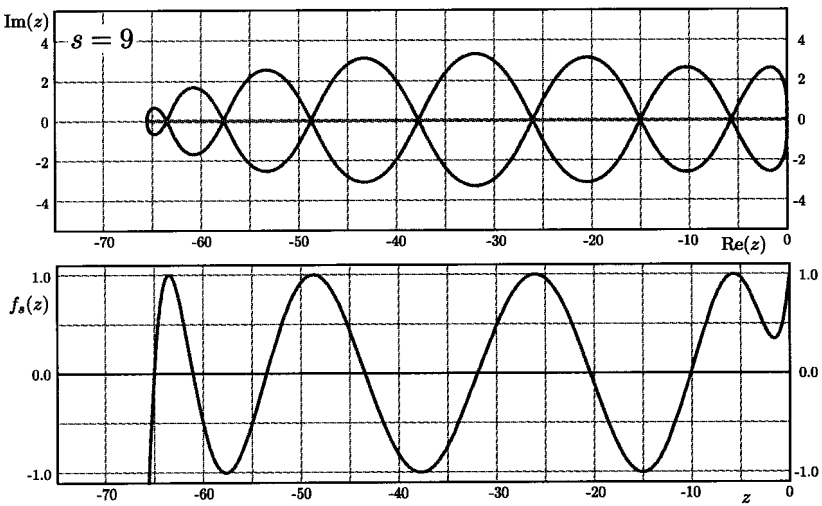


Figure 2: Second order stability polynomial of degree $s = 9$ containing maximal real interval. The second order condition introduces a minimum around $z = -2$ and two complex-conjugated roots which reduce the maximal possible real interval. Top: Boundary of the stability region. Bottom: $f_s(z)$ along the real axis.

4 Maximal disk

Another case of general interest is a stability domain which contains a maximal disk touching the origin. We define

$$G_r = \{z \in \mathbb{C} \mid |z + r| \leq r\} \tag{18}$$

for $r > 0$ which describes a disk in the complex plane with center at $(-r, 0)$ and radius r. Hence, the origin is a boundary point. The question of maximal contained disk is for example discussed by Jeltsch and Nevanlinna in [8] and [9].

4.1 Application: upwinded advection equation

The case of a disk is of particular interest when hyperbolic partial differential equations are solved with upwind methods. The advection equation

$$\partial_t u + a\,\partial_x u, \quad x \in [a,b], \quad t > 0$$

with advection velocity a is a typical example. The classical upwind method for this equation reads in a semi-discrete form

$$\partial_t u_i = a\frac{u_{i-1} - u_i}{\Delta x}, \quad i = 1, 2, \cdots$$

Here, again for periodic functions, the eigenvalues are situated on the circle $\frac{a}{\Delta x}(\exp(\mathbf{i}\,\varphi) - 1)$ with $\varphi \in [0, 2\pi]$. This circle represents the boundary of G_r with $r = a/\Delta x$. Again, the result of the optimal stability polynomials depends on the required order p of the method and we consider only $p = 1, 2, 3$.

4.2 1st order

The stability domain of the polynomial

$$f_s(z) = \left(\frac{z}{s} + 1\right)^s \tag{19}$$

has the shape of the disk G_s, hence we have

$$G_r = S(f_s) \quad \text{for } r = s. \tag{20}$$

The optimality follows for instance from the comparison theorem of Jeltsch and Nevanlinna [9], see also the text book [4]. According to this theorem no two stability domains with equal number of stages are contained in each other. Since, $S(f_s)$ is the disk G_s no other stability domain with s stages will contain this or a larger disk. The order conditions are given by $f_s'(0) = 1, \quad f_s''(0) < 1$, so we have a first order method.

Considering the zeros of f_s this polynomial exhibits the greatest possible symmetry since there is only one zero of multiplicity s located at the center of G_r. Obviously, this provides a value of $|f_s(z)|$ smaller than unity for a maximal radius.

Note, that the first order result does not bring any gain in efficiency since the first order s-stage method is equivalent to s simple Euler steps. This is slightly different when it comes to higher order methods.

4.3 2nd order

We briefly re-derive the stability polynomial containing a maximal disk in the second order case. This case was studied by Owren and Seip in [13].

According to the discussion of the second order case in Sec. 3.3 any second order stability polynomial has at least one complex conjugated pair of roots. Thus, the perfectly symmetric solution of the first order case with an p-fold zero in the center of the maximal disk is not possible. Highest possible symmetry is now obtained by distributing the zeros symmetrically around the center of the disk. The polynomial

$$f(z) = \alpha z^s + \beta, \quad \alpha, \beta \in \mathbb{R},\ \alpha, \beta > 0 \tag{21}$$

has s zeros symmetrically around the origin in the corners of a regular s-gon. The condition $\left|f\left(r\,e^{\mathrm{i}\varphi}\right)\right| \leq 1$ for an unknown radius r yields

$$\left|\alpha r^s e^{\mathrm{i}\,s\varphi} + \beta\right| \leq \alpha r^s + \beta \overset{!}{=} 1 \tag{22}$$

which, together with the shifted order conditions $f(r) = f'(r) = f''(r) = 1$, gives explicit relations for r, α, and β in dependence of s. We find

$$\alpha = \frac{1}{s\,(s-1)^{s-1}}, \quad \beta = \frac{1}{s}, \quad r = s - 1 \tag{23}$$

and after shifting by r

$$f_s(z) = \frac{s-1}{s}\left(\frac{z}{s-1} + 1\right)^s + \frac{1}{s}. \tag{24}$$

This second order stability polynomial satisfies

$$G_r \subset S(f_s) \quad \text{with } r = s - 1 \tag{25}$$

in an optimal way. A rigorous proof can be found in [13].

Fig. 3 shows the stability domains for increasing s with $s = 2, 3, 4, 5, 6$ and the boundaries of the included disk. In accordance to the symmetry of the stability polynomial the domains have the shape of smoothened regular s-gons for $s \geq 3$. The middle points of the edges coincide with points of the disk G_{s-1}.

Note, that the comparison theorem of Jeltsch and Nevanlinna cannot be used here since the stability domain is not given by the disk itself. Furthermore, the maximal included disk is smaller than in the first order case.

For the second order case the methods for higher s are more efficient since the s-stage method requires s function evaluations for a second order time step for which $2(s-1)$ function evaluations of the simple second order 2-stage method are necessary. Hence, formally these methods are asymptotically twice as fast for a large time step.

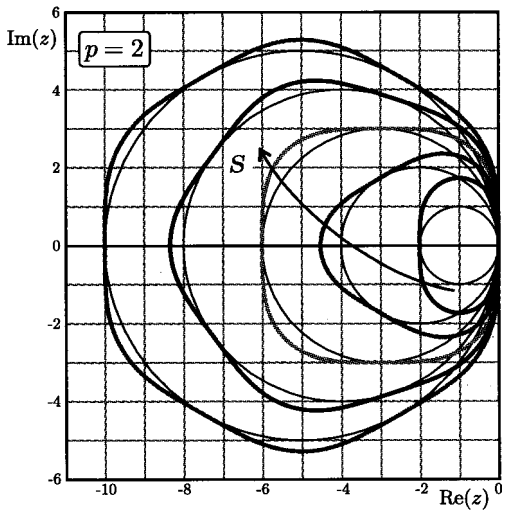


Figure 3: Optimal stability regions for a s-stages second order Runge-Kutta method including a largest possible disk with $s = 2, 3, 4, 5, 6$. The regions have the shapes of smoothened, regular s-edged polygons.

4.4 3rd order

The higher order case has also been studied in [13]. In the lower order cases above an increase of the number of stages by one also resulted in a larger disk with radius increased by one. This behavior extends to higher order so that an p-order, s-stage method allows a maximal disk of radius $r = s - p + 1$, at least asymptotically for large s.

Here, we present the polynomials for $p = 3$ and $s = 4, 5, 6$ for the maximal disk case. They have been constructed empirically to be essentially optimal. The general shape is

$$f_s(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \sum_{k=4}^{s} \alpha_k^{(s)} z^k \tag{26}$$

where the free coefficients have been fixed by specifying additional roots of f inside G_r. Again the highest symmetry yields the best result. The stability domains are depicted in Fig. 4 together with the maximal disk

included. The coefficients are given by

$$\begin{aligned} &\alpha_4^{(4)} = 0.023805 \\ &\alpha_4^{(5)} = 0.030651, \quad \alpha_5^{(5)} = 0.0022911 \\ &\alpha_4^{(6)} = 0.032771, \quad \alpha_5^{(6)} = 0.0034763, \quad \alpha_6^{(6)} = 0.00015648 \end{aligned} \tag{27}$$

and the possible radii of the included disks are found to be

$$r^{(3)} = 1.25, \quad r^{(4)} = 2.07, \quad r^{(5)} = 2.94, \quad r^{(6)} = 3.79. \tag{28}$$

While the cases $s = 3, 4$ exhibit a bigger radius than $s-p+1$, the higher stage methods do not reach this bound.

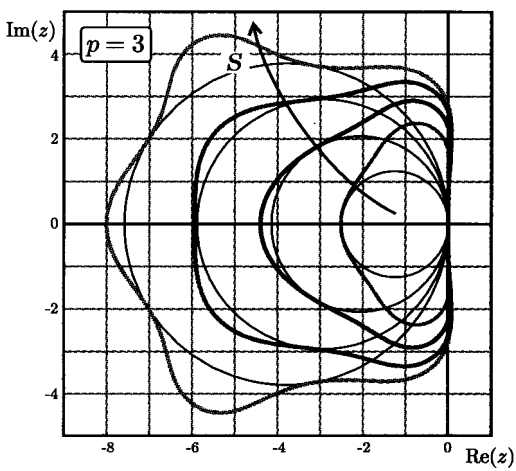


Figure4: Essentially optimal stability regions for a s-stages third order Runge-Kutta method including a largest possible disk with $s = 3, 4, 5, 6$. They have been found empirically. In the case $s = 5, 6$ the possible disk has a radius slightly smaller than $s - p + 1$.

5 Maximal imaginary interval

It is also possible to ask for a maximal interval on the imaginary axis to be included in the stability domain. We define

$$G_r = \{z \in \mathbb{C} \mid |\mathrm{Im}\, z| \leq r, \ \mathrm{Re}\, z = 0\} \tag{29}$$

for $r > 0$ which describes a symmetric section of the imaginary axis around the origin of length $2r$.

5.1 Application: central differences for advection

A purely imaginary spectrum arises when hyperbolic parabolic partial differential equations are discretized with fully symmetric stencils. In that case the advection equation

$$\partial_t u + a\,\partial_x u, \quad x \in [a,b], \quad t > 0$$

is turned into the semi-discrete equation

$$\partial_t u_i = a\frac{u_{i-1} - u_{i+1}}{2\Delta x}, \quad i = 1, 2, \cdots$$

For periodic functions, the eigenvalues are found in the interval $[-\frac{a}{\Delta x}\mathbf{i}, \frac{a}{\Delta x}\mathbf{i}]$ on the imaginary axis.

5.2 1st and 2nd order

A possible heuristic strategy to construct a stability domain that includes a large imaginary interval is to locate roots of the stability polynomial along the imaginary axis. A similar case is also discussed in the text book [4]. Since the coefficients of the polynomial need to be real, the imaginary roots have to occur in complex conjugated pairs. Furthermore, the order conditions can not be satisfied with purely imaginary roots, hence, an additional factor will be included in the polynomial.

The first order polynomial is defined for odd values of s and has the shape

$$f_s^{(1)}(z) = (1 + \alpha\, z) \prod_{k=1}^{(s-1)/2} \left(1 + (\frac{z}{z_k^{(s)}})^2\right) \tag{30}$$

with $(s-1)/2$ pairs of roots $\pm z_k^{(s)}\mathbf{i}$ (s odd). The coefficient α is fixed by the order condition $f_s'(0) = 1$. Similarly, we have for the second order polynomial

$$f_s^{(2)}(z) = (1 + \alpha\, z + \beta\, z^2) \prod_{k=1}^{(s-2)/2} \left(1 + (\frac{z}{z_k^{(s)}})^2\right) \tag{31}$$

with $(s-2)/2$ pairs of roots (s even). The conditions $f_s'(0) = f_s''(0) = 1$ define α and β.

These polynomials mimic the case of a maximal real interval where more and more roots are distributed on the real axis. However, in the imaginary case this approach is heuristic and might only be essentially optimal.

Here, we present the first cases $s = 3, 4, 5, 6, 7, 8$ for the first and second order polynomial, which have been constructed by trial and error.

Fig. 5 shows the respective stability domains. The roots which are placed along the imaginary axis are given by

$$\begin{aligned}
z_1^{(3)} &= 1.51 \\
z_1^{(5)} &= 1.65,\ z_2^{(5)} = 2.95 \\
z_1^{(7)} &= 1.73,\ z_2^{(7)} = 3.45,\ z_3^{(7)} = 4.36 \\
z_1^{(4)} &= 2.44 \\
z_1^{(6)} &= 2.81,\ z_2^{(6)} = 4.32 \\
z_1^{(8)} &= 2.95,\ z_2^{(8)} = 5.01,\ z_3^{(8)} = 6.04
\end{aligned} \tag{32}$$

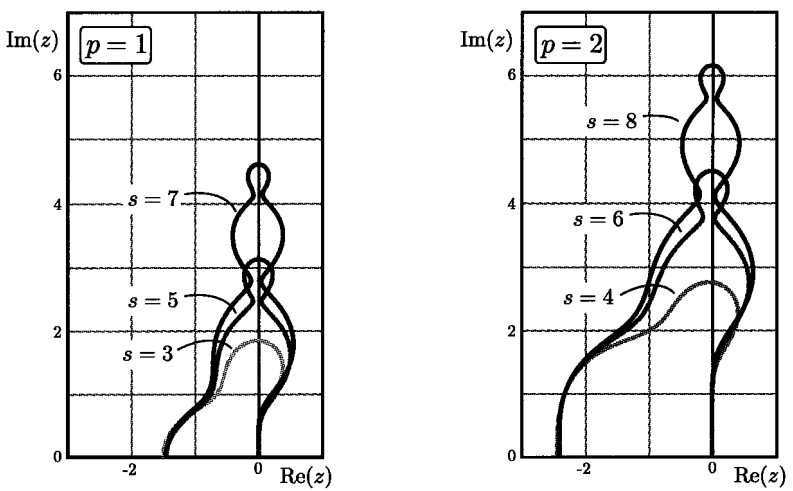


Figure 5: Stability regions that includes a maximized section of the imaginary axis. Left: first order, $s = 3, 5, 7$. Right: second order, $s = 4, 6, 8$. The respective polynomials follow the ansatz (30)/(31).

and the maximal extension along the imaginary axis is

$$r^{(3)} = 1.83, \quad r^{(5)} = 3.12,\ r^{(7)} = 4.51, \tag{33}$$

$$r^{(4)} = 2.79, \quad r^{(6)} = 4.47,\ r^{(8)} = 6.17. \tag{34}$$

Note, that in the case of a real interval we have $r \sim s^2$ and a quickly growing interval is included. Here, we find a clearly slower growth of the section with increasing s, presumably only linear.

6 Spectral gaps

Many real spectra come with gaps, that is, they decompose into two or more distinct intervals of specific widths. This represents scale separation in the respective application, since some phenomena happen on a distinctly faster time scale than others. This occurs in ODE systems of chemical kinetics, or molecular dynamics. A similar spectrum is found in discretizations of diffusion-reaction equations like

$$\partial_t u - D\,\partial_{xx} u = -\nu\, u \tag{35}$$

where the diffusive spectrum as given above is shifted along the real axis by the value ν.

Here, we are looking at the case of a spectrum in the form

$$G_{\delta,\lambda} = [-\lambda - \delta/2, -\lambda + \delta/2] \cup [-1, 0] \tag{36}$$

with two real positive numbers λ, δ. This spectrum has two real parts, one at the origin and one situated at $z = \lambda$ with symmetric width δ. In order two formulate an optimal stability domain for such a spectrum, we fix λ and ask for a stability polynomial which allows maximal width δ. Following the ideas of the sections above we construct a polynomial which allows to place roots in the vicinity of λ. Restricting ourselves to the second order case, the simplest idea is

$$f_s^{(2)}(z) = (1 + \alpha\, z + \beta\, z^2)(1 + \frac{z}{\lambda})^{s-2} \tag{37}$$

with $s \geq 3$. The order conditions $f_s'(0) = f_s''(0) = 1$ determine α and β. Here, one additional root is introduced at $-\lambda$ and all additional stages increase only the multiplicity of the root $-\lambda$. As a result the stability domain will allow bigger widths of the spectrum section around $-\lambda$.

Alternatively, it is possible to distribute additional roots around the value $-\lambda$ to allow increased widths. Again for $p = 2$ we write

$$f_s^{(2)}(z) = (1 + \alpha\, z + \beta\, z^2) \prod_{k=1}^{s-2} (1 + \frac{z}{\lambda + \Delta_k}) \tag{38}$$

with $s - 2$ adjustable constants Δ_k. For $\Delta_k = 0$ this form reduces to th case above with multiple roots at λ.

We continue to investigate four cases: The polynomial (37) with $s = 3$, 4 and 5, as well as the polynomial (38) with $s = 4$. The two necessary constants $\Delta_{1,2}$ can be fixed such that the width of the available stable interval around λ is maximal. The stability domains of these four polynomials for the special case of $\lambda = 30$ are shown in Fig. 6. All domains include the interval $[-1, 0]$ near the origin due to consistency.

The polynomial (37) produces an almost circular shape around λ which grows with higher multiplicity of the root λ. Correspondingly, larger intervals on the real axis are included around the value λ. On the other hand, the polynomial (38) shows an eight-shaped stability domain. This has to be compared with the case $s = 4$ and double-root at λ. Proper adjustment of the constants Δ_k allows a bigger real interval than the polynomial with only a double root at λ.

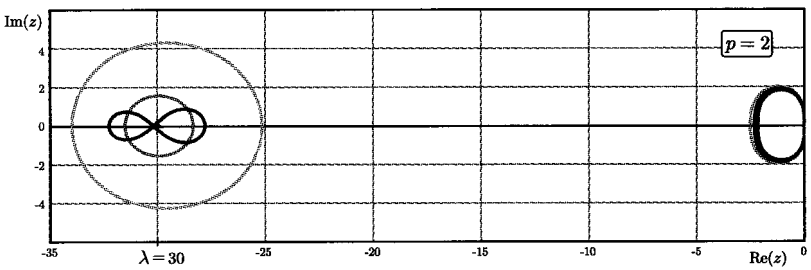


Figure 6: Stability domains for a spectrum with gap. The circular domains are realized with the polynomial (37) with $s = 3, 4, 5$ while the eight-shaped domain stems from (38) with $s = 4$. The aim is to produce a stability domain which allows a maximal width of a real interval around $\lambda = 30$.

It is interesting to see how the possible maximal width of the real interval around λ increases if λ increases. Fig. 7 shows the corresponding result for the four cases considered here. The plot shows the possible width of the stability domain over different values of λ for different poly-

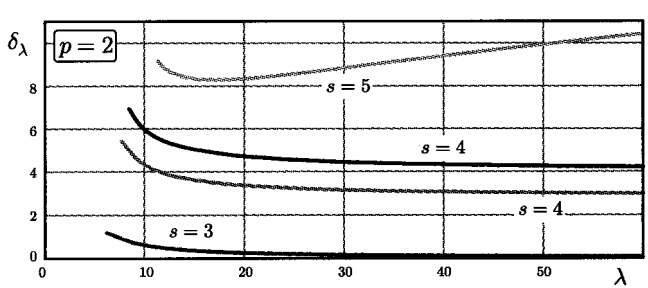


Figure 7: Maximal stable interval width δ around a given value λ in stability domains for spectral gaps. The higher curve for $s = 4$ corresponds to the polynomial (38) with optimized constants $\Delta_{1,2}$, while all other curves relate to the polynomial form (37).

nomials. The stability polynomial with a single root at λ (lowest curve, $s = 3$) allows only very small widths which are decaying for larger λ. In the plot only the case with a triple root ($s = 5$) shows an increasing width for larger values of λ. In the plot, the third curve from below corresponds to the polynomial (38) with $s = 4$ and roots $\lambda + \Delta_{1,2}$ optimized for a maximal width. Clearly, this yields larger widths than the case with a double root, i.e. $\Delta_{1,2} = 0$, depicted in the curve below for (37) and $s = 4$.

Optimizing the roots $\lambda + \Delta_{1,2}$ for a maximal width is related to the maximal real interval case in Sec. 3. The result of Sec. 3 can be used to construct even larger widths with polynomials with higher s.

7 Maximal thin regions

We note that in applications like compressible, viscous flow problems it is necessary to combine the situation of maximal real interval and the disk into, what we call a *thin region* G_r. The two main parameters of a thin region are r which is given by the largest interval $[-r, 0]$ contained in G_r and δ which is $\max(\operatorname{Im} z \mid z \in G_r)$.

The following definition assumes that a thin region is symmetric and is generated by a continuous real function.

Definition 1 (thin region). *The region $G_r \subset \mathbb{C}$ is called a* thin region, *if there exists a real continuous function $g_r(x)$, $x \in [-r, 0]$ with $g_r(0) = g_r(-r) = 0$, $\max_{x \in [-r,0]} g_r(x) = \delta$ and $r > 0$ such that*

$$G_r = \{z \in \mathbb{C} \quad | \quad |\operatorname{Im} z| \leq g_r(\operatorname{Re} z)\,,\ \operatorname{Re} z \in [-r, 0]\} \tag{39}$$

and $\delta / r \ll 1$.

The case $g_r \equiv 0$ produces the real interval as degenerated thin region. If a continuous function $\hat{g} : [-1, 0] \to [0, 1]$ is given, the thin region constructed by $g_r(a) = \delta\, \hat{g}(\frac{a}{r})$ is an affine mapping of $\hat{g}$ with $\hat{g}(-1) = \hat{g}(0) = 0$. For example $\hat{g}(x) = \sqrt{-x(1+x)}$ leads to a stretched ellipse with halfaxes r and δ. In the definition, g_r is generally parametrized by r. Hence, a family of thin regions G_r for different values of r may exhibit a different shape for different values of r and not only a shape obtained by affine mappings. However, the maximal thickness δ shall remain the same for all values of r. Fig. 8 shows a general case of a family of thin regions.

The real axis extension r of a thin region will be our main parameter. In the following we will describe how to derive optimal stability domains in the sense of (6) for thin regions. The stages will be optimized such that the stability domain allows a thin region G_r with maximal r. We

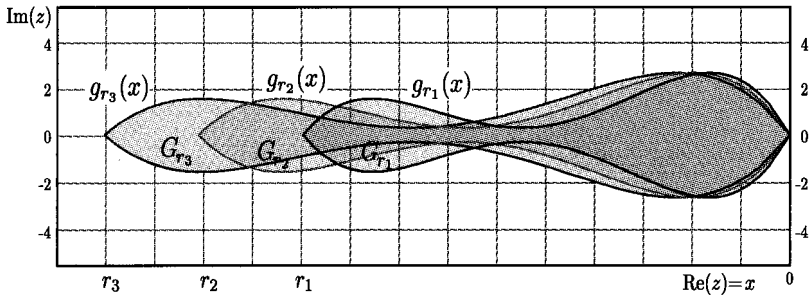


Figure 8: Example of thin regions G_r spanned by $g_r(x)$ for different values of r. In general G_r may have different shapes for different values of r.

will speak of a *maximal* thin region, which refers to a maximal extension r along the real axis at a given value of δ. A stability polynomial f_s with given order p and stages s that includes a maximal thin region in its stability domain will be called *optimal* for this thin region.

In [15] a theory is developed how to calculate optimal stability polynomials for thin regions. The theory relies on the hypothesis that in the optimal case the denting points of the boundary of the stability domain touch the boundary of the thin region. This leads to a direct characterization of the optimal polynomial. In the next section we only give the condensed algorithm how to compute the optimal polynomial for a given thin region with boundary $g_r(x)$. For details see [15].

7.1 Algorithm

The polynomial f_s will be uniquely described by $s-2$ extrema at real positions labelled $x_1 < x_2 < \cdots < x_{s-2} < 0$. The following algorithm determines these initially unknown positions

$$X = \{x_k\}_{k=1}^{s-2} . \tag{40}$$

The derivative f_s' has the form

$$f_s'(z; X) = 1 + z + \sum_{k=2}^{s-1} \beta_k z^k \stackrel{!}{=} (1 - \frac{z}{x_{s-1}}) \prod_{k=1}^{s-2} (1 - \frac{z}{x_k}), \tag{41}$$

from which the remaining extremum

$$x_{s-1} = -\frac{1}{1 + \sum_{k=1}^{s-2} \frac{1}{x_k}} \tag{42}$$

follows as function of the given extrema X. The stability polynomial is now given by

$$f_s(z;\, X) = 1 + \int_0^z f_s'(\zeta;\, X)d\zeta \tag{43}$$

based on the $s-2$ extrema X.

It remains to formulate an expression for the value of r in dependence of X. We will assume that the boundary of the stability and the thin region coincide at $z = -r$. If f_s is constructed from X the boundary point on the real axis can easily be calculated by solving $|f_s(r, X)| = 1$ which gives a function $r = R(X)$.

Finally, we have to solve the following equations in order to obtain an optimal stability polynomial.

Problem 2 (maximal thin region stability). *Given $g_r(z)$ and the unknowns $X = \{x_k\}_{k=1}^{s-2}$, solve the system of equations*

$$g_{R(X)}(x_k) = \sqrt{\frac{1 + f_s(x_k;\, X) sign(f_s''(x_k;\, X))}{|f_s''(x_k;\, X)|}}, \quad k = 1, 2, \cdots, s-2 \tag{44}$$

where $R(X) < x_1$ such that

$$|f_s(R(X);\, X)| = 1, \tag{45}$$

for the unknown extrema positions $X = \{x_k\}_{k=1}^{s-2} \subset \mathbb{R}$.

Note, that the current formulation does not require any form of optimization since it is based on a direct characterization by a single system of equations. This system of non-linear equations was implemented in C and solved with the advanced quasi-Newton method provided by [12]. An appropriate initial guess is found by choosing $g_r \equiv 0$ and the first order or second order maximal real interval result. For various shapes of thin regions a continuation method was employed. To avoid round-off errors the derivative (41) was converted into a representation by Chebyshev-polynomials on an sufficiently large interval for each evaluation of the residual. The necessary differentiation, integration and evaluation was then performed on the Chebyshev coefficients. This method proved to be efficient and stable also for large values of s.

Due to approximations that entered the equations (44) the resulting polynomial will be only essentially optimal. However, in actual applications this is sufficient.

7.2 Examples

In this section we show several examples of optimal thin region stability polynomials in order to demonstrate the flexibility and usefulness of the

proposed algorithm. We only present results for $p = 2$. Some of the examples must be considered as extreme cases of possible spectra.

Fig. 9 shows two optimal stability regions with $s = 9$ for two different thin regions. The upper case is that of a real interval with no imaginary part. In both results the denting points reach down to the boundary of the thin region. The deeper they reach the longer the real extension. Hence, the lower example has a smaller value for r.

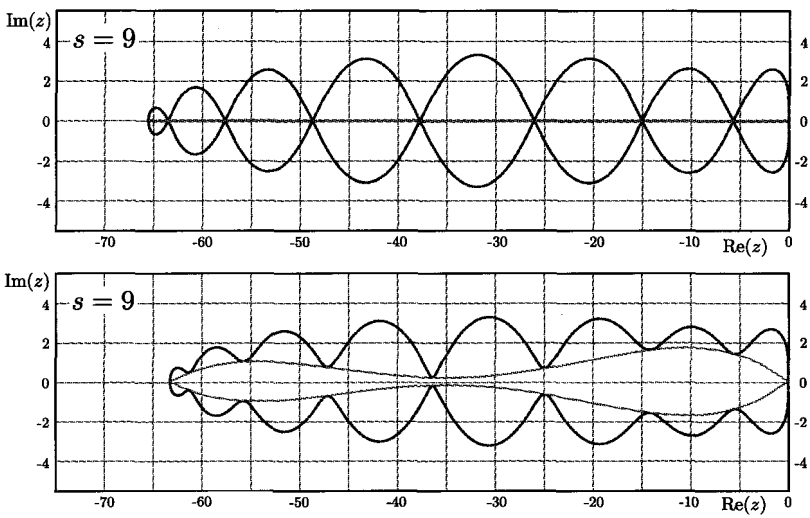


Figure 9: Two examples of thin regions, the real interval and a non-convex domain, together with their respective optimal stability region in the case $s = 9$ and $p = 2$. The stability domains allow a maximal extention r along the real line for the particular shape. Note, that the second case requires a smaller value of r.

The thin region can be of almost arbitrary shape, even though the framework presented above is developed for well behaved, smooth regions. In Fig. 10 the thin region has been subdivided into relative regions of different thickness. In the upper plot the thin region is subdivided into parts with relations 1:2:1. In the lower plot the five parts have relations 1:3:2:3:1. The small parts have a thickness of 0.1, in contrast to 1.6 of the thick parts. The algorithm manages to find the optimal stability region in which the denting points touch the boundary of the thin region. Problems can occur when the side pieces of the rectangles cut the stability domain boundary. For that reason the first derivative of g_r should be sufficiently small in general.

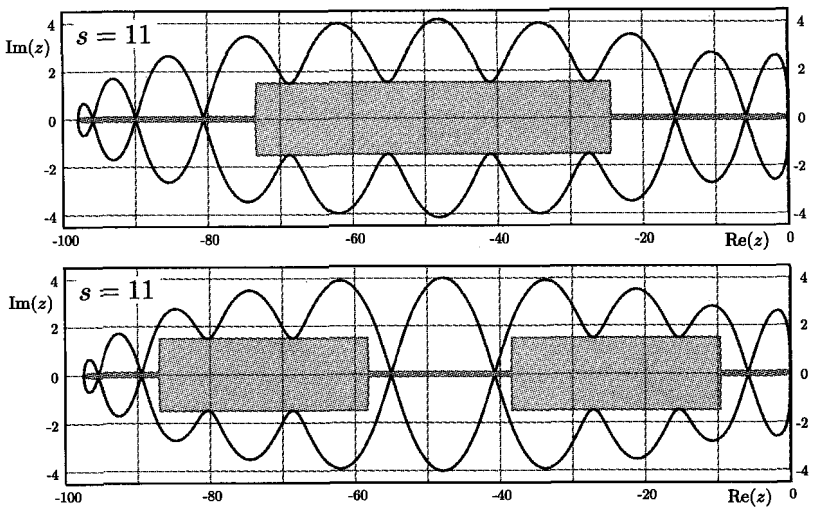


Figure 10: Two examples to demonstrate the ability of the proposed algorithm to produce highly adapted essentially optimal stability regions. The rectangles occupy relative parts of the real extension and have a thickness of 1.6.

8 Stabilized advection-diffusion

Spectra in the form of a thin region occur in semi-discretization of upwind methods for advection-diffusion. We briefly describe the differential equations, resulting spectra and optimal stability domains. A detailed discussion can be found in [15].

8.1 Semi-discrete advection-diffusion

We will consider the scalar function $u : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ and the advection-diffusion equation

$$\partial_t u + a\partial_x u = D\,\partial_{xx} u \tag{46}$$

with a constant advection equation $a \in \mathbb{R}$ and a positive diffusion constant $D \in \mathbb{R}$.

For advection with $a > 0$ the standard upwind method gives $\tilde{F}^{(hyp)}_{i+\frac{1}{2}} = a\,u^{(-)}_{i+\frac{1}{2}}$ for the transport part where $u^{(-)}_{i+\frac{1}{2}}$ is some reconstructed value of u on the left hand side of the interface $i + 1/2$. The diffusive gradients are discretized by central differences around the interface. We obtain as

semi-discrete numerical scheme

$$\partial_t u_i(t) = \frac{1}{\Delta x}\left(\tilde{F}^{(D)}_{i-\frac{1}{2}}(\hat{u}) - \tilde{F}^{(D)}_{i+\frac{1}{2}}(\hat{u})\right) \tag{47}$$

$$\text{with:} \quad \tilde{F}^{(D)}_{i+\frac{1}{2}}(\hat{u}) = a\left(u_i + \frac{1}{4}(u_{i+1} - u_{i-1})\right) - \frac{D}{\Delta x}(u_{i+1} - u_i), \tag{48}$$

which is second order in space, see e.g. the text book [11] for more informations about finite volume methods.

8.2 Optimal stability regions

The spectrum of the system (47) can be obtained analytically and can be written as a thin region G_r with the shape of a distorted ellipse, see [7], [15] or [16] for details. The thickness δ is given by $\approx 1.7\lambda$ with the courant number $\lambda = \frac{a\,\Delta t}{\Delta x}$ for given time step Δt and the real extension r is given by $2(1+\kappa)\lambda$ with an inverse grid Reynolds number $\kappa = \frac{2D}{a\,\Delta x}$. Hence, for large diffusion constant D or fine grids the thin region becomes longer, while the thickness stays the same. For a given number of stages s we are now looking for an optimal stability polynomial that includes the advection-diffusion spectrum with a maximal value of κ. In the following we will assume $\lambda = 1$ which means that the time step shall resolve the advection scale on the current grid, i.e., $\Delta t \approx a\,\Delta x$. In principle, $\lambda > 1$ is possible allowing time steps larger than those of the traditional CFL condition.

The optimal stability polynomials f_s for fixed s for the second order diffusive upwind method (47) are calculated by the algorithm described in Sec. 7.1 with $s = 3, \cdots, 101$. Except for the lower cases $s = 3, 4$ all polynomials were obtained from solving the equations in Sec. 7.1. The lower cases do not exhibit a thin region due to small values of κ and the optimal polynomials have been found by a separate optimization. In principle stability polynomials for higher values of s could also be obtained. As example the result for $s = 9$ is displayed in Fig. 11. For $s = 9$ the maximal real interval $[-r_{\max}, 0]$ included is $r_{\max} \approx 62.2$ which allows $\kappa \approx 30.1$.

For the case of a pure real interval the relation $r_{\max} \approx s^2$ has been reported, e.g., in the work of [1]. For the present results the maximal value $r_{\max}$ and the quotient $r_{\max}/s^2$ are displayed in Table 1. The numbers suggest the relations $r_{\max} \approx 0.79\,s^2$, respectively. In [15] also the spatially first order case is considered. The spectrum is thinner and correspondingly allows for larger $r_{\max} \approx 0.81 s^2$.

Table 1: Maximal real interval $[-r_{\max}, 0]$ included in the stability regions of f_s of the thin region for the spatially second order case $g^{(2)}$.

s	$r_{\max}$	$r_{\max}/s^2$	s	$r_{\max}$	$r_{\max}/s^2$
2	2.0	0.5	10	77.321	0.7732
3	4.520	0.5022	20	315.949	0.7898
4	10.552	0.6595	30	713.359	0.7926
5	17.690	0.7076	40	1269.691	0.7935
6	26.447	0.7346	50	1984.962	0.7939
7	36.782	0.7507	70	3892.310	0.7943
8	48.707	0.7610	90	6435.433	0.7944
9	62.220	0.7682	100	7945.410	0.7945

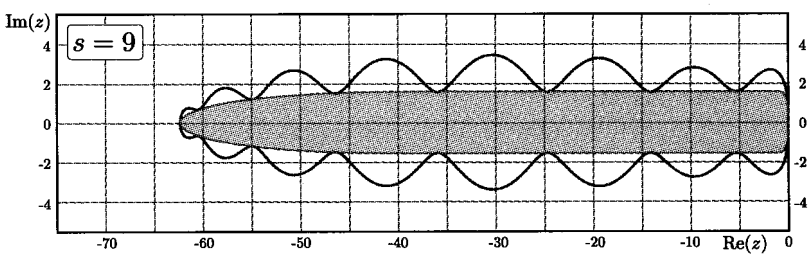


Figure11: The optimal second order stability domains for semi-discretized advection-diffusion for $s = 9$ in the spatially second order case.

8.3 Method construction

Once the stability polynomials are known it remains to construct practical Runge-Kutta methods from them. In principle, it is possible to conduct all internal steps with a very small time step $\tau\,\Delta t$, where τ is the ratio between allowable Euler step and full time step. For an ODE

$$y'(t) = F\left(y(t)\right), \tag{49}$$

we formulate the following algorithm for one time step.

Algorithm 1 (extrapolation type). *Given initial data y_n at time level n. Let $y^{(0)} = y_n$.*

$$k_j = F\left(y^{(j)}\right), \qquad y^{(j+1)} = y^{(j)} + \tau\Delta t\, k_j, \quad j = 0, 1, 2, \cdots, s-1$$

$$y_{n+1} = y_n + \sum_{j=1}^{s} \alpha_j\, k_j = \sum_{j=0}^{s} \tilde{\alpha}_j\, y^{(j)}. \tag{50}$$

The parameters $\tilde{\alpha}_j$, $j = 0, 1, \cdots, s$, can be calculated from any stability polynomial f_s by the solution of a linear system once τ is chosen. Since the time span $s\,\tau\,\Delta t$ is much smaller than Δt for the current methods, this algorithm can be viewed as extrapolation of the final value y_{n+1} from the shorter steps. Note, that it may be implemented with only one additional variable vector for temporary storage.

Another possibility is a variant of an algorithm given in [1], where the recursive formula for an orthogonal representation of the stability polynomial was used supplemented by a second order finishing procedure. Here, we simplify this method by using a combination of single Euler steps of increasing step sizes and the finishing procedure.

Algorithm 2 (increasing Euler steps). *Given initial data y_n at time level n. Let $y^{(0)} = y_n$.*

$$\begin{aligned} y^{(j+1)} &= y^{(j)} + \alpha_{j+1}\Delta t\, F\left(y^{(j)}\right), \quad j = 0, 1, 2, \cdots, s-2 \\ y_{n+1} &= y^{(s-1)} + \alpha_{s-1}\Delta t\, F\left(y^{(s-1)}\right) + \sigma\left(F\left(y^{(s-1)}\right) - F\left(y^{(s-2)}\right)\right). \end{aligned} \tag{51}$$

The parameters become obvious when the form

$$f_s(z) = (1 + \beta_1\, z + \beta_2\, z^2) \prod_{k=1}^{s-2} \left(1 - \frac{z}{z_k}\right) \tag{52}$$

of the stability polynomial is used. The Euler steps are given by the real zeros $\alpha_j = -\frac{1}{z_j}$, $j = 1, 2, \cdots, s-2$, while the second order procedure represents the part containing the complex zeros and we find $\alpha_{s-1} = \beta_1/2$ and $\sigma = 2\beta_2/\beta_1 - \beta_1/2$. Again, an implementation with only one temporary storage variable is possible. This method conducts time steps of different size. It can be viewed as multi-scale time stepping in which the different time steps damp the unstable high frequencies in such a way that a large time step is achievable in the finishing procedure.

Both methods are practical but have advantages and drawbacks in terms of internal stability and robustness. While the first one proceeds with only making very small time steps, the extrapolation procedure in the end may be difficult to evaluate in a numerically stable way. On the other hand the second method does not have any extrapolation, but conducts time steps which grow from very small to almost $\frac{1}{3}\Delta t$. Half of the time steps made will be using step sizes bigger than the allowable step size for a single explicit update (Euler method). Only the overall update will be stable. However, in real flow applications a single time step with large step size could immediately destroy the physicality of the solution, e.g. negative densities and force the calculation to break

down. Hence, special care is needed when designing and implementing the Runge-Kutta method.

In order to relax the problem of internal instabilities, a special ordering of the internal steps during one full time step is preferable in the second method. This is investigated in the work [10] from Lebedev, see also the discussion in [4]. Here we interchange steps with large and small step sizes and start with the largest one. The result yields a practical and efficient method as shown in the numerical examples in the next section for advection-diffusion and viscous, compressible flow, see [15].

8.4 Numerical experiments

The parameters of the explicit Runge-Kutta methods derived above have been calculated with high precision and implemented in order to solve an instationary problem of advection-diffusion. Due to the special design of the method and the possibility of choosing the optimal number of stages according to the strength of the diffusion, i.e., the value of κ, the time step during the simulation is fully advection-controlled. In the following we present some numerical experiments for the derived scheme for advection-diffusion equations. The implementation considers the scheme (47) and the stabilized Runge-Kutta method uses increasing Euler steps as in Algorithm 2.

For fixed s the time step of the method has to satisfy

$$\frac{a\,\Delta t}{\Delta x} \leq CFL \cdot \lambda_{\max}^{(s)}(\kappa) \tag{53}$$

with

$$\lambda_{\max}^{(s)}(\kappa) = \min\left(1, \frac{r_{\max}^{(s)}}{2(\kappa+1)}\right), \tag{54}$$

where $\kappa = 2D/(a\Delta x)$ as above. For time and space depending values of a and κ, this procedure provides an adaptive time step control as proposed, e.g., in [11] for hyperbolic problems. The value of $r_{\max}^{(s)}$ is given for each method. The number $CFL \leq 1$ allows to increase the robustness of the method by reducing the time step below the marginally stable value. We suggest the usage of $CFL \approx 0.9$, which is common when calculating hyperbolic problems. In Fig. 12 the graphs of $\lambda_{\max}^{(s)}$ for $s = 2, 3, 4, 5$ are drawn. We can see that the range of the diffusion parameter κ in which a pure advection time step $a\Delta t/\Delta x = 1$ is allowed grows with s. However, for larger s also more internal stages are needed. Hence, in a *stage-adaptive calculation* the number of stages s is chosen such that the method just reaches the kink in Fig. 12 for the current value for κ. The optimal s is given by

$$s^{(\mathrm{opt})} = \min\left\{ s \mid \lambda_{\max}^{(s)}(\kappa) = 1 \right\}. \tag{55}$$

This assures maximal efficiency. The source code is available online, see [14].

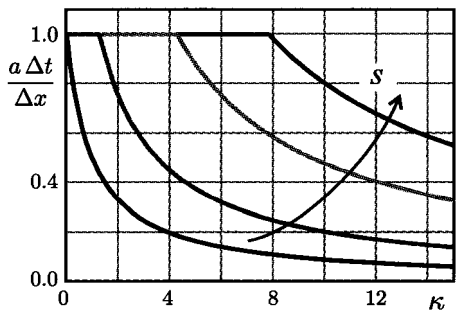


Figure 12: Time step constraints for advection-diffusion for stabilized explicit Runge-Kutta methods with stages $s = 2, 3, 4, 5$ drawn over the diffusion parameter $\kappa = 2D/(a\Delta x)$.

As an example we solved the time evolution for smooth periodic data on the interval $x \in [-2, 2]$ with periodic boundary conditions up to time $t = 0.8$, see [15] for details. Advection velocity is $a = 1$ and various diffusion coefficients in the advection dominated regime between $D = 0.001$ and $D = 1.0$ have been considered. The exact solution for these cases are easily found by analytic methods. For values of $CFL = 0.95$ or $CFL = 0.99$ all methods for various s were verified empirically to be second order convergent and stable.

It is interesting to compare the standard explicit time integration with $s = 2$ and the new adaptive procedure in which the number of stages is chosen according to the grid and value of diffusion coefficient, i.e. the value of κ. The method in which the number of stages is chosen adaptively integrates the equation with a time step which is purely derived from the advection. This time step is much larger than that required from a non-stabilized classical method as the method with $s = 2$, especially when D and/or the grid resolution is large. Also the efficiency increases since fewer function evaluations are needed as shown above. For the present case with $D = 0.1$ the two plots in Fig. 13 compare the stage-adaptive stabilized method with the classical method $s = 2$ in terms of efficiency. Both plots show the number of grid update evaluations for a calculation up to $t = 1$ on the ordinate. The first plot relates the number of evaluations to the grid resolution and the second to the achieved error. For high resolution or small errors the adaptive method requires an order of magnitude less work. For the adaptive method the work is approximately $\approx \mathcal{O}(N)$ which shows the linear scaling of an advection time step. The speed-up against the classical scheme is even

increased for higher values of the diffusion coefficient or finer grids.

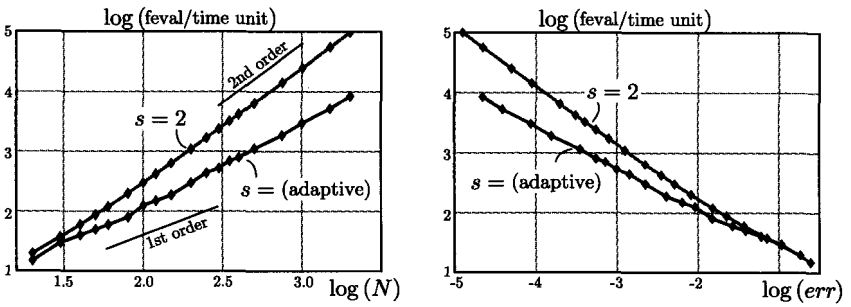


Figure 13: Comparison of neccessary work for a specific resolution (left) or a specific error (right) in the case of a classical method $s = 2$ and the new stabilized adaptive time steping.

9 Conclusion

In this report we presented families of stability polynomials for explicit Runge-Kutta methods that exhibit some optimality. For fixed number of stages s and order p they either include a maximal real interval, a maximal disk, a maximal imaginary interval, a maximal thin region, or a spectral gap with a spectrum part of maximal width separated from the origin. These families can be used to construct Runge-Kutta methods that adaptively follow a spectrum given in a respective application without the need of reducing the time step. Instead the number of stages of the method is increased in a specific way to take care of a specific spectrum.

The case of maximal thin regions is considered in greater detail following [15]. A *thin region* is a symmetric domain in the complex plane situated around the real line with high aspect ratio. Stability polynomials f that include a thin region with maximal real extension can be computed from a direct characterization with nonlinear equations for the coefficients of f.

Spectra in the form of thin regions occur in semi-discretizations of advection-diffusion equations or hyperbolic-parabolic systems. We presented optimal stability polynomials for explicit Runge-Kutta methods for advection-diffusion. For strong diffusion or fine grids they use more stages in order to maintain a time step controlled by the advection alone. Some numerical experiments demonstrate the efficiency gain over standard explicit methods.

Acknowledgement: The authors thank Ernst Hairer (University of Geneva) for pointing out reference [13] to us.

References

[1] A. Abdulle and A. A. Medovikov, *Second Order Chebyshev Methods Based on Orthogonal Polynomials*, Numer. Math. **90**, (2001), p.1-18

[2] A. Abdulle, *On roots and error constants of optimal stability polynomials*, BIT **40**(1), (2000), p.177-182

[3] E. Hairer, S. P. Norsett, and G. Wanner, Solving Ordinary Differential Equations, Volume I. Nonstiff Problems, Springer Series in Comput. Math. **8**, 2nd ed. Springer, Berlin (1993)

[4] E. Hairer and G. Wanner, Solving Ordinary Differential Equations, Volume II. Stiff and Differential-Algebraic Problems, Springer Series in Comput. Math. **14**, 2nd ed. Springer, Berlin (1996)

[5] P. J. van der Houwen and B. P. Sommeijer, *On the internal stability of explicit m-stage Runge-Kutta methods for large m-values*, Z. Angew. Math. Mech. **60**, (1980) p.479-485

[6] W. Hundsdorfer and J. G. Verwer, Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer Series in Computational Mathematics, Vol. 33, Springer, Berlin (2003)

[7] H.-O. Kreiss and H. Ulmer-Busenhart, Time-dependant Partial Differential Equations and Their Numerical Solution, Birkhäuser, Basel (2001)

[8] R. Jeltsch and O. Nevanlinna, *Largest Disk of Stability of Explicit Runge-Kutta Methods*, BIT **18**, (1978) p.500-502

[9] R. Jeltsch and O. Nevanlinna, *Stability of Explicit Time Discretizations for Solving Initial Value Problems*, Numer. Math. **37**, (1981) p.61-91

[10] V. I. Lebedev, *How to Solve Stiff Systems of Differential Equations by Explicit Methods*, in Numerical Methods and Applications, ed. by G. I. Marchuk, p.45-80, CRC Press (1994)

[11] R. J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Cambridge University Press, Cambridge (2002)

[12] U. Nowak,and L. Weimann, *A Family of Newton Codes for Systems of Highly Nonlinear Equations - Algorithms, Implementation, Applications*, Zuse Institute Berlin, technical report TR 90-10, (1990), code available at `www.zib.de`

[13] B. Owren and K. Seip, *Some Stability Results for Explicit Runge-Kutta Methods*, BIT **30**, (1990), p.700-706

[14] M. Torrilhon, Explicit method for advection-diffusion equations, Example Implementation in C, code available online at `www.math.ethz.ch/~matorril/ExplCode`, (2006)

[15] M. Torrilhon and R. Jeltsch, *Essentially Optimal Explicit Runge-Kutta Methods with Application to Hyperbolic-Parabolic Equations*, Num. Math. (2007), in press

[16] P. Wesseling, Principles of Computational Fluid Dynamics, Springer Series in Computational Mathematics, Vol. 29, Springer, Berlin (2001)

MS-Stability Analysis for Numerical Solutions of Stochastic Differential Equations —Beyond Single-step Single Dim—

Taketomo Mitsui
Graduate School of Information Science, Nagoya University, Nagoya, Japan.

Yoshihiro Saito
Gifu Shotoku Gakuen University, Gifu, Japan.

Abstract

Stability analysis for numerical solutions of stochastic differential equations (SDEs) is discussed. Similar to deterministic ordinary differential equations (ODEs), various numerical schemes, mainly generating a discretized random sequence to approximate the exact solution, are proposed for SDEs. Although convergence issue has been discussed in many literatures, a few results have been known about stability analysis in spite of its significance for numerical SDEs as well. We have proposed the *mean-square* stability (MS-stability) of numerical *single-step* schemes for a *scalar* SDE, that is, the numerical stability with respect to the mean-square norm when it is applied to the linear test equation. A big barrier has been observed to extend the concept to multi-step schemes or to multi-dimensional test equations, different from the ODE case. We will show a way to overcome the difficulties by tackling the two-dimensional case as well as two-step numerical schemes. The underlying idea of considering the self- or mutual-correlation of the solution components is promising for the MS-stability analysis beyond the single-step methods of one-dimensional case.

1 Introduction

Many phenomena, incorporated with randomness, in science and engineering can be mathematically modelled with stochastic differential equations (SDEs). Initial value problem of an SDE is expressed by the formula

$$dX(t) = f(X(t))\,dt + g(X(t))\,dW(t) \quad (t > 0), \qquad X(0) = X_0. \quad (1.1)$$

Here f and g are functions of $\mathbb{R}^d \to \mathbb{R}^d$ and $W(t)$ denotes the standard Wiener process, a Gaussian random process satisfying the conditions $P(W(0) = 0) = 1$, $\mathbf{E}(W(t)) = 0 \quad \forall t \in (0, \infty)$ and $\mathbf{C}(t,s) \equiv \mathbf{E}(W(t)W(s)) = \min(t,s)$. Here $\mathbf{E}$ means the mathematical expectation. Therefore, the solution $X(t)$ of SDE (1.1) is a random process. Note that Eq. (1.1) should be interpreted through the equivalent integral form given by

$$X(t) = X(0) + \int_0^t f(X(s))\,ds + \int_0^t g(X(s))\,dW(s). \tag{1.2}$$

According to the interpretation of the stochastic integral appearing in the second term of the right-hand side of (1.2), SDEs (and the stochastic integral) split into the Itô- and Stratonovich-type. Hereafter we will take the Itô-type for SDEs.

The stochastic analysis gives conditions for the unique existence of the solution of (1.1). However, generally it is very hard to obtain an analytical (closed) form of solution, especially in nonlinear cases of f and/or g. Similar to the ordinary differential equation (ODE) case, numerical solutions of SDEs have thus much significance. In fact, various numerical solutions have been proposed. One can refer to the book of P. KLOEDEN and E. PLATEN [2] and the literature therein. Major numerical schemes generate a discretized random sequence $\{X_n\}$, each of which approximates $X(t_n)$, along the step points $\{t_n\}$ $(n = 0, 1, 2, \cdots,\ t_n = nh)$ so that they are often called as discrete variable methods. Convergence issue has been discussed in many literatures. Stability analysis is significant for numerical SDEs as well, however a few results have been known. We have proposed the *mean-square* stability (MS-stability) of numerical schemes for a scalar SDE ([6]), that is, the numerical stability with respect to the mean-square norm when it is applied to the scalar linear test equation

$$dX(t) = \lambda X(t)\,dt + \mu X(t)\,dW(t) \quad (t > 0), \qquad X(0) = 1.$$

Since the equation has the supermartingale solution given by

$$X(t) = \exp\{(\lambda - \frac{1}{2}\mu^2)t + \mu W(t)\},$$

the condition for the asymptotic stability of the solution in the mean-square sense turns out to $2\Re\lambda + |\mu|^2 < 0$. The problem is whether or how the numerical solution inherits the stability. We already proposed the concept of MS-stability only for single-step schemes with the scalar test equation shown above.

A big barrier has been observed to extend the concept to multi-step schemes or to vector test equation, different from the ODE case. In the

present paper we will show a way to overcome the difficulties. Our results will be shown for the two-dimensional SDEs by single-step methods and for the scalar SDE by two-step methods, though they can suggest a way for more general cases.

2 Numerical schemes and MS-stability

Here we briefly review the discrete variable methods as well as their MS-stability for SDEs.

A discrete variable method stands for the constructive way to obtain a series of random variables $\{X_n\}$ approximating $\{X(t_n)\}$, the exact solution of (1.1) at t_n, along with the step-points $t_n = nh, \quad n = 1, 2, \cdots$, where a positive constant h means the stepsize. The simplest method is the Euler-Maruyama scheme given by

$$\begin{cases} \Delta W_n = W(t_{n+1}) - W(t_n), \\ X_{n+1} = X_n + f(X_n)h + g(X_n)\Delta W_n \quad (n = 1, 2, \ldots). \end{cases} \tag{2.1}$$

Here ΔW_n stands for an increment of the Wiener process, and is numerically realized as the random variable $\Delta W_n = \sqrt{h}\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ is the normal (Gaussian) random variable of 0 mean and the unit variance. Henceforth we can obtain an approximate generation of ΔW_n by a pseudo-random number generator on computer, which is a new factor in numerical implementation.

Of course the Euler-Maruyama scheme is the simplest. It is natural to seek more sophisticated schemes similarly as in the ODE case. The first point of view for improvements is to increase the accuracy of the methods, that is, the order of convergence. In numerical SDEs there are two concepts of order of convergence, hence we will touch on those.

As for (1.1), let T denote the terminal time of the solution, and set $T = Nh$. Then we call $e(h) = \mathbf{E}(|X(T) - X_N|)$ the absolute error of the numerical solution.

Definition 2.1. *X_N is said to converge strongly with order α if there exist $h_0 > 0$ and $K > 0$ satisfying*

$$e(h) \leq Kh^{\alpha} \quad \text{for} \quad h \in (0, h_0).$$

Intuitively, the strong order means a pathwise (trajectory) convergence to the random variable $X(T)$.

On the other hand, another concept can be introduced. Let $C_P^l(\mathbb{R}^d, \mathbb{R})$ denote the space of l times continuously differentiable function of $\mathbb{R}^d \to \mathbb{R}$ which, together with their partial derivatives up to and including order l, have polynomial growth.

Definition 2.2. *X_N is said to converge weakly with order q at time T as $h \to 0$ if for each $G \in C_P^{2(q+1)}(\mathbb{R}^d, \mathbb{R})$ there exists $C > 0$ independent of h and satisfying*

$$|\mathbf{E}[G(X(T))] - \mathbf{E}[G(X_T)]| = Ch^q.$$

Intuitively, we can regard the method with a weak order is approximating statistical quantities of $X(T)$ up to a certain order.

The Euler-Maruyama scheme is proved to be strongly convergent with order 1/2 and weakly convergent with order 1. To overcome this low order of the Euler-Maruyama scheme, SDE version of Runge-Kutta as well as linear multistep scheme can be thought of, and some preceding literature are devoted for it.

Here we stress that the numerical stability is significant in SDE as well as in ODE, for it is concerned with reliability of the scheme. For the stability analysis of numerical solutions, the following syntax diagram is taken in mind.

- Given a problem to be solved, we seek a criterion C_p which implies a certain stability behaviour of the analytical solution.
- A numerical scheme is applied to the problem.
- We seek another criterion C_m which implies an analogous stability behaviour as the analytical case when it is imposed on the scheme.

Now we concentrate our study on the asymptotic stability of the solution of SDEs in the mean-square sense. Consider the initial value problem of SDE (1.1) with $f(0) = g(0) = 0$. Then it is readily seen that the equation has the steady (or equilibrium) solution $X(t) \equiv 0$.

Definition 2.3. *The steady solution is said to be asymptotically stable in the mean-square sense if the conditions*

$$\forall \varepsilon > 0,\ \exists \delta > 0; \qquad \mathbf{E}\left(|X(t)|^2\right) < \varepsilon \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad |X_0| < \delta \tag{2.2}$$

and

$$\exists \delta_0; \qquad \lim_{t \to \infty} \mathbf{E}\left(|X(t)|^2\right) = 0 \quad \text{for all} \quad |X_0| < \delta_0 \tag{2.3}$$

hold. Here $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^d$.

By taking the linearized equation of (1.1) into consideration, the dominant part of stability behaviour can be analyzed through the following linear test equation in the scalar case ($d = 1$)

$$dX(t) = \lambda X dt + \mu X dW(t), \quad t \in [t_0, T], \qquad X(0) = 1, \tag{2.4}$$

where λ, $\mu \in \mathbb{C}$ with $\Re\lambda < 0$. This is also called the supermartingale equation and has the exact solution expressed with $X(t) = \exp\{(\lambda - \frac{1}{2}\mu^2)\, t + \mu W(t)\}$. As the condition C_p in the syntax diagram, the following result is obvious.

Lemma 2.1. *The linear test equation with the unit initial value is asymptotically stable in the mean-square sense iff*

$$2\Re\lambda + |\mu|^2 < 0.$$

Then what conditions for the numerical solution $\{X_n\}$ of the linear test equation (2.4) can be imposed so as the condition $|X_n| \to 0$ as $n \to \infty$ holds? In [6] we proposed the concept of the numerical mean-square stability (MS-stability in short) through the following steps. Define $Y_n = \mathbf{E}(|X_n|^2)$, apply a numerical single-step scheme to (2.4) and take the mean-square norm of the both sides. Then we obtain a one-step difference equation of the form $Y_{n+1} = R(\bar{h}, k)Y_n$, where $\bar{h} = h\lambda$, $k = -\mu^2/\lambda$ and R depends on them. We can call $R(\bar{h}, k)$ the stability function of the scheme, for the equivalence

$$Y_n \to 0 \quad \text{as} \quad n \to \infty \quad \Leftrightarrow \quad |R(\bar{h}, k)| < 1$$

is readily established. Hence we attain

Definition 2.4. *The single-step scheme is said to be MS-stable for $\bar{h}$ and k satisfying $|R(\bar{h}, k)| < 1$.*

It is natural to call $\mathcal{R}$ the domain of MS-stability of the scheme when

$$\mathcal{R} = \{(\bar{h}, k);\ |R(\bar{h}, k)| < 1 \text{ holds}\}.$$

Observe that the condition $2\Re\lambda + |\mu|^2 < 0$ in Lemma 2.1 implies $\Re\lambda < 0$ and, for real λ and μ, it also means $0 < k < 2$.

For instance, we can compute the stability function of the Euler-Maruyama scheme as

$$R(\bar{h}, k) = |1 + \bar{h}|^2 + |k\bar{h}|. \tag{2.5}$$

However, an extension of the concept of MS-stability beyond single-step methods for scalar SDEs is not straightforward. This is different from the ODE case, where the test equation $dx/dt = \lambda x$ ($\lambda \in \mathbb{C}$) is valid for the system of ODEs $d\boldsymbol{x}/dt = L\boldsymbol{x}$ (L is a square matrix) through a similarity transformation. SDEs cannot enjoy this convenience. In the following sections we will explain the situation and show a way to make a break-through as well.

3 • Two-dimensional MS-stability

First we will tackle an extension of MS-stability to the two-dimensional system of SDEs. According to the guideline of the linear stability analysis, we are led to the test equation of two-dimensional linear SDE system with one multiplicative noise given by

$$d\boldsymbol{X}(t) = \boldsymbol{D}\boldsymbol{X}(t)\,dt + \boldsymbol{B}\boldsymbol{X}(t)\,dW(t), \qquad \boldsymbol{X}(0) = \mathbf{1}. \tag{3.1}$$

The difficulty lies on the fact that we cannot assume the real constant matrices $\boldsymbol{D}$ and $\boldsymbol{B}$ simultaneously diagonalizable. Thus alternatively we impose the matrices as

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix} \tag{3.2}$$

with $\beta_1 \neq 0$ and $\beta_2 \neq 0$ at best. In [7] we developed the MS-stability theory as follows. First we seek an analytical criterion C_p of MS-stability. Instead of $\mathbf{E}\left(|X(t)|^2\right)$ of the scalar case, it is natural to look at $\boldsymbol{P}(t) = \mathbf{E}(\boldsymbol{X}(t)\,\boldsymbol{X}(t)^T)$ as 2×2 matrix-valued second moment of the solution of the test equation (3.1), (3.2). Indeed $\boldsymbol{P}(t)$ obeys the initial value problem of the matrix ODE given by

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = D\boldsymbol{P} + \boldsymbol{P}D^T + B\boldsymbol{P}B^T \quad (t > 0), \quad \boldsymbol{P}(0) = \boldsymbol{X}_0\,\boldsymbol{X}_0^T. \tag{3.3}$$

Inspired by $\boldsymbol{P}(t)$, we introduce a three-dimensional vector function $\boldsymbol{Y}(t) = (Y^1(t), Y^2(t), Y^3(t))^T$ whose components are

$$Y^1(t) = \mathbf{E}[(X^1(t))^2], \quad Y^2(t) = \mathbf{E}[(X^2(t))^2], \quad Y^3(t) = \mathbf{E}[X^1(t)\,X^2(t)].$$

Here the superfix denotes the component number. Then $\boldsymbol{P}$ is governed by ODE of three-dimension

$$\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}t} = \mathcal{M}\boldsymbol{Y} \tag{3.4}$$

by virtue of the symmetry of the matrix $\boldsymbol{P}(t)$. Here the matrix $\mathcal{M}$ for the test equation (3.1), (3.2) is given as

$$\mathcal{M} = \begin{bmatrix} 2\lambda_1 + \alpha_1^2 & \beta_1^2 & 2\alpha_1\beta_1 \\ \beta_2^2 & 2\lambda_2 + \alpha_2^2 & 2\alpha_2\beta_2 \\ \alpha_1\beta_2 & \alpha_2\beta_1 & \lambda_1 + \lambda_2 + \alpha_1\alpha_2 + \beta_1\beta_2 \end{bmatrix}. \tag{3.5}$$

For the establishment of the condition C_p in the syntax diagram of stability, a necessary and sufficient condition is desirable. Consequently we introduce the logarithmic norm of a matrix. For a square matrix A,

$$\mu_p[A] = \lim_{\delta \to 0+} (\|I + \delta A\|_p - 1)/\delta$$

is called the logarithmic norm associated with the norm $\|A\|_p$. Under the notation the following lemma is readily shown.

Lemma 3.1. *The linear test system with the unit initial value* (3.1), (3.2) *is asymptotically MS-stable w.r.t. the logarithmic norm μ_p iff*

$$\mu_p[\mathcal{M}] < 0.$$

For instance, in the ∞-norm case we have

Theorem 3.1. *The system* (3.1), (3.2) *with all the real components is MS-stable* w.r.t. μ_∞ *if the estimation*

$$\max\left\{2\lambda_1 + (|\alpha_1| + |\beta_1|)^2, \quad 2\lambda_2 + (|\alpha_2| + |\beta_2|)^2\right\} < 0 \tag{3.6}$$

holds.

Now we turn to a numerical criterion of MS-stability. By applying a numerical single-step scheme to the test equation and by calculating the components $\boldsymbol{X}_n$, we obtain the one-step difference equation

$$\boldsymbol{Y}_{n+1} = \overline{\mathcal{M}}\,\boldsymbol{Y}_n \tag{3.7}$$

where, as the discretized counterpart of $\boldsymbol{Y}(t)$, the three-dimensional vector $\boldsymbol{Y}_n$ is given by

$$\boldsymbol{Y}_n = (Y_n^1, Y_n^2, Y_n^3),\ Y_n^1 = \mathbf{E}[(X_n^1)^2],\ Y_n^2 = \mathbf{E}[(X_n^2)^2],\ Y_n^3 = \mathbf{E}[X_n^1 X_n^2] \tag{3.8}$$

and the components of the matrix $\overline{\mathcal{M}}$ is a function of the parameters $\lambda_i, \alpha_i, \beta_i$ and the stepsize h. Hence we call $\overline{\mathcal{M}}$ the stability matrix of the scheme.

Definition 3.1. *The single-step scheme for the two-dimensional equation is said to be MS-stable* w.r.t. $\|\cdot\|_p$ *if it has $\overline{\mathcal{M}}$ satisfying $\|\overline{\mathcal{M}}\|_p < 1$.*

According to the above criterion we studied MS-stability of the Euler-Maruyama scheme of the two-dimensional case and obtained a criterion for the ∞-norm ([7]). For instance, corresponding to Theorem 3.1, we have

Theorem 3.2. *For the test equation* (3.1), (3.2) *the Euler-Maruyama scheme gives*

$$\overline{\mathcal{M}} = \begin{bmatrix} r^2(\lambda_1 h) + \alpha_1^2 h & \beta_1^2 h & 2\alpha_1\beta_1 h \\ \beta_2^2 h & r^2(\lambda_2 h) + \alpha_2^2 h & 2\alpha_2\beta_2 h \\ \alpha_1\beta_2 h & \alpha_2\beta_1 h & r(\lambda_1 h)r(\lambda_2 h) + (\alpha_1\alpha_2 + \beta_1\beta_2)h \end{bmatrix},$$

which yields its numerical stability criterion of the real matrix case w.r.t. $\|\cdot\|_\infty$ *as*

$$\max\{(1+\lambda_1 h)^2+(|\alpha_1|+|\beta_1|)^2 h, \quad (1+\lambda_2 h)^2+(|\alpha_2|+|\beta_2|)^2 h\} < 1. \quad (3.9)$$

Here the symbol $r(x)$ *denotes* $1+x$.

In [7] we showed the numerical stability criteria are slightly different by the employed norm. However, here we stress the fact that the MS-stability region for the real scalar case is still efficient for the real-coefficient two-dimensional case. Put $(\overline{h}_i, k_i) \equiv (\lambda_i h, -(|\alpha_i|^2+|\beta_i|^2)/\lambda_i)$ for $i = 1, 2$. Then the question in the two-dimensional case is whether $(\overline{h}_1, k_1)$ and $(\overline{h}_2, k_2)$ lie in the stability region $\mathcal{R}$ of the Euler-Maruyama scheme stated in the previous section. Since we took the parameters $\bar{h} = \lambda h$ and $k = -\mu^2/\lambda$ for the scalar test equation, the two-dimensional case only appears to be indexed by looking the matrices $\boldsymbol{D}$ and $\boldsymbol{B}$ in (3.2) row-wise.

4 MS-stability of two-step methods

In the previous section we see that the idea taking the second moment into account is useful for the stability analysis of two-dimensional case. This can be carried over to the case of more dimensions. Here we will study that the underlying idea is applicable to two-step schemes of scalar SDE as well.

Like as for ODEs, extension over the Euler-Maruyama scheme for higher order is done in the directions of Runge-Kutta (single-step) and of linear multi-step. We will pay our attention for the MS-stability of multi-step schemes, for, in ODE case, linear multi-step extension generally reduces stability. Thus the question how is in SDEs attracts interests. Firstly we will analyze two-step schemes.

Several two-step schemes have been known (*e.g.*, [2]). One of them is the so-called stochastic mid-point scheme, which, for a scalar SDE $dX(t) = f(X(t))\,dt + g(X(t))\,dW(t)$, is given below.

$$X_{n+1} = X_{n-1} + 2f(X_n)h - f'(X_{n-1})\,g(X_{n-1})\Delta W_{n-1}h + V_n + V_{n-1}, \quad (4.1)$$

where

$$\begin{aligned} V_n = {} & g(X_n)\Delta W_n + \left[fg' + (g^2 g'')/2\right](X_n)\left\{\Delta W_n h - \Delta Z_n\right\} \\ & + [f'g](X_n)\Delta Z_n + [gg'](X_n)\left\{(\Delta W_n)^2 - h\right\}/2 \\ & + [g(gg')'](X_n)\left\{(\Delta W_n)^2/3 - h\right\}\Delta W_n/2. \end{aligned}$$

Here ΔZ_n denotes another Gaussian random variable satisfying $\mathbf{E}[\Delta Z_n] = 0$, $\mathbf{E}[(\Delta Z_n)^2] = h^3/3$ and $\mathbf{E}[\Delta W_n\,\Delta Z_n] = h^2/2$. Although it requires the derivatives f', g' and g'', it attains strong order 1.5.

Applying the scheme to the linear test equation (2.4), we obtain

$$X_{n+1} = X_{n-1} + 2\lambda h X_n - \lambda h \mu X_{n-1} \Delta W_{n-1} + V_n + V_{n-1}$$

where

$$V_n = \mu X_n \Delta W_n + \lambda \mu X_n \Delta W_n h + \frac{1}{2}\mu^2 \left((\Delta W_n)^2 - h\right) + \frac{1}{2}\mu^3 X_n \left(\frac{1}{3}(\Delta W_n)^2 - h\right) \Delta W_n.$$

Note that the random variable ΔZ_n no longer appears. Taking the expectation of the squared recurrence formula given above and putting $\bar{h} = \lambda h,\ k = -\mu^2/\lambda$ as for the single-step case, we arrive at

$$\mathbf{E}[(X_{n+1})^2] = \mathbf{E}[(X_n)^2]\left(4\bar{h}^2 - (1+\bar{h})^2 k\bar{h} + \frac{1}{2}k^2\bar{h}^2 - \frac{1}{6}k^3\bar{h}^3\right) + \mathbf{E}[(X_{n-1})^2]\left(1 - k\bar{h} + \frac{1}{2}k^2\bar{h}^2 - \frac{1}{6}k^3\bar{h}^3\right) + \mathbf{E}[X_n X_{n-1}]4\bar{h}.$$

This can be regarded as a linear recurrence w.r.t. $\mathbf{E}[(X_n)^2]$, $\mathbf{E}[(X_{n-1})^2]$ and $\mathbf{E}[X_n X_{n-1}]$. By virtue of the equation $\mathbf{E}[X_{n+1} X_n] = \mathbf{E}[X_n X_{n-1}] + 2\bar{h}\mathbf{E}[(X_n)^2]$, we can establish the linear transition relation as follows:

$$\begin{bmatrix} \mathbf{E}[(X_{n+1})^2] \\ \mathbf{E}[(X_n)^2] \\ \mathbf{E}[X_{n+1} X_n] \end{bmatrix} = \begin{bmatrix} r_{11}(\bar{h},k) & r_{12}(\bar{h},k) & 4\bar{h} \\ 1 & 0 & 0 \\ 2\bar{h} & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{E}[(X_n)^2] \\ \mathbf{E}[(X_{n-1})^2] \\ \mathbf{E}[X_n X_{n-1}] \end{bmatrix}, \tag{4.2}$$

where the components r_{11} and r_{12} are polynomials of the arguments. In particular

$$r_{12}(\bar{h},k) = 1 - k\bar{h} + \frac{1}{2}k^2\bar{h}^2 - \frac{1}{6}k^3\bar{h}^3.$$

Now the question results in the eigenvalue distribution of the coefficient matrix of (4.2), which decides whether the magnitude of the vector reduces or not. The following is a two-step extension of MS-stability.

Definition 4.1. *A two-step method is said to be MS-stable if all the eigenvalues of the transition matrix of* $(\mathbf{E}[(X_n)^2],\ \mathbf{E}[(X_{n-1})^2],\ \mathbf{E}[X_n X_{n-1}])^T$ *to* $(\mathbf{E}[(X_{n+1})^2], \mathbf{E}[(X_n)^2], \mathbf{E}[X_{n+1} X_n])^T$ *are less than unity in magnitude.*

Calculations show that the characteristic polynomial of the matrix in 4.2 is given by

$$z^3 + (\text{coeff.})z^2 + (\text{coeff.})z + r_{12}(\bar{h},k).$$

Note the conditions $\bar{h} < 0$ and $0 \le k \le 2$ hold for real λ and μ due to the analytical stability requirement. Then, as the product $k\bar{h}$ is negative, all the eigenvalues cannot be less than unity in magnitude even for real λ and μ. Henceforth we have

Theorem 4.1. *The stochastic mid-point scheme cannot be MS-stable in the extended sense to two-step case.*

Now we will investigate another two-step scheme a bit simpler but with more parameters than the mid-point one. Then we may adjust the parameters so as to attain MS-stability. Here we take the explicit two-step scheme given below. Refer to Eq. (12.4.4) in Kloeden-Platen [2]. It was shown that the scheme has strong order 1.0.

$$\begin{aligned} X_{n+1} &= (1-\gamma)X_n + \gamma X_{n-1} + \{(1+\gamma\alpha)f(X_n) \\ &\quad +\gamma(1-\alpha)f(X_{n-1})\}h + V_n + \gamma V_{n-1}, \\ V_n &= g(X_n)\Delta W_n + [gg'](X_n)\left\{(\Delta W_n)^2 - h\right\}/2. \end{aligned} \tag{4.3}$$

Here γ and α are parameters lying in $[0,1]$, and the selection $\gamma = 0$ gives the so-called Milstein scheme, which turns out to a single-step scheme.

Applying the scheme to the linear scalar test equation leads to

$$\begin{bmatrix} \mathbf{E}[(X_{n+1})^2] \\ \mathbf{E}[(X_n)^2] \\ \mathbf{E}[X_{n+1}\,X_n] \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 1 & 0 & 0 \\ D_1 & 0 & D_3 \end{bmatrix} \begin{bmatrix} \mathbf{E}[(X_n)^2] \\ \mathbf{E}[(X_{n-1})^2] \\ \mathbf{E}[X_n\,X_{n-1}] \end{bmatrix}, \tag{4.4}$$

where the matrix components are given by

$$\begin{aligned} C_1 &= \left\{(1-\gamma) + (1+\gamma\alpha)\bar{h}\right\}^2 - k\bar{h} + k^2\bar{h}^2/2, \\ C_2 &= \gamma^2\left\{1 + (1-\alpha)\bar{h}\right\}^2 - k\bar{h} + k^2\bar{h}^2/2, \\ C_3 &= 2\gamma\left\{(1-\gamma) + (1+\gamma\alpha)\bar{h}\right\}\left\{1 + (1-\alpha)\bar{h}\right\}, \\ D_1 &= (1-\gamma) + (1+\gamma\alpha)\bar{h}, \\ D_2 &= \gamma\left\{1 + (1-\alpha)\bar{h}\right\} \end{aligned}$$

by introducing as usual $\bar{h} = \lambda h$ and $k = -\mu^2/\lambda$. How the eigenvalue distribution depends on the parameters α and γ? Complicated algebraic computations, however, still do not give much insight. Instead, we study the MS-stability of the scheme by drawing pictures of the region of MS-stability for several pairs of the parameters. Some results are shown below. For simplicity, we are studying only for real λ and μ.

For the case $\gamma = 0$, which yields the Milstein scheme as mentioned above, the region of MS-stability is in Fig. 4.1.

Next, we fix γ as $1/2$ and vary α from 0 to 1. The regions of MS-stability for $\alpha = 0, 1/2$ and 1 are given in Figs. 4.2, 4.3 and 4.4, respectively. We can observe that the increasing of α reduces the area of the region in these cases.

Our analysis can be confirmed through numerical examples. In the region for the parameters $\gamma = 0.5$ and $\alpha = 0.0$, we can see that the

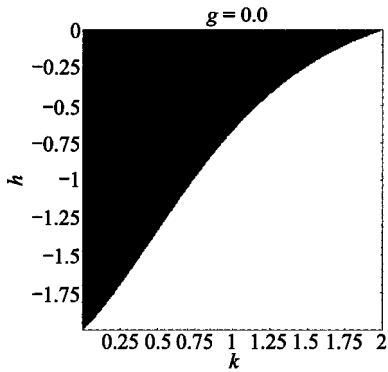


Figure 4.1 MS-stability region (Milstein scheme)

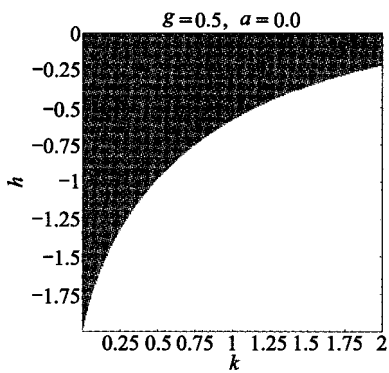


Figure 4.2 MS-stability region ($\gamma = 1/2, \alpha = 0$)

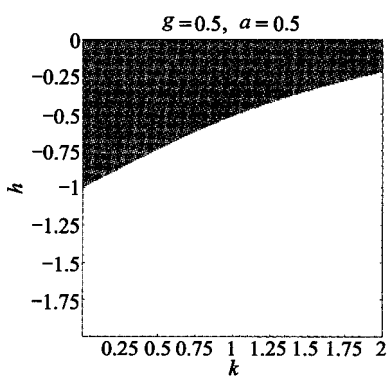


Figure 4.3 MS-stability region ($\gamma = 1/2, \alpha = 1/2$)

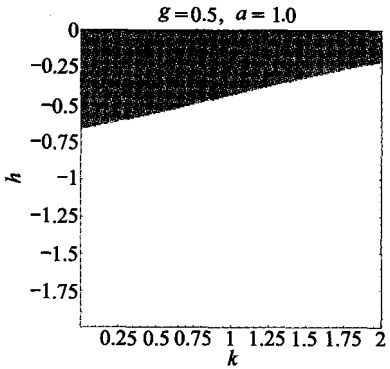


Figure 4.4 MS-stability region ($\gamma = 1/2, \alpha = 1$)

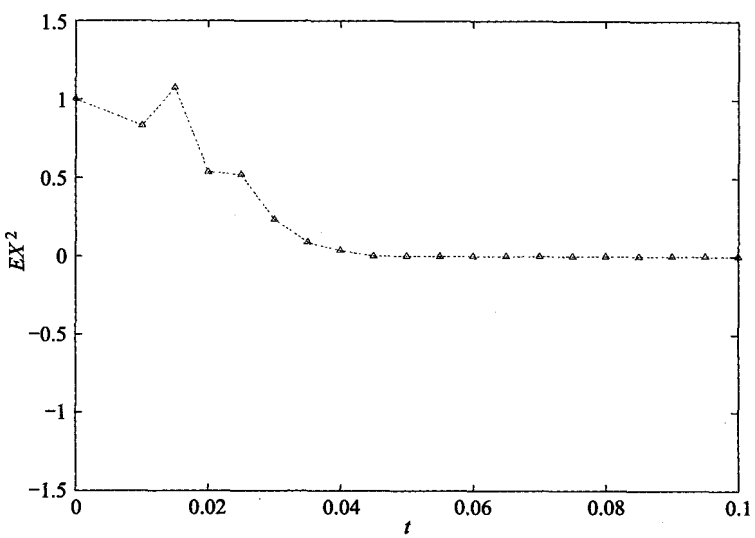


Figure 4.5 Stable solutions ($(\lambda, \mu, h) = (-100, 10, 0.005)$)

combination $(k, \bar{h}) = (1, -0.5)$ implies an MS-stable solution, while $(k, \bar{h}) = (1, -1)$ an MS-unstable one. They correspond, for example, to $(\lambda, \mu, h) = (-100, 10, 0.005)$ and $(\lambda, \mu, h) = (-100, 10, 0.01)$, respectively. The mean-square norm of numerical solution by 1000 samples with $(\lambda, \mu) = (-100, 10)$ by the scheme (4.3) of the stepsize $h = 0.005$ versus t is depicted in Fig. 4.5, while for the same λ and μ but with doubled h in Fig. 4.6.

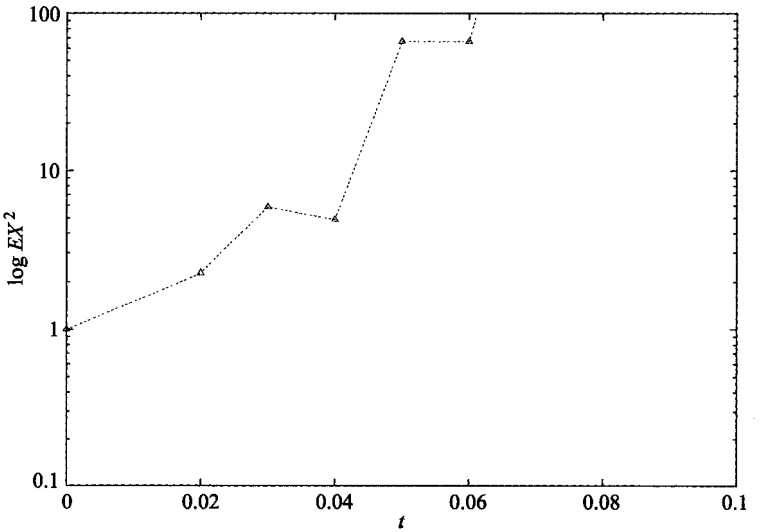


Figure 4.6 Unstable solutions $((\lambda, \mu, h) = (-100, 10, 0.01))$

5 Concluding remarks

An introduction of the concept of MS-stability as well as its analysis for numerical SDEs was carried out only for single-step methods through the scalar test equation. Behind this, some particular barriers have been observed in the SDE case, different from the ODE case. To overcome the difficulties, first we extend MS-stability analysis of the single step schemes to the two-dimensional test equation. Also the analysis is extended to two-step schemes based on the scalar test equation. In both cases the idea taking the self- or mutual correlations of solution components plays a key role. Therefore, we will try to establish stability criterion for more dimensions and more steps with the underlying idea.

References

[1] Burrage, K. Burrage, P. and Mitsui, T., Numerical solutions of stochastic differential equations — implementation and stability issues —, *J. Comput. Appl. Math.,* **125**(2000), 171-182.

[2] Kloeden, P. E. and Platen, E., *Numerical Solution of Stochastic Differential Equations,* Springer-Verlag, 1992.

[3] Komori, Y. and Mitsui, T., Stable ROW-type weak scheme for stochastic differential equations, *Monte Carlo Methods and Applic.,* **1**(1995), 279-300.

[4] Komori, Y., Saito, Y. and Mitsui, T., Some issues in discrete approximate solution for stochastic differential equations, *Computers Math. Applic.,* **28**(1994), No.10–12, 269–278.

[5] Saito, Y. and Mitsui, T., T-stability of numerical scheme for stochastic differential equations, WSSIAA **2** *"Contributions in Numerical Mathematics"* (ed. by R.P. Agarwal), World Scientific Publ., Singapore, 1993, pp. 333–344.

[6] Saito, Y. and Mitsui, T., Stability analysis of numerical schemes for stochastic differential equations, *SIAM J. Numer. Anal.,* **33**(1996), 2254-2267.

[7] Saito, Y. and Mitsui, T., Mean-square stability of numerical schemes for stochastic differential systems, *Vietnam Journal of Mathematics,* **30**(2002), Special Issue, 551-560.

[8] Schurz, H., *Stability, Stationarity, and Boundedness of Some Implicit Numerical Methods for Stochastic Differential Equations,* Logos Verlag, Berlin, 1997.

A Comparison of Direct Discretization of 4th Order Problems Versus System Reduction

R. D. Russell

J. F. Williams

X. Xu*

Department of Mathematics, Simon Fraser University
Burnaby, BC, Canada
E-mail: rdr@cs.sfu.ca, jfw@math.sfu.ca, xxub@sfu.ca

Abstract

In this paper, we compare a direct discretization of fourth order partial differential equations with two alternative discretizations of lower order systems. Numerical experiments show that caution must be taken when forming an ODE system by differentiating a DAE system on a moving grid. The mesh speed term must generally be included in order to obtain good numerical results. When solving a fourth order problem, it is recommended that one solves it directly instead of converting it into a mathematically equivalent system of lower order equations, especially when there is a solution singularity.

1 Introduction

Fourth order models are increasingly commonplace in the physical sciences, arising in thin-film theory, lubrication theory, convection-explosion theory, flame and wave propagation, phase transition and pattern formation. The recent book [8] introduces a number of analytical techniques for many such problems, but there is no corresponding treatise for the general numerical analysis of them. While this may of lesser concern when solving ordinary differential equations (ODEs) related to these models, it is often unclear how best to proceed when solving partial differential equations (PDEs).

More specifically, many successful boundary value codes simply discretize all equations as first order systems suggesting that there is in

*The first and third authors are supported in part by NSERC (Canada) through Grant OGP-0008781.

some sense little conceptual difference between fourth order and lower order equations. In this article we shall compare a direct discretization of fourth order equations with two alternative discretizations of lower order systems. Both codes in this investigation use moving collocation methods; the first is MOVCOL2 [7] and the second is MOVCOL4 [10]. Both MOVCOL2 and MOVCOL4 use a method of lines approach to solve general systems of the form

$$\begin{aligned} f(x,t,u_t,u,u_x) &= [g(x,t,u,u_x)]_x \quad \text{and} \\ f(x,t,u_t,u,u_x,u_{xx},u_{xxx}) &= [g(x,t,u,u_x,u_{xx},u_{xxx})]_x\,, \end{aligned} \tag{1.1}$$

respectively. MOVCOL2 uses a cubic approximation to the solution whereas MOVCOL4 uses a seventh order polynomial. Both codes solve (1.1) in integrated form preserving, when appropriate, the right hand side in the physically common case where $f \equiv u_t$. The DAE code DASSL [9] is used for the time integration in MOVCOL2 and MOVCOL4.

They use an adaptive strategy based on the so-called moving mesh PDE (MMPDE) approach as detailed in [6] and have been proven successful in the computation of difficult singular problems [2, 11]. The basic idea of this approach is to augment the physical PDE with a mesh equation such as MMPDE6:

$$\dot{x}_{\xi\xi} = -\frac{1}{\varepsilon}\left(M(u,u_x,x)x_\xi\right)_\xi\,, \tag{1.2}$$

where a fixed mesh in the computational coordinate ξ determines the mesh in x through the mapping $x(\xi,t)$. In this equation the *monitor function* $M > 0$ determines the mesh motion. Using a moving mesh, the physical PDE is modified to a quasi-Lagrangian frame, viz., $u_t = \dot{u} + u_x\dot{x}$. We shall return to this important point below.

1.1 Discretizations

In this article we are comparing the direct discretization of the fourth order PDE as outlined in [10] with alternative system strategies which reduce the PDE to a system of lower order equations and then solve using MOVCOL2. The reduction to lower order systems is a common technique for solving high order ODEs. Indeed, there are many ways to write fourth order equations as lower order systems, and it is our goal to compare them for the PDE case. For ease of exposition, we consider the simple problem

$$\begin{cases} u_t = -u_{xxxx} \\ u_x = u_{xxx} = 0, x = 0, \pi. \end{cases} \tag{1.3}$$

The most naive approach is simply to define $u_1 = u$ and $u_2 = u_{xx}$ giving rise to the system

$$\text{(DAE)} \begin{cases} u_{1t} = -u_{2xx} \\ u_2 = u_{1xx} \\ u_{1x} = u_{2x} = 0, x = 0, \pi. \end{cases} \tag{1.4}$$

This approximation leads to a differential algebraic equation (DAE) where on a moving grid at least two-fifths of the variables are algebraic. Interestingly, DDASSL has difficulty in most cases to start integrating this form even for simple problems. One well known strategy for DAEs is simply to differentiate the appropriate equations with respect to time yielding an ODE. Performing this gives

$$\text{(ODE1)} \begin{cases} u_{1t} = -u_{2xx} \\ u_{2t} = u_{1xxt} \\ u_{1x} = u_{2x} = 0, x = 0, \pi. \end{cases} \tag{1.5}$$

However, recalling that we have a quasi-Lagrangian formulation, this becomes

$$\text{(ODE1)} \begin{cases} \dot{u}_1 = -u_{2xx} + \dot{x}u_{1x} \\ \dot{u}_2 = \dot{u}_{1xx} + u_{1xxx}\dot{x} - u_{2x}\dot{x} \\ u_{1x} = u_{2x} = 0, x = 0, \pi. \end{cases} \tag{1.6}$$

As written, MOVCOL2 cannot immediately solve this problem as, being a code for second order problems, it does not compute the third derivative. To overcome this we write the problem as a three component system $u_1 = u, , u_2 = u_x$, and $u_3 = u_{xx}$ then

$$\text{(ODE2)} \begin{cases} \dot{u}_1 = -u_{2xx} + \dot{x}u_{1x} \\ \dot{u}_2 = \dot{u}_{1x} - \dot{x}(u_{2x} - u_{1xx}) \\ \dot{u}_3 = \dot{u}_{2x} - \dot{x}(u_{3x} - u_{2xx}) \\ u_{1x} = u_2 = u_{3x} = 0, x = 0, \pi. \end{cases} \tag{1.7}$$

Since u_2 and u_1 are approximated with different spline functions it does not follow from $(u_1)_{xx} = (u_2)_x$ that $(u_1)_{xt} = u_{2t}$. Numerical experiments in the following will show that the error can easily grow if this term is dropped. Consequently, both the third derivative term $(u_1)_{xxx}$ and the mesh speed $\dot{x}(t)$ must be dealt with, which precludes the possibility of using MOVCOL2 to solve it. Note that this complication only arises due to the use of a moving grid.

Our goal will be to run numerical experiments in both the fixed grid and moving grid contexts and observe whether or not there are reformulations of high order problems as systems of lower order ones which are sufficiently stable (and well-posed) to warrant their use with codes designed for such low order problems. For the experiments, we

solve each problem with $atol = rtol = 10^{-6}$ for DDASSL and MMPDE6 with $\tau = 10^{-3}$. For the mesh moving strategy, we use the common choice of an arc length monitor function for Experiment I, and variations of $1/u$ for the other 2 experiments where finite time singularities at $u = 0$ has to be dealt with and arc length monitor function does not suffice.

2 Experiment I

The first problem, used in [10] as a test problem to investigate the convergence properties of MOVCOL4, is

$$\begin{cases} u_t = -\dfrac{\sin(t)}{\cos(t)+3} u_{xxxx} & x \in (0,\pi), \quad 0 < t < \pi \\ u(x,0) = 4c\cos(x) \\ u_x(0,t) = u_{xxx}(0,t) = 0 \\ u_x(\pi,t) = u_{xxx}(\pi,t) = 0, \end{cases} \tag{2.1}$$

where $c = 0.3$ and the exact solution is $u(x,t) = c(\cos(t)+3)\cos(x)$.

Due to the nonconstant coefficient of the fourth order derivative, a small time step size is needed in order to see how each discretization performs. Here, for error tolerances $atol = rtol = 10^{-6}$ in the ODE solver, we set the maximal time stepsize to be 10^{-5} which ensures that the error generated by spatial discretization dominates.

First, we solve with MOVCOL4 and only 5 mesh points. Then we convert the PDE into a system of 2 second order equations as follows:

$$\begin{cases} u_{1t} = -\dfrac{\sin(t)}{\cos(t)+3} u_{2xx} & x \in (0,\pi), \quad 0 < t < \pi \\ u_2 = u_{1xx} \\ u_1(x,0) = 4c\cos(x) \\ u_2(x,0) = -4c\cos(x) \\ u_{1x} = u_{2x} = 0, x = 0, \pi. \end{cases} \tag{2.2}$$

This cannot be solved directly with MOVCOL2 since DDASSL fails to solve the DAE system.

To make DDASSL work, we do time differentiation to the DAE system once to recover an ODE system and solve the problem

$$\begin{cases} u_{1t} = -\dfrac{\sin(t)}{\cos(t)+3} u_{2xx} & x \in (0,\pi), \quad 0 < t < \pi \\ u_{2t} = u_{1xxt} \\ u_1(x,0) = 4c\cos(x) \\ u_2(x,0) = -4c\cos(x) \\ u_{1x} = u_{2x} = 0, x = 0, \pi, \end{cases} \tag{2.3}$$

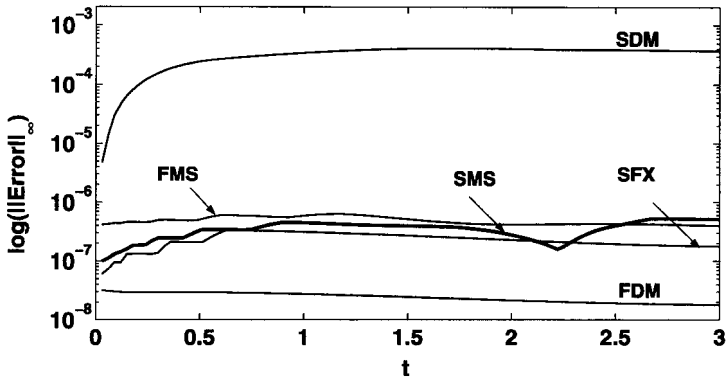


Figure 2.1 Maximum error as function of time. Error is computed as maximum difference between computed solution and exact solution at 101 equally spaced points on $[0, \pi]$ as time evolves. The discretization methods are described in Section 1.

Table 1 Comparisons of the efficiency of different methods

Method	Grid Points	System Size	Runtime (s)	Max. error
FDM	5	25	33.919	3.8862×10^{-8}
SFX	51	204	183.724	3.3390×10^{-7}
SMS	51	255	255.407	5.2592×10^{-7}
FMS	51	357	408.427	6.3577×10^{-7}

within the same parameters frame but increasing the number of mesh points to 51.

Figure 2.1 shows the error in time for the following five different approaches: FDM - Fourth order direct approach on a moving grid solved using MOVCOL4; SFX - System approach with 2 equations on a fixed grid using MOVCOL2; SDM - System approach with 2 equations on a moving grid using MOVCOL2, with no mesh speeds term involved; SMS - System approach with 2 equations on a moving grid using MOVCOL2, mesh speeds term included; FMS - System approach with 3 equations on a moving grid using MOVCOL2, mesh speeds term included. Note that when the mesh speeds term is dropped using a moving grid, large errors are generated.

Table 1 contains some technical information for this experiment, where for fixed time t, $\|\text{Error}\|_\infty = \max_{1 \le i \le npts}(|u_{\text{num}}(x_i) - u_{\text{exact}}(x_i)|$, the maximal error on the whole interval, is measured by sampling the numerical solution reconstructed using the collocation polynomials and

the exact solutions at 101 equally spaced points on $[0, \pi]$.

Both Figure 2.1 and Table 1 show that for small time step sizes, all other four approaches give good numerical results, except for the SDM approach which has dropped the mesh speed term. Among the other four approaches, the FDM approach with MOVCOL4 gives a better numerical solution than the other 3 approaches, although only 5 mesh points are used.

For the same number of mesh points, the fourth order system used by MOVCOL4 is stiffer and hence the computation is much slower than with the system approach and MOVCOL2. However, MOVCOL4 is a code of higher order accuracy, and normally fewer points are needed. This is clearly indicated in Table 1.

3 Experiment II

The non-regularized thin film equation with finite time extinction is studied in this experiment. Once again, the 2 equations approach on a moving grid without explicitly using the mesh speed term fails, and we are unable to observe the third derivative discontinuity of the solution (which is known to form). As this problem has no exact solution, we convert this fourth order problem to a system of 2 second order equations and solve it using MOVCOL2 on a nonuniform fixed grid with a large number of points (5001) in order to gather an idea of the actual time when the jump discontinuity forms and the profile of the third derivative right before the singularity forms. Since we know that the jump discontinuity happens at $x = 0$, to get good accuracy our nonuniform grid is concentrated in a very small interval $[-0.002, 0.002]$ with 4001 points equally distributed. Outside this interval, the remaining 1000 points are equally distributed on the two subintervals $[-1, -0.002)$ and $(0.002, 1]$.

The fourth order problem is solved directly using MOVCOL4, and as for the previous example, we also convert this problem into a system of 2 and 3 equations and solve it using MOVCOL2 on a moving grid with the mesh speeds involved. Numerical results are shown in Figure 3.1.

Figure 3.1 shows the computed third derivatives at specified times very close to the formation of the singularity by different approaches. For each approach, we compute the solution until the code breaks down due to the pinch-off singularity and then record the final time to which the code can integrate, which is associated with the top profile in each subfigure of Figure 3.1.

Among the four approaches, the pinch-off time and the profile of the third derivative for the SFX approach with a small stencil and for the FDM approach agree well with the numerical results found in [4, 5]. The pinch-off time for the SMS and FMS approaches agree well with the

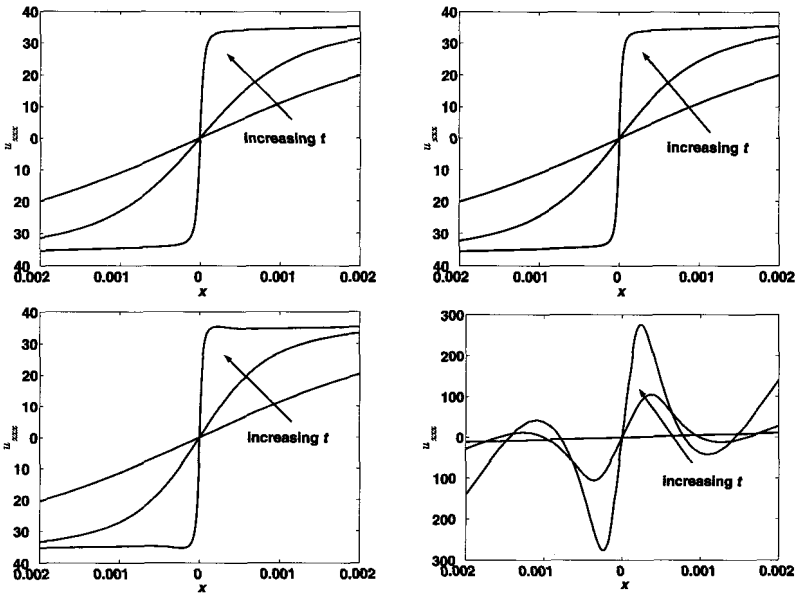


Figure 3.1 Third derivatives of regularized system at typical times right before the pinch-off singularity occurs, Upper Left: Adaptive MOVCOL4, t=0.0007300000, 0.0007302000, 0.0007302314; Upper Right: MOVCOL2 with fine static grid, t=0.0007300000, 0.0007302000, 0.0007302252; Lower Left: Adaptive MOVCOL2, 2 equations, t=0.0007300000, 0.0007302000, 0.0007302168; Lower Right: Adaptive MOVCOL2, 3 equations, t=0.0007290000, 0.0007300000, 0.0007300005.

other approaches; however, the profile of the third derivative undergoes qualitative structure changes. This is caused by the high speed of the mesh points (of magnitude around 10^5 at some points while resolving the singularity) together with the developing jump discontinuity in the third derivative. For the 3 equations system approach, where 2 mesh speed terms are involved, the error increases dramatically and the jump discontinuity can no longer be observed.

4 Experiment III

For the thin film equation which models the case where a thin film is draining at the boundary,

$$\text{(Ex III)} \begin{cases} u_t = -(u^n u_{xxx})_x & x \in (-1,1) \\ u(\pm 1, t) = 0 \\ u_x(\pm 1, t) = 0 \\ u(x,0) = (1-x^2)^3, \end{cases} \tag{4.1}$$

the zero-contact-angle boundary condition essentially overspecifies the boundary conditions for the system approach with 2 equations. The obvious formulation

$$u_1 = u, \qquad u_2 = u_x, \qquad u_3 = u_{xx}$$

leads very naturally to the boundary conditions

$$u_1(\pm 1, t) = 0, \qquad u_2(\pm 1, t) = 0, \qquad u_3(\pm 1, t) = u_{2x}(\pm 1, t).$$

It is well known that for given initial data uniqueness cannot be guaranteed as we are implicitly dealing with a free boundary problem for a fourth order equation. The function $u(x,t) = M(a^2 - x^2)_+$ is a stationary weak solution provided $\pm a \in (-1, 1)$ and M is a positive real number. It is still an open problem how to formulate the free boundary problem explicitly. In fact, given a numerical scheme which does not satisfy additional estimates which enforce the zero-contact-angle implicitly, the solution in the continuous setting is not unique and therefore no convergence results can be formulated in the case of the free boundary problem [1].

From (4.1), for the draining to happen, we have

$$\frac{\partial}{\partial t}\int_{-1}^{1} u(x,t)dx = \int_{-1}^{1} u_t(x,t)dx = -(u^n u_{xxx})|_{-1}^{1} < 0,$$

so $u_{xxx} < 0$ at $x = -1$ and $u_{xxx} > 0$ at $x = 1$. The film first spreads and then drains when u_{xxx} blows up to infinity. As MOVCOL4 requires the solution to be C^3, we expect that drainage is not possible for a direct discretization with MOVCOL4; however, numerical tests still reveal some interesting contrasts between the direct discretization and the systems approach.

For $n = 1$, we use a regularization $u^4/(10^{-14} + u^3)$ for u in the term uu_{xxx} in the main equation to avoid nonnegativity in the solution. MMPDE6 with a monitor function $1/(u + 0.1)$ is applied to control the mesh movement. When reformulated as a 3 equation system, one can only solve it using MOVCOL2 on a moving grid since it is designed to solve second order PDEs or PDE systems and will be under-determined if there are no second derivatives involved in any of the equations. In particular, on a fixed grid the solution with the 3 equation reformulation converges to a nonphysical profile.

Figure 4.1 shows that, as previously mentioned, on a moving grid with 201 points the solution converges to a parabola $u(x,t) = M((1 - \epsilon)^2 - x^2)_+$, where $\epsilon \to 0$. MOVCOL4 solves the equation until the third derivative is approximately 5×10^8. For the system approach, on a moving grid with 501 grid points MOVCOL2 can only solve the

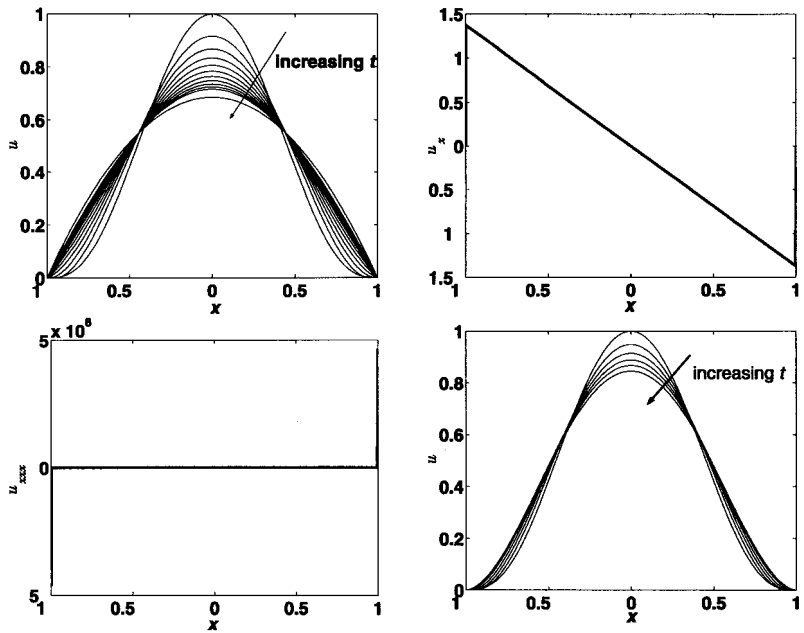


Figure 4.1 Upper Left: Solution at typical times, direct discretization, with adaptive MOVCOL4; converges to a stationary state - the parabola. Upper Right: First derivative of the direct discretization, with adaptive MOVCOL4. Lower Left: Third derivative, direct discretization, with adaptive MOVCOL4. Lower Right: Solution at typical times, system approach, adaptive MOVCOL2, codes breaks down at t=0.0049.

system until the third derivative is approximately 2×10^7. Many more points would be required for MOVCOL2 to integrate until the third derivative is comparable to 5×10^8, indicating that direct discretization with MOVCOL4 is capable of integrating further into the singularity. Meanwhile, in view of what we observed in the previous experiments, the effect of the mesh speeds worsens as the third derivative blows up.

It remains an interesting question how to compute the solution when draining occurs. Unlike the finite difference methods which have been used with some success on this problem, the boundary conditions $u = u_x = 0$ are strictly enforced with our methods.

5 Conclusion

We have demonstrated that caution must be taken when forming an ODE system by differentiating a DAE system on a moving grid. The mesh

speed term must generally be included in order to obtain good numerical results. When solving a fourth order problem, it is recommended that one solves it directly with MOVCOL4 instead of converting it into a (mathematically equivalent) system of 2 or 3 lower order equations. If the mesh moves with high speed in order to resolve spatial structure changes as time proceeds, large errors will be generated for the system approaches involving mesh speeds.

In summary, for fourth order problems, applying the moving mesh code MOVCOL2 to equivalent lower order systems does not in our experience generally warrant the associated cost nor achieve comparable accuracy to that of MOVCOL4 applied directly. Given the need for solving fourth order problems like the thin film problem with some form of adaptivity (e.g., see [10]), this begs the question of whether or not one can expect an adaptive code for the lower order systems case to be competitive at all for such problems, and if so, which adaptive method and which systems reformulation(s) would be the most promising.

References

[1] J. Becker and G. Grün, The thin-film equation: recent advances and some new perspectives, J. Phys.: Condens. Matter 17 (2005), 291–307.

[2] C. J. Budd, W. Huang, and R. D. Russell, Moving mesh methods for problems with blow-up, SIAM J. Sci. Comput. 17 (1996), 305–327.

[3] J.B. van den Berg, M. Bowen, J.R. King, M.M.A. El-Sheikh. The self-similar solution for draining in the thin film equation. Eur. J. Appl. Math. 15(2004), 329-346.

[4] A.L. Bertozzi, Loss and gain of regularity in a lubrication equation for thin viscous films, Free Boundary Problems: Theory and Applications, Proceedings of the International Colloquium on Free Boundary Problems, Toledo, Spain, June 1993 (J. I. Díaz, M. A. Herrero, A. Liñán, and J. L. Vázquez, eds.), Pitman Research Notes in Mathematics Series, vol. 323, 1995, 72–85.

[5] A. L. Bertozzi, The mathematics of moving contact lines in thin liquid films, Notice of the AMS 45(1998), 689–697.

[6] W. Huang, Y. Ren, and R. D. Russell, Moving mesh partial differential equations (MMPDEs) based on the equidistribution principle, SIAM J. Numer. Anal. 31(1994), no. 3, 709–730.

[7] W. Huang and R. D. Russell, A moving collocation method for solving time dependent partial differential equations, Appl. Num. Math. 20 (1996), no. 1-2, 101–116.

[8] L. Peletier and W. Troy, Spatial Patterns: Higher-order Models in Physics and Mechanics, Birkhäuser, New York, 2001.

[9] L. Petzold, A Description of DASSL: A Differential /Algebraic System Solver, SAND82-8637, Sandia Labs, Livermore, Cal., 1982.

[10] R. Russell, JF. Williams, X. Xu, MOVCOL4: a high resolution moving collocation method for evolutionary PDEs, submitted to SIAM J. Sci. Comput.

[11] J. M. Stockie, J. A. Mackenzie, and R. D. Russell, A moving mesh method for one-dimensional hyperbolic conservation laws, SIAM J. Sci. Comput. 22(2001), 1791–1813.

Subspace Techniques for Nonlinear Optimization*

Ya-xiang Yuan
LSEC, ICMSEC, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China.
E-mail: yyx@lsec.cc.ac.cn

Abstract

In this paper, we review various subspace techniques that are used in constructing of numerical methods for nonlinear optimization. The subspace techniques are getting more and more important as the optimization problems we have to solve are getting larger and larger in scale. The applications of subspace techniques have the advantage of reducing both computation cost and memory size. Actually in many standard optimization methods (such as conjugate gradient method, limited memory quasi-Newton method, projected gradient method, and null space method) there are ideas or techniques that can be viewed as subspace techniques. The essential part of a subspace method is how to choose the subspace in which the trial step or the trust region should belong. Model subspace algorithms for unconstrained optimization and constrained optimization are given respectively, and different proposals are made on how to choose the subspaces. As an example, we also present an interior point method based on subspace techniques.

1 Introduction

Nonlinear optimization is to minimize an objective function subject to some equality and inequality constraints, which can be written as follows.

$$\min_{x\in\Re^n} \quad f(x) \tag{1.1}$$

$$\text{subject to} \quad c_i(x) = 0, \qquad i = 1, \cdots, m_e \tag{1.2}$$

$$c_i(x) \geq 0, \qquad i = m_e + 1, \cdots, m, \tag{1.3}$$

where m and m_e are two non-negative integers satisfying $m \geq m_e$. Normally, $f(x)$ and $c_i(x)(i = 1, \cdots, m)$ are continuously differentiable functions. We use the notations $g(x) = \nabla f(x)$ and $a_i(x) = \nabla c_i(x)$.

*The author is partly supported by NSF of China (No. 10231060).

Large scale optimization problems (n is very large) have attracted much attention from researchers in the recent years. A very good recent review paper was given by Gould, Orban and Toint[5], where the authors discussed the main approaches for handling large scale problems. The step computation, active set, gradient projection, and interior point method were discussed there.

One of the straightforward technique for large scale problems is to solve the large scale subproblems efficiently by approximate methods. The overall cost for building a house will be reduced if each brick is cheaper. For example, the truncated conjugate gradient (CG) method is a very successful algorithm for the single ball trust region subproblem:

$$\min_{d\in\Re^n} \; Q(d) = g^T d + \frac{1}{2} d^T B d \tag{1.4}$$

$$s.t. \;\; \|d\|_2 \le \Delta. \tag{1.5}$$

The truncated CG method proposed independently by Steihaug[8] and Toint[11] defines an approximate solution by applying the standard conjugate gradient method until reaching the trust region bound or obtaining the exact solution. It was proved by Yuan[17] that:

$$Q(0) - Q(d_{TCG}) \ge \frac{1}{2}[Q(0) - Q(d^*)] \tag{1.6}$$

where d_{TCG} is the solution obtained by the truncated conjugate gradient method and d^* is the solution of (1.4)-(1.5).

Active set is also an efficient approach to large scale problems, particularly for problems having many redundant constraints. In a sequential quadratic programming (SQP) algorithm, the subproblem at the k-th iteration is

$$\min_{d\in\Re^n} \; Q_k(d) = \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d \tag{1.7}$$

$$s.t. \;\; c_i(x_k) + a_i^T(x_k)^T d = 0, \qquad i = 1, \cdots, m_e, \tag{1.8}$$

$$c_i(x_k) + a_i^T(x_k)^T d \ge 0, \qquad i = m_e + 1, \cdots, m. \tag{1.9}$$

By active set technique, instead of solving the above subproblem, we can solve the following subproblem:

$$\min_{d\in\Re^n} \; Q_k(d) = \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d \tag{1.10}$$

$$s.t. \;\; c_i(x_k) + a_i(x_k)^T d = 0, \qquad i \in \mathcal{A}_k, \tag{1.11}$$

where $\mathcal{A}_k$ is an active set which is a subset of $\{1, 2, \cdots, m\}$ and is an approximation to the active set at the solution. When the number of

elements of $\mathcal{A}$ is significantly smaller that m, the active set subproblem should be much easier to solve than the original subproblem.

One of our motivations for suggesting subspace algorithms for nonlinear optimization is the unbalance property shared by most line search algorithms. Consider any line search method having the form:

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.12}$$

where d_k is the search direction and $\alpha_k > 0$ is the step-length computed by certain line search technique. The search direction d_k is usually computed by solving a subproblem which is an approximation to the original nonlinear problem. Thus, each iteration of a line search algorithm composites of two parts, one is to find a d_k in the whole n dimensional space, while the other part is to search a suitable step-length in the fixed one dimensional space spanned by the computed d_k. Thus, the overall algorithm swings between n dimensional search and one dimensional search alternately.

Another motivation for us to consider subspace algorithms is that quite a few well known existing algorithms essentially have certain subspace features. For example, the conjugate gradient method uses a search direction in a two dimensional subspace spanned by the steepest descent direction and the previous step, the dog-leg method computes a step that is a convex combination of the steepest descent direction and the Newton's direction, and limited memory quasi-Newton algorithms produce search directions spanned in lower dimensional spaces.

This paper is organized as follows. In the next section, we will give some examples of algorithms that have certain subspace structures. In section 3, a model algorithm using subspace approach for unconstrained optimization is given and some possibilities for choices of the subspaces are also discussed. In section 4, subspace techniques for constrained optimization are presented, together with a subspace interior point trust region algorithm for box constrained optimization as an example. Finally, a brief discussion is given.

2 Examples of Subspace Approaches

It is well-known that nonlinear conjugate gradient methods use a linear combination of the steepest descent direction $-g_k$ and the previous search direction d_{k-1} to define the new search direction:

$$d_k = -g_k + \beta_k d_{k-1}. \tag{2.1}$$

Thus, one of the central tasks of nonlinear conjugate gradient methods has been how to define the suitable β_k based on certain conjugate

principles. The four leading contenders are the FR, HS, PRP and DY methods[7].

Instead of the conjugate property, Stoer and Yuan[9] suggests to look at the conjugate gradient method from the subspace point of view. As a conjugate gradient method uses a β_k to define d_k and then a step-size α_k to set $x_{k+1} = x_k + \alpha_k d_k$, no matter whatever β_k and α_k are used, the increment in the iterative point will be a linear combination of $-g_k$ and d_{k-1}. Hence it is natural to ask how to find a best point in the 2-dimensional subspace spanned by $-g_k$ and d_{k-1}. Namely, we can consider a model subproblem

$$\min_{d \in Span\{-g_k, d_{k-1}\}} Q_k(d) \approx f(x_k + d). \tag{2.2}$$

Let d_k be the solution of the above 2-dimensional subproblem. Stoer and Yuan[9] presents a successive 2-dimensional search algorithm, which is an example of algorithms using subspace models.

Another famous optimization technique has the subspace nature is the limited memory quasi-Newton method. Quasi-Newton updates (for example, see [3] and [10]) have the form

$$B_k = U(B_{k-1}, s_{k-1}, y_{k-1}) \tag{2.3}$$

which satisfies

$$B_k s_{k-1} = y_{k-1}, \tag{2.4}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})$. An example is the famous BFGS method

$$B_k = B_{k-1} - \frac{B_{k-1} s_{k-1} s_{k-1}^T B_{k-1}}{s_{k-1}^T B_{k-1} s_{k-1}} + \frac{y_{k-1} y_{k-1}^T}{s_{k-1}^T y_{k-1}}. \tag{2.5}$$

Limited memory quasi-Newton updates the approximate Hessian repeatedly:

$$B_k^{(i)} = U(B_k^{(i-1)}, s_{k-m-1+i}, y_{k-m-1+i}), \quad i = 1, \cdots, m, \tag{2.6}$$

with $B_k^{(0)} = \sigma_k I$. For example, see [6]. There are different formulae for σ_k with one particular choice being $s_{k-1}^T y_{k-1} / y_{k-1}^T y_{k-1}$. It can be shown that the limited memory quasi-Newton matrix can be written as

$$B_k = B_k^{(m)} = \sigma_k I + [S_k \quad Y_k] T_k \begin{bmatrix} S_k^T \\ Y_k^T \end{bmatrix}, \tag{2.7}$$

where T_k is a $2m \times 2m$ matrix and

$$[S_k \quad Y_k] = [s_{k-1}, s_{k-2}, \cdots, s_{k-m}, y_{k-1}, y_{k-2}, \cdots, y_{k-m}] \in \Re^{n \times 2m}. \tag{2.8}$$

In a line search type method, we have that $s_k = \alpha_k d_k = -\alpha_k B_k^{-1} g_k$, while for a trust region type algorithm one would have $s_k = -(B_k + \lambda_k I)^{-1} g_k$. Therefore, in either case, it follows that

$$\begin{aligned} s_k &= -\left(\rho_k I + [S_k \quad Y_k] T_k \begin{bmatrix} S_k^T \\ Y_k^T \end{bmatrix}\right)^{-1} g_k \\ &\in Span\{g_k, s_{k-1}, \cdots, s_{k-m}, y_{k-1}, \cdots, y_{k-m}\}. \end{aligned} \tag{2.9}$$

Consequently, we have shown that limited memory quasi-Newton algorithms no matter with line search or trust region will always produce a step in the subspace $Span\{g_k, s_{k-1}, \cdots, s_{k-m}, y_{k-1}, \cdots, y_{k-m}\}$. Indeed, a trust region algorithm using this subspace is given by Wang, Wen and Yuan[13].

Even for the standard quasi-Newton updates, we can also establish subspace property results. The following is a result for trust region type algorithms.

Lemma 2.1. *(Wang and Yuan, [14]) Suppose $B_1 = \sigma I$, $\sigma > 0$. The matrix updating formula is any one chosen from amongst SR1, PSB and Broyden family, and B_k is the k-th updated matrix. s_k is the solution of*

$$\min_{d \in \Re^n} \; g_k^T d + \frac{1}{2} d^T B_k d \tag{2.10}$$

$$s.t. \;\; \|d\|_2 \le \Delta_k, \tag{2.11}$$

$x_{k+1} = x_k + s_k$, $g_k = \nabla f(x_k)$. Let $\mathcal{G}_k = Span\{g_1, g_2, \cdots, g_k\}$. Then

$$s_k \in \mathcal{G}_k, \tag{2.12}$$

and we have

$$B_k z \in \mathcal{G}_k, \quad B_k u = \sigma u \tag{2.13}$$

for any $z \in \mathcal{G}_k$, $w \in \mathcal{G}_k^{\perp}$.

The above lemma is an extension of similar results for line search type algorithms, which were discussed by Gill and Leonard[4] and Vlcek and Luksan[12].

A modification to the standard trust region subproblem (1.4)-(1.5) is given by Burdakov and Yuan[1]. Assume that the matrix B is generated by limited memory quasi-Newton updates. Namely we have

$$B = \sigma I + PDP^T, \qquad P \in \Re^{n \times l}. \tag{2.14}$$

Instead of using the Euclidean norm as in (1.5), Burdakov and Yuan[1] suggested the following norm

$$\|d\|_{P,\infty} = \max\{\|P^T d\|_\infty, \; \|P_\perp^T d\|_2\}, \tag{2.15}$$

where $P_\perp^T$ is a matrix such that $P_\perp^T P = 0$ and $(P, P_\perp)$ is nonsingular. Such a choice of the norm makes the subproblem

$$\min_{\|d\|_{P,\infty} \le \Delta} \quad Q(d) \tag{2.16}$$

easy to solve.

3 A model subspace algorithms

Based on the above observations, we can suggest a model subspace algorithm for unconstrained optimization, which is a slightly modification of the standard trust region algorithm for unconstrained optimization (for example, see Yuan[16]).

Algorithm 3.1. *(A model subspace algorithm for unconstrained optimization)*

Step 1 Given x_1, Define $\mathcal{S}_1$, $\epsilon > 0$, $k := 1$.

Step 2 Solve a subspace subproblem:

$$\min_{d \in \mathcal{S}_k} Q_k(d) = g_k^T d + \frac{1}{2} d^T B_k d \tag{3.1}$$

obtaining s_k. If $\|s_k\| \le \epsilon$ then stop.

Step 3 Define

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } f(x_k + s_k) < f(x_k) \\ x_k & \text{otherwise} \end{cases} \tag{3.2}$$

Step 4 Generate $\mathcal{S}_{k+1}$ and $Q_{k+1}(d)$.

Step 1 $k := k + 1$, Go to Step 2.

The main difference between the above algorithm and standard whole space algorithms is the constraint for the step s_k to be in the subspace $\mathcal{S}_k$. Thus, the key issue here is how to choose the subspace $\mathcal{S}_k$.

One obvious choice for the subspace $\mathcal{S}_k$ is a generalization of the 2-dimensional subspace studied by Stoer and Yuan[9], namely $\mathcal{S}_k = Span\{-g_k, s_{k-1}, \cdots, s_{k-m}\}$. As all the points in $\mathcal{S}_k$ can be expressed by

$$d = \alpha g_k + \sum_{i=1}^{m} \beta_i s_{k-i},$$

using the following approximations

$$s_{k-i}^T \nabla^2 f(x_k) s_{k-j} \approx s_{k-i}^T y_{k-j}, \quad s_{k-i}^T \nabla^2 f(x_k) g_k \approx y_{k-i}^T g_k \tag{3.3}$$

we can write the quadratic function $Q_k(d) \approx f(x_k + d)$ in the subspace by a quadratic function of $\alpha, \beta_1, \cdots, \beta_m$:

$$\bar{Q}_k(\alpha, \beta_1, \cdots, \beta_m) = (-\|g_k\|^2, g_k^T s_{k-1}, \cdots, g_k^T s_{k-m})^T \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$$

$$+\frac{1}{2} \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T \begin{pmatrix} \rho_k & -g_k^T y_{k-1} & \cdots & -g_k^T y_{k-m} \\ -g_k^T y_{k-1} & y_{k-1}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -g_k^T y_{k-m} & y_{k-m}^T s_{k-1} & \cdots & y_{k-m}^T s_{k-m} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad (3.4)$$

where ρ_k should be an approximation to $g_k^T \nabla^2 f(x_k) g_k$. There are many possible ways to approximate it. First, we can write this value in a different form:

$$g_k^T \nabla^2 f(x_k) g_k = \frac{1}{\cos^2 \theta_k} \frac{(s_{k-1}^T \nabla^2 f(x_k) g_k)^2}{s_{k-1}^T \nabla^2 f(x_k) s_{k-1}} \approx \frac{1}{\cos^2 \theta_k} \frac{(y_{k-1}^T g_k)^2}{s_{k-1}^T y_{k-1}}, \quad (3.5)$$

where θ_k is the angle between $(\nabla^2 f(x_k))^{1/2} s_{k-1}$ and $(\nabla^2 f(x_k))^{1/2} g_k$. Similar to [9], due to the mean value to $\cos^2 \theta = 1/2$, we can let

$$\rho_k = 2 \frac{(y_{k-1}^T g_k)^2}{s_{k-1}^T y_{k-1}}. \quad (3.6)$$

Because $g_k^T \nabla^2 f(x_k) g_k$ is the magnitude of the Hessian matrix in the direction g_k, which can be estimated by the average over the m directions $s_{k-i} (i = 1, \cdots, m)$, it is reasonable to let

$$\rho_k = \frac{\|g_k\|_2^2}{m} \sum_{i=1}^{m} \frac{s_{k-i}^T y_{k-i}}{s_{k-i}^T s_{k-i}}. \quad (3.7)$$

Of course, we can use a limited memory matrix B_k and set $\rho_k = g_k^T B_k g_k$. If an extra function value at $x_k + t g_k$ is calculated, one can also let

$$\rho_k = \frac{2(f(x_k + t g_k) - f(x_k) - t\|g_k\|_2^2)}{t^2}. \quad (3.8)$$

Another possible choice for $\mathcal{S}_k$ is $Span\{-g_k, y_{k-1}, \cdots, y_{k-m}\}$. If we express the solution of (3.1) as

$$d = \alpha g_k + \sum_{i=1}^{m} \beta_i y_{k-i} = W_k \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad (3.9)$$

the coefficients $\alpha, \beta_1, \cdots, \beta_m$ should be

$$-\left[W_k^T \nabla^2 f(x_k) W_k\right]^{-1} W_k^T \nabla f(x_k).$$

One approach to estimate the above vector is to replace $\left[W_k^T \nabla^2 f(x_k) W_k\right]^{-1}$ by $\left[W_k^T (\nabla^2 f(x_k))^{-1} W_k\right]$. In this way, all we need to do is to estimate $g_k (\nabla^2 f(x_k))^{-1} g_k$. Estimations for such a value can be obtained by similar techniques that were just discussed above for estimating $g_k^T \nabla^2 f(x_k) g_k$.

Let P_k be a matrix whose columns spanned $\mathcal{S}_k$. The solution of (3.1) d_k can be written as $P_k z_k$, where z_k solves

$$P_k^T g_k + P_k^T B_k P_k z = 0. \tag{3.10}$$

A generalization of the subspace subproblem (3.1) is to find a step d such that

$$\bar{P}_k^T (g_k + B_k d) = 0, \quad d \in \mathcal{S}_k, \tag{3.11}$$

where $\bar{P}_k^T$ is a projection from $\Re^n$ to a lower dimensional subspace $\bar{\mathcal{S}}_k$. Normally (3.11) exists a solution if the dimension of $\bar{\mathcal{S}}_k$ is not larger than that of $\mathcal{S}_k$. (3.1) is a special case of (3.11) when $\bar{\mathcal{S}}_k = \mathcal{S}_k$. It is easy to see that the Gauss-Seidel method for linear equations is just the case when $\mathcal{S}_k = \bar{\mathcal{S}}_k$ are the one dimensional subspaces spanned by the co-ordinate directions.

4 Subspace Techniques for Constrained Optimization

Now we discuss subspace techniques for constrained optimization. For simplicity, we only consider equality constrained problems here, as the techniques discussed below can be easily extended to general inequality constraints. The equality constrained optimization has the form:

$$\min_{x \in \Re^n} \; f(x) \tag{4.1}$$

$$s.t. \;\; c(x) = 0. \tag{4.2}$$

An SQP subproblem for the above problem is

$$\min_{d \in \Re^n} \; Q_k(d) \tag{4.3}$$

$$s.t. \;\; c(x_k) + A_k^T d = 0. \tag{4.4}$$

The null space approach obtains a solution of the above subproblem by computing a Range Space step (vertical step) v_k and a Null Space step

(horizontal step) h_k and setting $d_k = h_k + v_k$. For example, let the Q-R factorization of A_k be

$$A_k = [Y_k \quad Z_k] \begin{bmatrix} R_k \\ 0 \end{bmatrix}, \tag{4.5}$$

we can let $v_k = -Y_k R_k^{-T} c_k$ and $h_k = Z_k z$ with z being the solution of

$$\min Q_k(v_k + Z_k z). \tag{4.6}$$

One of the nice properties of the SQP subproblem is that its solution d_k is a superlinearly convergent step, namely

$$x_k + d_k - x^* = o(\|x_k - x^*\|) \tag{4.7}$$

if B_k is a good approximate to the Hessian of the Lagrangian. Unfortunately, Marotos effect may happen. That is, even when relation (4.7) holds, it is possible that

$$f(x_k + d_k) > f(x_k), \qquad \|c(x_k + d_k)\| > \|c(x_k)\|. \tag{4.8}$$

The increases in both the objective function and constraint violation will make all the standard globalization techniques such as line search, trust region, and filter reject the new point $x_k + d_k$!

A remedy for the Marotos effect is the second order correction step technique which solves another subproblem

$$\min_{d \in \Re^n} \; Q_k(d_k + d) \tag{4.9}$$

$$s.t. \;\; c(x_k + d_k) + A_k^T d = 0 \tag{4.10}$$

to obtain an additional step $\hat{d}_k$. Under certain conditions, it can be shown that $\hat{d}_k = O(\|d_k\|^2)$, therefore we call $\hat{d}_k$ a second order correction step. Such a second order step would make the new point $x_k + d_k + \hat{d}_k$ acceptable.

Though the second order correct step is a very good technique having both strong theoretical properties and nice computational performance, it has an undesirable feature explained as follows. From the subspace point of view, the standard QP step d_k is the sum of two steps, a range space step v_k and a null space step h_k. When Marotos effect happens, it can be proved that the second order correction step $\hat{d}_k$ is almost a range space step as well. The range space steps v_k and $\hat{d}_k$ are quadratical steps, because basically the Newton-Raphson method is used to compute the range space step. The null space step h_k is normally a quasi-Newton step as B_k is usually updated by quasi-Newton formulae. Though the quasi-Newton step is superlinearly convergent, its Q-order of convergence can

ba arbitrary close to 1 (see [15]). Thus, the convergence rate of the null space step can be much slower than the range space step. This is not surprising as it is quite normal to observe iterate points sinking into the feasible region. Consequently, the second order correction step technique leads to two range space steps and one null space step. There exists a strong unbalance, because in the range space we apply a faster algorithm (Newton's method) and make two iterations while in the null space we apply a slower algorithm for only one iteration. Just suppose there were a chariot with a larger wheel on one side and a smaller wheel on the other side. Obviously, such a chariot would be far from perfect. What the second order correction technique would make it even worse is to install a gearbox to this chariot so that the smaller wheel turns only one round while the larger wheel turns two rounds. Such a chariot has been driven on the road of optimization for two decades without being noticed.

The above paragraph shows that subspace analysis will provide us additional insight into an algorithm. At least, for null space type algorithms, we need to give another think on the null space step and the range space step.

Similar to Algorithm 3.1, we can give a subspace algorithm for constrained problem as follows.

Algorithm 4.1. *(A model subspace method for equality constrained optimization)*

Step 1 *Given* x_1, *Define* $\mathcal{S}_1$, σ_1, $\epsilon > 0$, $k = 1$.

Step 2 *Solve a subspace subproblem:*

$$\min \ Q_k(d) \tag{4.11}$$

$$s.t. \ c_k + A_k^T d = 0, \qquad d \in \mathcal{S}_k. \tag{4.12}$$

obtaining s_k. *If* $\|s_k\| \leq \epsilon$ *then stop.*

Step 3 *Define*

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } P(x_k + s_k, \sigma_k) < P(x_k, \sigma_k) \\ x_k & \text{otherwise} \end{cases} \tag{4.13}$$

where $P(x, \sigma)$ *is a penalty function.*

Step 4 *Generate* σ_{k+1}, $\mathcal{S}_{k+1}$ *and* $Q_{k+1}(d)$.

Step 1 $k := k + 1$, *Go to Step 2.*

Similar to the unconstrained case, there are many choices for $\mathcal{S}_k$. For example, a possible choice for $\mathcal{S}_k$ is $Span\{-g_k, s_{k-1}, \cdots, s_{k-m}, -\nabla c_{k_i}(x_k)\}$, where $|c_{k_i}(x_k)| = \|c(x_k)\|_\infty$.

Another subspace approach is try to find a subspace solution to the linear system:

$$\begin{bmatrix} B_k & A_k \\ A_k^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \lambda \end{bmatrix} = - \begin{bmatrix} g_k \\ c_k \end{bmatrix}. \tag{4.14}$$

Namely we use a subspace $\mathcal{S}_k$ to compute d_k by either solving

$$\min_{d \in \mathcal{S}_k} \left\| \begin{bmatrix} B_k & A_k \\ A_k^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \lambda \end{bmatrix} + \begin{bmatrix} g_k \\ c_k \end{bmatrix} \right\|_2^2, \tag{4.15}$$

or solving

$$\min_{d \in \mathcal{S}_k} \left\| P_k \left(\begin{bmatrix} B_k & A_k \\ A_k^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \lambda \end{bmatrix} + \begin{bmatrix} g_k \\ c_k \end{bmatrix} \right) \right\|_2^2, \tag{4.16}$$

where P_k is a projection from $\Re^{n+m}$ to some lower dimensional subspace.

For box constrained optimization, Yuan[18] presents a trust region interior point method based on subspace technique, which is an generalization of the algorithm given by Coleman and Li[2].

Consider the box constrained problem

$$\min \ f(x) \tag{4.17}$$

$$s.t. \ \ l \le x \le u. \tag{4.18}$$

Denote the current iterate point at the k-th iteration by x_k, and let

$$\Omega_k = \mathrm{Diag}((\omega_k)_1, \cdots, (\omega_k)_n), \tag{4.19}$$

where

$$(\omega_k)_i = \min\{(x_k)_i - l_i, u_i - (x_k)_i\} \tag{4.20}$$

and $\Delta_k = \min_{1 \le i \le m} (\omega_k)_i = \mathrm{Dist}(x_k, \Gamma)$.

The subspace subproblem given by Yuan[18] is defined by

$$\min \ \phi_k(d) = g_k^T d + \frac{1}{2} d^T B_k d \tag{4.21}$$

$$s.t. \ \ d^T \Omega_k^{-1} d \le \rho_k \Delta_k, \qquad d \in \mathcal{S}_k, \tag{4.22}$$

where $\rho_k \in (0,1)$ and the subspace $\mathcal{S}_k$ updated from iteration to iteration.

Algorithm 4.2. *(An interior point method with subspace techniques)*

Step 1 *Given an interior point x_1, $\epsilon > 0$, given a subspace $\mathcal{S}_1$, $k := 1$.*

Step 2 *Solve (4.21)-(4.22) obtaining s_k.*
if $\|s_k\| \le \epsilon$ then stop.

Step 3 Let

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } r_k > 0 \\ x_k & \text{otherwise} \end{cases} \tag{4.23}$$

$$\rho_{k+1} = \begin{cases} \frac{\rho_k+1}{2} & \text{if } r_k \geq 0.1 \\ \frac{\rho_k}{2} & \text{otherwise} \end{cases} \tag{4.24}$$

Step 4 Define the subspace $\mathcal{S}_{k+1}$ for the next interation $k := k+1$, go to Step 2.

Yuan[18] shows that the above algorithm converges to a stationary point of (4.17)-(4.18) if $\Omega_k g_k \in \mathcal{S}_k$ for all k.

5 Discussions

Various subspace techniques for constructing numerical methods for nonlinear optimization have been discussed in the paper. Subspace techniques are suitable for large scale problems, particularly when function and gradient values are difficult to compute and when functions are highly nonlinear. For constrained optimization, Subspace depending on constraints are not easy to be defined. We believe that subspace methods will attract more and more attention in the future.

References

[1] O. Burdakov and Y. Yuan, Limited memory trust region methods for solving large scale unconstrained optimization, Preprint, 2002.

[2] T. Coleman and Y.Y. Li, An interior point trust region approach for nonlinear minimization subject to bounds, SIAM J. Optim. 6(1996) pp. 418–445.

[3] R. Fletcher, Practical Methods of Optimization, Second Edition, (John Wiley and Sons, 1987).

[4] P.E. Gill and M.W. Leonard, Reduced-Hessian quasi-Newton methods for unconstrained optimization, SIAM J. Optim., 12(2001), pp.209–237.

[5] N.I.M. Gould, D. Orban and Ph.L. Toint, Numerical methods for large-scale nonlinear optimization, Acta Numerica (2005), pp. 299–361.

[6] D.C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical Programming 45(1989), pp. 503–528.

[7] J.L. Nazareth, Conjugate-gradient methods, in C.A. Floudas and P.M. Pardalos, eds., Encyclopedia of Optimizaiton, Vol 1. (Kluwer, 2001) pp.319–323.

[8] T. Steihaug, The conjugate gradient method and trust regions in large scale optimization, SIAM J. Numerical Analysis, 20(1983), pp. 626–637.

[9] J. Stoer and Y. Yuan, A subspace study on conjugate gradient algorithms, ZAMM Z. angew. Math. Mech., 75(1995),pp. 69–77.

[10] W.Y. Suan and Y. Yuan, Optimization Theory and Mehtods: Nonlinear Programming, Springer Optimization and Its Application, Vol. 1, (Springer, 2006).

[11] Ph.L. Toint, Towards an efficient sparsity exploiting Newton method for minimization, in: I. Duff, ed. Sparse Matrices and Their Uses, (Academic Press, 1981), pp. 57–88.

[12] J. Vlček and L. Lukšan, New variable metric methods for unconstrained minimization covering the large-scale case, Technical report No. V876, October 2002, Institue of Computer Science, Academy of Sciences of the Czech Republic.

[13] Z.H. Wang, Z.W. Wen and Y. Yuan, A subspace trust region method for large scale unconstrained optimization, in: Y. Yuan, ed. Numerical Linear Algebra and Optimization (Science Press, 2004) pp. 265–274.

[14] Z.H. Wang and Y. Yuan, A subspace implementation of a quasi-Newton trust region method for unconstrained optimization, Report, ICMSEC, AMSS, Chinese Academy of Sciences, 2004.

[15] Y. Yuan, On the least Q-order of convergence of variable metric algorithms, IMA J. Numerical Analysis 4(1984), pp. 233–239.

[16] Y. Yuan, A review of trust region algorithms for optimization, in ICM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics, J. M. Ball and J. C. R. Hunt, eds., (Oxford University Press, 2000), pp. 271–282.

[17] Y. Yuan, On the truncated conjugate gradient method, Math. Prog., 87(2000), pp. 561–571.

[18] Y. Yuan, Subspace techniques for nonlinear optimization, talk presented at the International Conference on High Performance Scientific Computing, March 6-10, 2006.

Some Topics in Industrial and Applied Mathematics

Series in Contemporary Applied Mathematics
CAM 8

The Shanghai Forum on Industrial and Applied Mathematics was organized in May 2006 on the occasion that many famous industrial and applied mathematicians gathered in Shanghai from different countries to participate in the Officers' Meeting and the Board Meeting of the ICIAM (International Council for Industrial and Applied Mathematics). This volume collects the material covered by the majority of the lectures of which reflects panoramically recent results and trends in industrial and applied mathematics. This book will be very useful for graduate students and researchers in industrial and applied mathematics.

Higher Education Press
www.hep.com.cn

World Scientific
www.worldscientific.com

6552 hc
ISBN-13 978-981-270-934-9
ISBN-10 981-270-934-7
9 789812 709349