Boletín de la Sociedad Española de Matemática Aplicada S$\vec{e}$MA

Página web de S$\vec{e}$MA

`http://www.uca.es/sema/`

Dirección Editorial: Boletín de S$\vec{e}$MA. Dpto. de Matemáticas. Universidad de Oviedo. Avda. de Calvo Sotelo, s/n. 33007-Oviedo. `boletin_sema@orion.ciencias.uniovi.es`

ISSN 1575-9822
Depósito Legal: AS-1442-2002

Imprime: Grupo Bitácora. C/ Instituto, 17, Entresuelo. 33201 Gijón (Asturias)

Diseño de portada: Ana Cernea

Estimados socios:

Al iniciar la serie de publicaciones del Boletín de S$\vec{e}$MA, correspondientes al año 2003, deseamos recordar que a partir de ahora esta publicación tendrá una periodicidad trimestral (Marzo, Junio, Septiembre y Diciembre). Enumeramos aquí también las distintas secciones que se han ido configurando hasta este momento:

- Mensajes y opiniones.
- Notas biográficas y efemérides.
- Artículos.
- Matemáticas e Industria.
- Educación matemática.
- Resúmenes de libros, tesis, artículos y "software".
- Noticias.
- Anuncios.

Agradecemos profundamente a cuantos autores han participado, pues ellos son los que hacen posible esta publicación. Reiteramos, una vez más, nuestro llamamiento a una participación más activa en todas las secciones del Boletín.

Grupo Editor
`boletin_sema@orion.ciencias.uniovi.es`

Review on adaptativity in finite elements and multilevel methods

B. Achchab[1], O. Axelsson[2], L. Laayouni[3] and A. Souissi[4]

[1,3] Faculté des Sciences Juridiques, Economiques et Sociales,
Université Hassan I$^{\text{er}}$-Settat, Maroc
[2,3] Department of Mathematics, University of Nijmegen,
Netherlands
[4] Département de Mathématiques et d'Informatique, Faculté des
Sciences, Université Mohamed V-Agdal, Maroc

[1]achchab@hotmail.com, [2]axelsson@sci.kun.nl,
[3]laayouni@sci.kun.nl, [4]souissi@fsr.ac.ma

Abstract

In this summary, the contents of the five papers that comprise the thesis [34] are reviewed. Adaptivity in finite elements and multilevel methods are the main focus of the thesis [34]. They have a most important aim in common, which is to lead to a robust and efficient strategy of resolution when dealing with partial differential equations. The work presented in [34] is intended to contribute to providing robust and efficient a posteriori error estimates and preconditioning methods for a large class of different problems.

Key words: *A posteriori error estimates, anisotropic finite elements, fluid-structure problems, multilevel methods, preconditioning.*
AMS subject classifications: *65N15, 65N25, 65N30, 65N50, 65N55.*

1 Introduction

Chapter 1 of [34] is devoted to the study of indefinite problems satisfying a discrete variant of Gårding inequality. It presents a general framework for a priori and a posteriori estimation analysis for this class of problems. Some relevant applications are briefly indicated. This framework is used for developing hierarchical a posteriori error estimater for a coupled problem in Chapter 2 of [34]. The summarize of this work leads to the publication [7].

Fecha de recepción: 19 de Junio de 2002

Chapter 2 of the thesis provides a detailed analysis of a fluid-structure interaction problem. It deals with an elastoacoustic problem. Pressure and displacement variables are used respectively for the fluid and the structure. Using the framework developed in Chapter 1 we give a complete a priori and a posteriori error analysis for an approximation by the element $P_1 - [P_1]^2$. This study was summarized as a paper [4].

Some boundary value problems yield so-called anisotropic solutions (e.g. with boundary layers). Then anisotropic meshes (i.e. meshes with stretched elements) can be advantageous to reduce the computational effort substantially. Chapter 3 of [34] is devoted to a posteriori error estimation for anisotropic tetrahedral or triangular finite element meshes for a convection-diffusion problem with dominant convection. It provides an a posteriori residual error estimator for a stabilized finite elements formulation. We also present some numerical experiments for the bidimensional case. This work leads to a publication [5, 1, 2].

The constant γ of the strengthened Cauchy-Bunyakowski-Schwarz (C.B.S) inequality plays a fundamental role in the convergence rate of multilevel iterative methods. The main purpose of Chapter 4 of [34] is to give an estimate of the constant γ for a three dimensional elasticity system. The theoretical results obtained are of practical importance for the successful implementation of the finite element method to large scale modeling of complicated structures. They allow us to construct optimal order algebraic multilevel iterative solvers for a wide class of real-life elasticity problems. This work was as summarized as a published paper [3].

In Chapter 5 in [34] the recently proposed algebraic multilevel iteration method for iterative solution of systems of partial differential equations is considered. Based on a special approximation of the blocks corresponding to the new nodes at every discretization level, an optimal order preconditioner with respect to the arithmetic cost independent of both the parameters of the discretization and the coefficients of the system is constructed. The results are derived in the framework of a two-dimensional hierarchical basis using a linear finite element discretization on arbitrary triangular meshes. This study leads to a publication [6].

In this paper we will concentrate only on the last four chapters of the thesis [34].

2 A posteriori error estimations for fluid structure interaction problem

The numerical simulation of coupled phenomena has made great progress in recent years. This development is due in particular to the performance increase of our computers. Among these coupled phenomena one finds fluid-structure interaction problems. They describe a mobile, rigid or deformable structure and a liquid or gas flowing around or against a part of the structure. These phenomena are known as coupled problems, because the evolution of each of the two elements depends on the other. Thus, for instance, the shape of the sail of a boat depends on the flow of air around the sail. On the other hand, this flow depends on the shape of the sail. We can cite a great number of examples of the same type. Among these, one can mention for instance the hydroelastic phenomena (fluid in liquid phase): flows around a ship, submarine, dam in a port or piles of bridge; fluid flows in the interior of pipes; movements of fluids in a reservoir, blood flows in the arteries, etc. We distinguish also the aeroelastic phenomena, where the fluid is in the gas phase: flows around air vehicles (planes, missiles, etc.) and vehicles (high-speed trains, cars, etc.) wind influence on flexible constructions (suspension bridges, cooling agents of nuclear thermal power station, etc.).

We consider an elastic body in a rectangular domain Ω_s containing a fluid cavity represented by the domain Ω_f; the two mediums are separated by an interface Σ. The system is schematized by Figure 1. Under the assumption of small movements the boundary value problem is given by the following equations system:

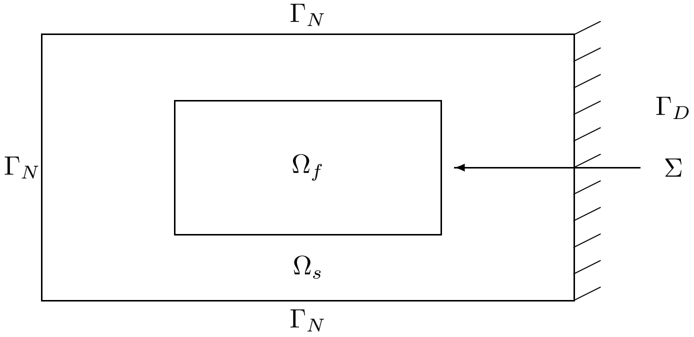


Figure 1: Bidimensional fluid-structure interaction problem

$$\left\{\begin{array}{rcll} -c^2\Delta p &=& \omega^2 p & \text{in } \Omega_f, \\ -\operatorname{div}\sigma(u) &=& \rho_s\omega^2 u & \text{in } \Omega_s, \\ \sigma(u).n &=& g & \text{on } \Gamma_N, \\ u &=& 0 & \text{on } \Gamma_D, \\ -\dfrac{1}{\rho_s}p.n+\sigma(u).n &=& 0 & \text{on } \Sigma, \\ \dfrac{\partial p}{\partial n} &=& \rho_f\omega^2 u.n & \text{on } \Sigma, \end{array}\right.$$

where p is the pressure in the fluid domain Ω_f and u is the displacement field of the structure Ω_s, ρ_f and c denote, respectively, the density of the fluid and the velocity of the sound in vacuum, g is a surface density of forces given on Γ_N,

ρ_s is the volumic density of the structure, n is the unit normal vector directed towards the outside of Ω_s, ω is the frequency, σ is the tensor of the constraints of Cauchy related to the linearized tensor of the deformations ε by the following law of behavior:

$$\sigma_{ij}(u) = \sum_{k,l=1}^{2} a_{ijkl}\varepsilon_{kl}(u) \qquad \text{with } i,j = 1,2,$$

where a_{ijkl} $(i,j,k,l = 1,2)$ are the coefficients of elasticity which have the usual properties of positivity and symmetry with respect to the permutation of indices. A natural way to tackle this problem consists in describing the fluid by the field of pressure p while the structure is described by the field of displacement u. We begin by establishing the classical variational formulation of this problem, namely

$$(\mathcal{P}) \quad \begin{cases} \text{Find } \mathcal{U} = (p,u) \in W \text{ such that} \\ B(\mathcal{U},\mathcal{V}) = \displaystyle\int_{\Gamma_N} g\, v\, d\sigma, \quad \forall\, \mathcal{V} = (q,v) \in W, \end{cases}$$

where $W = \left\{(q,v) \in H^1(\Omega_f) \times [H^1(\Omega_s)]^2;\ v|_{\Gamma_D} = 0\right\}$. The bilinear form $B(\cdot,\cdot)$ is given by:

$$\begin{aligned} B(\mathcal{U},\mathcal{V}) &= c^2\int_{\Omega_f} \nabla p \nabla q\, dx + \int_{\Omega_f} p.q\, dx + \int_{\Omega_s} \sigma(u) : \varepsilon(v)\, dx + \rho_s \int_{\Omega_s} u.v\, dx \\ &+ c^2\rho_f \int_{\Sigma} q.u.n\, d\sigma - \frac{1}{\rho_s}\int_{\Sigma} p.v.n\, d\sigma - \lambda \int_{\Omega_s} \rho_s u.v\, dx \\ &- \lambda \int_{\Omega_f} p.q\, dx - \lambda c^2 \rho_f \int_{\Sigma} q.u.n\, d\sigma, \quad \forall\, \mathcal{U} = (p,u), \mathcal{V} = (q,v), \end{aligned} \tag{1}$$

with $\sigma(u) : \varepsilon(v) = \sum_{i,j=1}^{2} \sigma_{ij}(u)\varepsilon_{ij}(v)$ and $\lambda = 1 + \omega^2$.

This formulation has, on the computational side, the advantage of introducing only one unknown variable per node to describe the fluid. It has, on the other hand, the drawback of leading always to nonsymmetric matrices. This formulation, where numerical simulation by finite elements methods is described in [43], has been the subject of many investigations (c.f. [30, 25, 26, 44, 41, 45, 27, 42]). Also let us mention [40] in the field of the car industry. A way to obtain a symmetric variational formulation consists in describing the fluid by means of the field of displacement u^F (c.f. [31, 24]). Indeed in this case it is easy to establish a symmetric variational formulation of this problem. It is however sufficient to consider the fluid as a particular elastic medium whose law of behavior is $p = -\rho_f c^2 \operatorname{div} u^F$. We obtain then a symmetric formulation. This latter presents the disadvantage of requiring the discretization of the relation $\operatorname{rot} u^F = 0$ which constitutes a complex numerical problem by itself (C.f. [39]).

The bilinear form $B(\cdot,\cdot)$ satisfies an alternative of the Gårding inequality given by the following:

Lemma 1 *The bilinear form $B(\cdot,\cdot)$ defined on $W \times W$ by (1) satisfies the following inequality:*

$$B(\mathcal{U},\mathcal{U}) \geq c_0 \, \|\mathcal{U}\|_W^2 - \mu_0 \, (1+\omega^2) \, \|\mathcal{U}\|_V^2, \quad \forall \, \mathcal{U} \in W, \tag{2}$$

where c_0 and μ_0 are strictly positive constants independent of ω.

Based essentially on this lemma we prove that the problem $(\mathcal{P})$ admits a unique solution, this result is given in the following theorem:

Theorem 2 *For $g \in [L^2(\Gamma_N)]^2$, problem $(\mathcal{P})$ admits a unique solution. Furthermore, there exists a constant C independent of g such that:*

$$\|\mathcal{U}\|_W \leq C \, \|g\|_{\Gamma_N}. \tag{3}$$

The associated discrete variational formulation of the problem $(\mathcal{P})$ is to:

$$(\mathcal{P}_h) \quad \left\{ \begin{array}{l} \text{Find } \mathcal{U}_h = (p_h, u_h) \in W_h \text{ such that} \\ B(\mathcal{U}_h, \mathcal{V}_h) = \displaystyle\int_{\Gamma_N} g \, v_h \, d\sigma, \quad \forall \, \mathcal{V}_h = (q_h, v_h) \in W_h, \end{array} \right.$$

where $W_h = W_{1,h} \times W_{2,h}$ and $W_h \hookrightarrow W$, with

$$W_{1,h} = \left\{ v_h \in [C^0(\overline{\Omega}_s)]^2 \, ; \, v_{h|_T} \in [P_1(T)]^2 \; \forall T \subset \Omega_s ; \; v_{h|_{\Gamma_D}} = 0 \right\},$$

and

$$W_{2,h} = \left\{ q_h \in C^0(\overline{\Omega}_f) \, ; \, q_{h|_T} \in P_1(T) \; \forall T \subset \Omega_f \right\}.$$

We consider the following two properties (P_1) and (P_2) defined by:

$$(P_1) \quad \lim_{h \to 0} \left(\sup_{\mathcal{U} \in W} \inf_{\mathcal{U}_h \in W_h} \frac{\|\mathcal{U} - \mathcal{U}_h\|_W}{\|\mathcal{U}\|_W} \right) = 0$$

$$(P_2) \quad \left\{ \begin{array}{l} \text{For every } \mathcal{U}_h \in W_h \text{ we have} \\ \displaystyle\sup_{\|\mathcal{V}_h\|_W = 1} B(\mathcal{U}_h, \mathcal{V}_h) \geq \eta \, \|\mathcal{U}_h\|_W - \alpha \, \|\mathcal{U}_h\|_V , \end{array} \right.$$

where α and η are two constants such that $\alpha \geq 0$ and $\eta > 0$.

Based essentially on the previous properties we were able to prove the following theorem.

Theorem 3 *For h small enough and under the properties (P_1) and (P_2), there exists a strictly positive constant γ_h such that:*

$$\inf_{\mathcal{U}_h \in W_h} \sup_{\mathcal{V}_h \in W_h} \frac{B(\mathcal{U}_h, \mathcal{V}_h)}{\|\mathcal{U}_h\|_W \, \|\mathcal{V}_h\|_W} \geq \gamma_h. \tag{4}$$

We introduce the following spaces:

$$\overline{W}_h = \Big\{(q,v) \in W\,;\, q_{|_T} \in P_2(T)\ \forall T \subset \Omega_f,\ v_{|_T} \in [P_2(T)]^2\ \forall T \subset \Omega_s,\ v_{|_{\Gamma_D}} = 0\Big\},$$

$$\widehat{W}_h = \big\{(q,v) \in \overline{W}_h\,;\, q(a_i) = 0\ \forall a_i \text{ vertex of } T \subset \Omega_f,\ v(b_i) = 0\ \forall b_i \text{ vertex of } T \subset \Omega_s\big\}.$$

We enrich the space W_h by the space $\widehat{W}_h$ to obtain the space $\overline{W}_h$ such that:

$$\overline{W}_h = W_h \oplus \widehat{W}_h.$$

Now we introduce the intermediate problem associated with problem $(\mathcal{P}_h)$, definite on space $\overline{W}_h$, by:

$$(\overline{\mathcal{P}}_h)\quad \left\{\begin{array}{l} \text{Find } \overline{\mathcal{U}}_h = (\overline{p}_h, \overline{u}_h) \in \overline{W}_h \text{ such that} \\ B(\overline{\mathcal{U}}_h, \overline{\mathcal{V}}_h) = \displaystyle\int_{\Gamma_N} g\,\overline{v}_h\,d\sigma, \quad \forall\, \overline{\mathcal{V}}_h = (\overline{q}_h, \overline{v}_h) \in \overline{W}_h. \end{array}\right.$$

The global hierarchic a posteriori error estimator is the solution of problem $(\widehat{\mathcal{P}}_h)$ defined on the space $\widehat{W}_h$ by:

$$(\widehat{\mathcal{P}}_h)\quad \left\{\begin{array}{l} \text{Find } \widehat{\mathcal{E}}_h = (\widehat{p}_h, \widehat{u}_h) \in \widehat{W}_h \text{ such that} \\ B(\widehat{\mathcal{E}}_h, \widehat{\mathcal{V}}_h) = \displaystyle\int_{\Gamma_N} g\,\widehat{v}_h\,d\sigma - B(\mathcal{U}_h, \widehat{\mathcal{V}}) \quad \forall\, \widehat{\mathcal{V}}_h = (\widehat{q}_h, \widehat{v}_h) \in \widehat{W}_h. \end{array}\right.$$

By introducing the assumption of saturation:

(H_1) $\|\mathcal{U} - \overline{\mathcal{U}}_h\|_W \le \beta\, \|\mathcal{U} - \mathcal{U}_h\|_W, \quad \beta \in (0,1)$ is a constant independent of h,

and the assumption of existence of the constant of the strengthened Cauchy-Bunyakovski-Schwarz (C.B.S) inequality, independent of h:

$$(H_2)\quad \left\{\begin{array}{l} \exists \gamma < 1 \text{ independent of } h \text{ such that} \\ |(\mathcal{U}, \mathcal{V})_W| \le \gamma\, \|\mathcal{U}\|_W\, \|\mathcal{V}\|_W, \quad \forall\, \mathcal{U} \in W_h,\ \forall\, \mathcal{V} \in \widehat{W}_h, \end{array}\right.$$

we obtain the following result which gives the asymptotic equivalence of the constructed estimator, with the exact error.

Theorem 4 *If $\mathcal{U}$ is the solution of problem $(\mathcal{P})$, $\mathcal{U}_h$ the solution of problem $(\mathcal{P}_h)$ and $\widehat{\mathcal{E}}_h$ is the solution of problem $(\widehat{\mathcal{P}}_h)$, then under assumptions (H_1) and (H_2) the a posteriori estimator $\widehat{\mathcal{E}}_h$ satisfies the following inequalities:*

$$\|\mathcal{U} - \mathcal{U}_h\|_W \le \frac{\widehat{c}}{\overline{\gamma}_h\,(1-\beta)\,\sqrt{1-\gamma^2}}\, \|\widehat{\mathcal{E}}_h\|_W, \tag{5}$$

$$\|\widehat{\mathcal{E}}_h\|_W \le \frac{\overline{c}}{\widehat{\gamma}_h}\,(1+\beta)\, \|\mathcal{U} - \mathcal{U}_h\|_W, \tag{6}$$

where $\widehat{c}$ and $\overline{c}$ are, respectively, the constants of continuity of $B(\cdot,\cdot)$ on $\widehat{W}_h$ and $\overline{W}_h$.

The originality of this study is the introduction of a posteriori error estimates of hierarchical type for the coupled problem of elastoacoustics. The hierarchical a posteriori error estimator for the problem of elastoacoustics even if it is global one leads to the resolution of a linear system, which is generally well conditioned, and can be used for the adaptation of the grid. However a local version of this estimator is a interesting question.

3 Anisotropic error estimation for convection diffusion problem with dominant convection

In this section we consider a convection-diffusion problem given by the following elliptic boundary value problem in a bounded polygonal domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with Lipschitz boundary $\partial\Omega$.

$$(\mathcal{P}) \quad \left\{ \begin{array}{rcll} L_\varepsilon u := -\varepsilon\Delta u + \beta\nabla u + \sigma u & = & f & \text{in } \Omega, \\ u & = & 0 & \text{on } \Gamma_D, \\ \varepsilon\dfrac{\partial u}{\partial n} & = & g & \text{on } \Gamma_N, \end{array} \right.$$

where $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. We assume that the diffusion constant ε satisfies $0 < \varepsilon \ll 1$. Furthermore, we assume that:

(A_1) $\beta \in W^{1,\infty}(\Omega)^d, \sigma \in L^\infty(\Omega)$,

(A_2) $-\frac{1}{2}\nabla\beta + \sigma \geq 1$,

(A_3) $\Gamma_- := \{x \in \partial\Omega\,;\, \beta(x)n(x) < 0\} \subset \Gamma_D$.

The standard weak formulation of problem $(\mathcal{P})$ is to find $u \in H^1_D(\Omega)$ such that

$$B_\varepsilon(u, v) = (f, v) + (g, v)_{\Gamma_N}, \qquad \forall\, v \in H^1_D(\Omega), \tag{7}$$

where

$$B_\varepsilon(u, v) := \varepsilon \int_\Omega \nabla u\, \nabla v\, dx + \int_\Omega \beta\, \nabla u\, v\, dx + \int_\Omega \sigma\, u\, v dx.$$

Due to the Lax-Migram lemma, the problem (7) admits a unique solution u. Moreover, assumptions $(A_1) - (A_3)$ and integration by parts imply that

$$B_\varepsilon(v, v) \geq |||v|||^2_\varepsilon, \qquad \forall\, v \in H^1_D(\Omega) \tag{8}$$

and

$$\begin{array}{rcl} B_\varepsilon(v, w) & \leq & |||v|||_\varepsilon\, |||w|||_\varepsilon\, (1 + \|\sigma\|_\infty) \\ & + & |||v|||_\varepsilon\, \|w\|_{0,\Omega}\, \varepsilon^{-1/2}\, \|\beta\|_\infty, \quad \forall\, v, w \in H^1_D(\Omega). \end{array} \tag{9}$$

We denote by $\mathcal{F} = \{\mathcal{T}_h\}$ a family of partitions $\mathcal{T}_h$ of Ω into tetrahedra. We enumerate the vertices $P_0, \ldots, P_3$ of a given tetrahedron such that P_0P_1 is the longest edge, the triangle $P_0P_1P_2$ has the largest area of the two triangles adjacent to P_0P_1, and P_0P_2 is the shortest edge of $P_0P_1P_2$, c.f. Figure 2. We

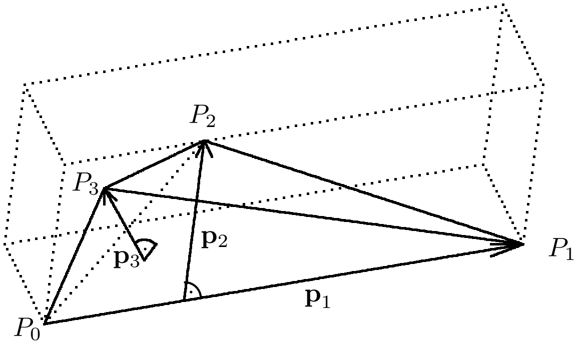


Figure 2: Notation for a tetrahedron T

define three vectors $\mathbf{p}_1, \dots, \mathbf{p}_3$ as follows: $\mathbf{p}_1 = \overrightarrow{P_0P_1}$; $\mathbf{p}_2$ is the vector in the plane $P_0P_1P_2$ which is perpendicular to P_0P_1 and which points to P_2; $\mathbf{p}_3$ is the vector which is perpendicular to the plane $P_0P_1P_2$ and which points to P_3. Set $h_{j,T} := |\mathbf{p}_j|$ and $h_{min,T} := \min\{h_{1,T}, h_{2,T}, h_{3,T}\} = h_{3,T}$.

Given any edge $E \in \mathcal{E}_h$, where $\mathcal{E}_h$ is the set of all faces in $\mathcal{T}_h$, we denote by $T_E \in \mathcal{T}_h$ the element which is adjacent to E and which has minimal $h_{min,T}$.

Set

$$h_{min,E} := h_{min,T_E}, \quad h_E := 3|T_E|/|E|.$$

The standard Galerkin method associated with the problem $(\mathcal{P})$ is:

$$\text{Find } u_h \in Y^0_{h,k} \text{ such that } B_\varepsilon(u_h, v_h) = (f, v_h) + (g, v_h)_{\Gamma_N}, \quad \forall\, v_h \in Y^0_{h,k}. \quad (10)$$

We recall the well known fact that the solution u_h of problem (10) may suffer from non-physical oscillations unless the element-wise numbers

$$P_T := \varepsilon^{-1}\, h_T\, \|\beta\|_\infty \quad \text{and} \quad \Gamma_T := \varepsilon^{-1}\, h_T^2\, \|\sigma\|_\infty,$$

are sufficiently small. As a remedy, one can use stabilized finite element methods, which consist of finding a solution u_h to the following problem:

$$\text{Find } u_h \in Y^0_{h,k} \text{ such that } B_\delta(u_h, v_h) = L_\delta(v_h), \quad \forall\, v_h \in Y^0_{h,k}, \quad (11)$$

where

$$\begin{aligned} B_\delta(u, v) &= B_\varepsilon(u, v) + \widetilde{B}_h(u, v), \\ L_\delta(v) &= (f, v) + (g, v)_{\Gamma_N} + \widetilde{L}_h(v). \end{aligned} \quad (12)$$

The bilinear form $\widetilde{B}_h(\cdot, \cdot)$ and the linear form $\widetilde{L}_h(\cdot)$ are the added stabilized terms corresponding to the considered stabilized method.

For the GLS method the corresponding added stabilized terms are given by:

$$\widetilde{B}_h(u,v) = \sum_{T\in\mathcal{T}_h} \delta_{h,T}\,(L_\varepsilon u, L_\varepsilon v)_T, \quad \forall\, u,v \in Y^0_{h,k},$$

and

$$\widetilde{L}_h(v) = \sum_{T\in\mathcal{T}_h} \delta_{h,T}\,(f, L_\varepsilon v)_T, \quad \forall\, u,v \in Y^0_{h,k}.$$

The non-negative numerical diffusion parameters $\delta_{h,T}$ have to satisfy the additional special properties $\delta_{h,T} = \dfrac{h_T}{2\|\beta\|_\infty}\xi(P_T)$, see [29], where P_T is the mesh-Peclet number and $\xi(\cdot)$ is a function defined by

$$\xi(\chi) = \begin{cases} \chi & \text{if} \quad \chi < 1, \\ 1 & \text{if} \quad \chi \geq 1. \end{cases}$$

In general we don't need to use extremely fine meshes when dealing with stabilized finite element formulations, i.e. $\varepsilon \ll h_T$, thus $\delta_{h,T} = \mathcal{O}(h_T)$.

We give here the definition of the matching function $m_1(v,\mathcal{T}_h)$. The matching function measures the correspondence (or the degree of alignment) of an anisotropic mesh $\mathcal{T}_h$ with an anisotropic function v. One advantage of this definition is that we can isolate the influence of the mesh alignment on various estimates by means of m_1, the worse the alignment, the larger m_1 becomes, and the less accurate the estimates are.

Definition 1 (Matching function m_1) *Let $v \in H^1(\Omega)$ be an arbitrary non-constant function, and $\mathcal{F}$ be a family of triangulations of Ω. Define the matching function $m_1(\cdot,\cdot) : H^1(\Omega) \times \mathcal{F} \to \mathbb{R}$ by*

$$m_1(v,\mathcal{T}_h) := \frac{\left(\displaystyle\sum_{T\in\mathcal{T}_h} h^{-2}_{min,T}\,\|C^t_T \nabla v\|^2_T\right)^{1/2}}{\|\nabla v\|}.$$

The following anisotropic estimates describe the main interpolation results

Lemma 5 *The following interpolation estimates hold for any $v \in H^1_0(\Omega)$*

$$\begin{aligned} \sum_{T\in\mathcal{T}_h} \mu_T^{-2}\,\|v - I_h v\|^2_{0,T} \quad &\leq \quad m_1(v,\mathcal{T}_h)^2\,|||v|||^2_{\varepsilon,\Omega}, \\ \varepsilon^{1/2} \sum_{T\in\mathcal{T}_h} \sum_{E\subset\partial T\setminus\Gamma_D} \mu_T^{-1}\,\frac{h_{E,T}}{h_{min,E}}\,\|v - I_h v\|^2_E \quad &\leq \quad m_1(v,\mathcal{T}_h)^2\,|||v|||^2_{\varepsilon,\Omega}. \end{aligned}$$

We introduce the definition of a residual error estimator that is based on an anisotropic mesh.

Definition 2 (Residual error estimator) *Define the local residual error estimator $\eta_{\varepsilon,R,T}(u_h)$ for a tetrahedron T by*

$$\begin{aligned}\eta_{\varepsilon,R,T}(u_h) \quad &:= \quad (\mu_T^2\,\|f_h+\varepsilon\Delta u_h-\beta_h\nabla u_h-\sigma_h u_h\|_{0,T}^2 \\ &+ \quad \frac{1}{2}\sum_{E\subset\partial T\cap\Omega}\varepsilon^{-1/2}\,\mu_T\,\frac{h_{min,E}}{h_{E,T}}\,\|[\varepsilon\partial_{n_E}u_h]_E\|_{0,E}^2 \\ &+ \quad \sum_{E\subset\partial T\cap\Gamma_N}\varepsilon^{-1/2}\mu_T\,\frac{h_{min,E}}{h_{E,T}}\,\|g_h-\varepsilon\partial_{n_E}u_h\|_{0,E}^2)^{1/2},\end{aligned}$$

where f_h, β_h, σ_h and g_h are arbitrary approximations of f, β, σ and g by piecewise polynomials of degree at most k with respect to $\mathcal{T}_h$ and with respect to the partition of Γ induced by $\mathcal{T}_h$, respectively.

The main result is summarize in the following theorem.

Theorem 6 *Let $u \in H_D^1(\Omega)$ be the solution of problem (7) and $u_h \in Y_{h,k}^0$ be the finite element solution of problem (11), corresponding to the GLS method. Then, the error is bounded globally from above by*

$$\begin{aligned}|||u-u_h|||_\varepsilon \le m_1(u-u_h,\mathcal{T}_h)\, &\Bigg\{ \{\sum_{T\in\mathcal{T}_h}\eta_{\varepsilon,R,T}{}^2(u_h)\}^{1/2} \\ +\Bigg\{ \sum_{T\in\mathcal{T}_h}\mu_T^2\,\|f-f_h\|_{0,T}^2 &+ \sum_{T\in\mathcal{T}_h}\mu_T^2\,\|\sigma u_h-\sigma_h u_h\|_{0,T}^2 \\ + \sum_{T\in\mathcal{T}_h}\mu_T^2\,\|\beta\nabla u_h-\beta_h\nabla u_h\|_{0,T}^2 &+ \sum_{E\in\mathcal{E}_{h,N}^2}\varepsilon^{-1/2}\,\mu_E\,\|g_h-g\|_{0,E}^2\Bigg\}^{1/2}\Bigg\}.\end{aligned}$$

The error is bounded locally from below by

$$\begin{aligned}\eta_{\varepsilon,R,T}(u_h) \le &\left(1+\|\sigma\|_{\infty,\omega_E}+\mu_T\,\varepsilon^{-1/2}\,\|\beta\|_{\infty,\omega_E}\right)|||u-u_h|||_{\varepsilon,\omega_E} \\ &+\mu_T\,\|f-f_h\|_{0,\omega_E}+\mu_T\,\|\sigma u_h-\sigma_h u_h\|_{0,\omega_E} \\ &+\mu_T\,\|\beta\nabla u_h-\beta_h\nabla u_h\|_{0,\omega_E}+\left(\sum_{E\subset\partial T\cap\Gamma_N}\varepsilon^{-1/2}\,\mu_T\,\|g_h-g\|_{0,\omega_E}^2\right)^{1/2},\end{aligned}$$

for all $T \in \mathcal{T}_h$.

Remark 1 *The upper error bound contains the matching function $m_1(u-u_h,\mathcal{T}_h)$. Since $u-u_h$ is not known, $m_1(u-u_h,\mathcal{T}_h)$ cannot be computed exactly. This can be remedied by using an approximation m_1^R by means of a recovered* *gradient* $\nabla^R u_h \approx \nabla u$ *:*

$$\begin{aligned} m_1(u-u_h,\mathcal{T}_h) &= \left(\sum_{T\in\mathcal{T}_h} h_{min,T}^{-2}\, \|C_T^t \nabla(u-u_h)\|_T^2\right)^{1/2} / \|\nabla(u-u_h)\| \\ &\approx \left(\sum_{T\in\mathcal{T}_h} h_{min,T}^{-2}\, \|C_T^t (\nabla^R u_h - \nabla u_h)\|_T^2\right)^{1/2} / \|\nabla^R u_h - \nabla u_h\| \\ &:= m_1^R(u_h,\mathcal{T}_h). \end{aligned}$$

For simplicity the recovered gradient $\nabla^R u_h$ *is chosen to be the linear Lagrange interpolate at the nodes of the mesh. The value at a node* a *is given by*

$$\nabla^R u_h(a) := \sum_{T:a\in\mathcal{N}_T} \frac{|T|}{|\omega_a|} \nabla u_{h|_T}, \qquad \textit{with} \quad \omega_a := \bigcup_{T:a\in\mathcal{N}_T} T,$$

for more details see e.g. [33].

As a model problem we consider a singularly perturbed convection-diffusion problem governed by the following partial differential equations:

$$\left\{ \begin{aligned} -\varepsilon\Delta u + \beta\nabla u + \sigma u &= f \quad \text{in } \Omega=[0,1]^2, \\ u &= 0 \quad \text{on } \Gamma=\partial\Omega. \end{aligned} \right.$$

The diffusion constant ε of the problem is chosen to be equal to 10^{-4}. In that case the solution of the problem exhibits a boundary layer behavior. For our consideration we take

$$\beta = \left[\frac{\sqrt{2}-x}{2}, \frac{\sqrt{2}-y}{2}\right] \quad \text{and} \quad \sigma = 1.$$

The right-hand f is chosen properly such that the exact solution is explicitly given by:

$$u(x,y) = x^2y^2\left(1-e^{-(1-x)/\sqrt{\varepsilon}}\right)\left(1-e^{-(1-y)/\sqrt{\varepsilon}}\right).$$

The numerical experiments have been performed using Freefem. For different meshes with different sizes we compare the energy norm of the exact error with the modified residual error estimator on uniform and anisotropic meshes. Anisotropic meshes are exponentially graded meshes adapted in a way to capture the boundary layer behavior of the solution. For this purpose the domain is discretized by a sequence of meshes, each one being the tensor product of two one dimensional Bakhvalov-like meshes [21] with 2^k intervals in $[0,1]$, $k=2,\ldots,8$.

The values of the matching functions $m_1(u-u_h,,\mathcal{T}_h)$ and $m_1^R(u_h,\mathcal{T}_h)$ are summarized in Tables 1 and 2 for uniform and exponentialy graded meshes respectively and for different mesh sizes. As it was expected the matching

http://www.freefem.org

Size	$m_1(u-u_h,\mathcal{T}_h)$	$m_1^R(u_h,\mathcal{T}_h)$
25	1.45549	1.44013
81	1.51033	1.50580
289	1.54716	1.54580
1089	1.56729	1.56615
4225	1.57513	1.57298
16641	1.57870	1.57262
66049	1.58050	1.56607

Size	$\|u-u_h\|$	$\eta_{\varepsilon,R}(u_h)$	$\|u-u_h\|/m_1\eta_\epsilon$
25	1.480947e-01	2.246904e+00	0.04528409
81	1.489938e-01	2.087869e+00	0.04724883
289	1.272738e-01	2.103825e+00	0.03910146
1089	7.722742e-02	1.676989e+00	0.02938262
4225	3.171651e-02	7.549109e-01	0.02667298
16641	9.583700e-03	1.561212e-01	0.03888390
66049	2.531673e-03	3.546737e-02	0.04516287

Table 1: Uniform mesh

function is small on uniform meshes compared to exponentially graded meshes, though it is still reasonably small on anisotropic meshes. The tables shows also the coincidence that exist between m_1 and m_1^R. Hence, one can use m_1^R instead of m_1 without loss of the efficiency of the estimator. In Tables 1 and 2 we give comparison between the energy error norm and the modified global estimator respectively for uniform and graded meshes. The values showing in those tables confirm the main result, namely the error bounds of Theorem 2.3. The efficiency of the modified residual estimator depends on how the mesh behaves with the anisotropy of the solution. This can be seen for instance from the values of the quotient of the error over the matching function multiplied by the estimator. On isotropic meshes this quotient remains constant for all different mesh sizes. Similar values can be seen also on anisotropic meshes with small sizes, but this quotient increases with large sizes, this is due to the fact that the matching function in that case increases as well.

4 Multilevel methods

4.1 Strenghtened Cauchy-Bunyakowski-Schwarz (C.B.S.) inequality for a three dimensional elasticity system

Over the past two decades many authors have presented multigrid methods in the context of multilevel finite element spaces for self adjoint coercive problems. The main tool in the analysis of such methods is the strengthened Cauchy-

Size	$m_1(u-u_h,\mathcal{T}_h)$	$m_1^R(u_h,\mathcal{T}_h)$
25	1.23077	1.30592
81	1.44595	1.30898
289	1.95260	1.33164
1089	4.01102	1.83591
4225	6.15498	3.40029
16641	8.09934	4.45940
66049	8.11395	3.58645

Size	$\|u-u_h\|$	$\eta_{\varepsilon,R}(u_h)$	$\|u-u_h\|/m_1\eta_\epsilon$
25	4.038918e-02	7.024847e-01	0.04671429
81	1.186747e-02	2.349178e-01	0.03493717
289	4.534163e-03	8.647979e-02	0.02685147
1089	2.338822e-03	4.823673e-02	0.01208827
4225	1.272625e-03	3.089789e-02	0.00669182
16641	5.065755e-04	1.363573e-02	0.00833085
66049	1.508516e-04	3.259980e-03	0.00570299

Table 2: Exponentially graded mesh

Bunyakowski-Schwarz (C.B.S.) inequality:

$$|(\mathbf{u},\mathbf{v})| \leq \gamma\sqrt{(\mathbf{u},\mathbf{u})}\sqrt{(\mathbf{v},\mathbf{v})}, \qquad \mathbf{u}\in U,\ \mathbf{v}\in V,\ U\cap V=\{0\},$$

where U and V are two linear subspaces, $(\cdot,\cdot)$ is the bilinear form corresponding to the variational formulation of the problem and $\gamma\in(0,1)$ depends only on the spaces U and V. The strengthened C.B.S. inequality has been used in two-level methods by Bank and Dupont [18], Axelsson [10], Axelsson and Gustafsson [12], Braess [22, 23], Maître and Musy [35]. The important role of the C.B.S. inequality constant in multilevel methods was reported by Axelsson and Vassilevski [15, 16] and by Eijkhout and Vassilevski [28]. The following theorem shows the main role of the constant(C.B.S.).

Theorem 7 *Let A be a symmetric, positive semidefinite 2×2 block matrix*

$$A=\begin{pmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}$$

where A_{11} is invertible and where the partitioning corresponds to U,V, which are the spaces of vectors with only nonzero first, respectively, second components, i.e. $\mathbf{u}\in U$ is of the form $\mathbf{u}=\binom{\mathbf{u}_1}{0}$ and $\mathbf{v}\in V$ of the form $\mathbf{v}=\binom{0}{\mathbf{v}_2}$. Assuming that the kernel of A , $N(A)$, is included in V, then the number γ given by

$$\gamma^2=\sup_{\substack{\mathbf{u}\in U\\ \mathbf{v}\in V\setminus N(A)}}\frac{(\mathbf{u}^TA\mathbf{v})^2}{(\mathbf{u}^TA\mathbf{u})\,(\mathbf{v}^TA\mathbf{v})}$$

satisfies

$$\gamma < 1 \quad \textit{and} \quad (1-\gamma^2)\mathbf{v}_2^t A_{22}\mathbf{v}_2 \leq \mathbf{v}_2^t S \mathbf{v}_2 \leq \mathbf{v}_2^t A_{22}\mathbf{v}_2, \quad \forall\, \mathbf{v}_2 \in V, \tag{13}$$

where $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ *is known as the Schur complement.*

Let Ω be a bounded open connected subset of $\mathbb{R}^3$ with a Lipschitz-continuous boundary Γ. Let Γ_0 be a measurable subset of Γ and let Γ_1 be its complement on Γ. We consider the following three-dimensional boundary value problem of linear elasticity :

$$\left\{ \begin{array}{rcll} \displaystyle\sum_{j=1}^{3} \frac{\partial \sigma_{ij}(\mathbf{u})}{\partial x_j} + f_i & = & 0 & \text{in } \Omega, \quad i = 1,2,3, \\ u_i & = & 0 & \text{on } \Gamma_0, \quad i = 1,2,3, \\ \displaystyle\sum_{j=1}^{3} \sigma_{ij}(\mathbf{u})\nu_j & = & g_i & \text{on } \Gamma_1, \quad i = 1,2,3. \end{array} \right.$$

$\nu = (\nu_1, \nu_2, \nu_3)$ is the external normal to Ω, with the classical constitutive law for isotropic material with Lamé moduli λ and μ:

$$\sigma_{ij}(\mathbf{u}) = \lambda\,(\sum_{k=1}^{3} \varepsilon_{kk}(\mathbf{u}))\,\delta_{ij} + 2\mu\varepsilon_{ij}(\mathbf{u})$$

where:

$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}) \quad i,j = 1,2,3.$$

We introduce the space:

$$V(\Omega) = \{\mathbf{v} \in [H^1(\Omega)]^3\,;\, \mathbf{v} = 0 \quad \text{on } \Gamma_0\}.$$

For $f \in [L^2(\Omega)]^3$, $g \in [L^2(\Gamma_0)]^3$, the variational formulation of the problem is given by:

$$(\mathcal{P}) \qquad \left\{ \begin{array}{l} \text{Find } \mathbf{u} \in V(\Omega) \text{ such that} \\ a(\mathbf{u},\mathbf{v}) = L(\mathbf{v}), \quad \forall\, \mathbf{v} \in V(\Omega) \end{array} \right.$$

where

$$L(\mathbf{v}) = \int_\Omega f\,\mathbf{v}\,dx + \int_{\Gamma_1} g\,\mathbf{v}\,d\sigma, \quad \forall\, \mathbf{v} \in V(\Omega)$$

and

$$a(\mathbf{u},\mathbf{v}) = \lambda \int_\Omega \operatorname{div}(\mathbf{u})\,\operatorname{div}(\mathbf{v})\,dx + 2\mu \sum_{i,j=1}^{3} \int_\Omega \varepsilon_{ij}(\mathbf{u})\,\varepsilon_{ij}(\mathbf{v})\,dx.$$

Let $\mathcal{T}_1$ be an initial discretization of the domain Ω. We assume that $\mathcal{T}_1$ is a set of non-degenerate arbitrary tetrahedral elements which cover Ω and have a mutually disjoint interior. The finer discretization $\mathcal{T}_2$ is constructed by refining

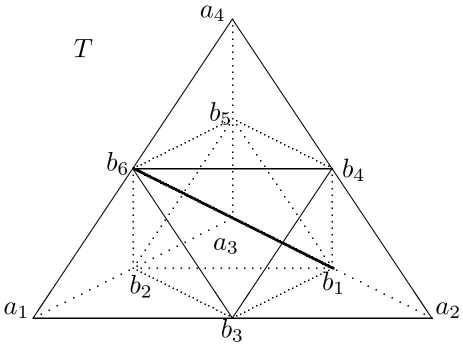


Figure 3: Connect midpoints of the tetrahedron

each tetrahedron $T \in \mathcal{T}_1$ one level using the regular refinement strategy as shown in Figure 3.

We define the space H by:

$$H = \{\mathbf{u} = (u_1, u_2, u_3) \in (C^0(\bar{\Omega}))^3\,;\, u_{i|T} \in P_1(T),\ i = 1, 2, 3,\ \ \forall T \in \mathcal{T}_2\}.$$

H is a finite-dimensional Hilbert space equipped with the scalar product $a(\cdot, \cdot)$ that can be written in the direct sum decomposition:

$$H = U \oplus V$$

where:

$$U = \{\mathbf{u} = (u_1, u_2, u_3) \in (C^0(\bar{\Omega}))^3\,;\, u_{i|T} \in P_1(T),\ i = 1, 2, 3,\ \ \forall T \in \mathcal{T}_1\}$$

and

$$V = \{\mathbf{u} \in H\,;\, u_i(s) = 0 \text{ for all vertices } s \text{ of each } T \in \mathcal{T}_1\}.$$

The constant of the strengthened C.B.S. inequality related to the pair (U, V) is defined by:

$$\gamma^2 = \sup_{\substack{\mathbf{u} \in U \\ \mathbf{v} \in V}} \frac{(a(\mathbf{u}, \mathbf{v}))^2}{a(\mathbf{u}, \mathbf{u})\, a(\mathbf{v}, \mathbf{v})}.$$

The estimate of the constant γ of the strengthened C.B.S. inequality is of fundamental practical importance, being necessary for computing the optimal parameters of the algebraic multilevel preconditioners as well as in the evaluation of the efficiency index of some hierarchical a posteriori error estimators. Recently Margenov [36] gave estimates of the 2D elasticity problem on a triangular mesh, and proved for a uniform mesh of right isosceles triangles, that the constant γ^2 is bounded above by 3/4 uniformly on the mesh. More recently B. Achchab and J.F.Maître [9, 1, 2] (see also Axelsson [11]) proved that this result remains true for every triangular mesh. A first investigation of the

3D problem was given by [32], where it was reported that numerical experiments on a particular triangulation, namely the refinement of a standard tetrahedron, show that the constant γ^2 is bounded above by 9/10.

In [34] we prove for an arbitrary tetrahedral mesh that the constant γ^2 associated with the form $a(\mathbf{u}, \mathbf{v})$ remains bounded above by 9/10. Moreover, we prove that the constant γ plays a crucial role in some a posteriori error estimators such as that of [8, 1, 2, 17, 20].

4.2 A preconditioning method for systems of linear partial differential equations

In many problems in mathematical modeling in natural sciences, engineering and in other areas as well where second order boundary value problems must be solved numerically, large scale linear systems arise which furthermore, frequently must be solved a number of times for each modeling case.

Often, the arising systems are severely ill-conditioned due to some problem parameters taking near certain limit values. Examples of such parameters are ratio of coefficient jumps, anisotropy, aspect ratio of the mesh and domain geometry, Poisson ratio for nearly incompressible materials, etc. Furthermore, the condition number may increase rapidly when the discretization mesh is refined (due to both a smaller mesh parameter and possible irregularity of the mesh elements). Therefore, instead, in finding a good solution method one should preferably search for efficient preconditioners for the, parameter free, conjugate gradient iterative solution method.

The method presented in [34, 6] is a block matrix approximate factorization preconditioning of the algebraic multilevel iteration, AMLI type. It is based on two or multilevel finite element meshes and can handle arbitrary coefficient jumps on the coarsest mesh used and also ratio of anisotropy, using newly developed finite element based preconditioners for the block corresponding to the added nodes. The condition number is bounded for any ratio of coefficient jumps and anisotropy.

Algebraic multilevel preconditioners were first presented in [15, 16] and are multilevel extensions of the two-level methods in [19] and [12]. Here block matrix approximate factorizations were considered and it was shown that by recursively extending the two level method using certain matrix polynomial approximations of the arising Schur complement matrices, one can derive a preconditioning with a condition number which is bounded independently on the number of levels and on jumps in the coefficients, assuming the coarsest mesh used had no jumps inside any element. As, in practice, one can use a coarsest mesh which is still quite fine, a significant number of jumps in coefficients, i.e. different materials in the physical model can be allowed. Similarly, preconditioners in additive form, i.e., using block diagonal preconditioners, but with stabilization at certain levels (see [11]) were developed with the same properties.

In the above methods the block matrix corresponding to the on each level added degrees of freedom gets increasingly ill-conditioned with increasing degree of anisotropy. Until recently, no efficient generally applicable method to handle this problem has been given. In [38], a preconditioner to this matrix in multiplicative form and in [14] an element by element preconditioner in additive form were suggested. The first method considered either x- or y-dominated anisotropy while the latter considered the general case with arbitrary coefficients in the differential operator. It was shown that the preconditioner is spectrally equivalent to the given matrix with bounds which hold uniformly in the number of levels and in the coefficients of the operator. A preliminary idea of the present work has been considered in [13], in particular for a scalar case.

In [6] we consider possible improvements of these methods. In particular it is shown that for the considered new element by element preconditioners improvements in the condition number can be achieved. The analysis of the computational complexity related to the constructed preconditioners is also considered for different model problems.

Consider the elliptic problem

$$\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) = f \qquad \text{in } \Omega,$$

where Ω is a polyhedral domain, with proper boundary conditions on $\partial\Omega$. For systems of *pde*'s of dimension d, u is a d-dimensional vector and a_{ij} are $d \times d$ matrices. Its variational formulation is: find $u \in H_g^1(\Omega)$ such that

$$a(u,v) = \int_\Omega f\,v\,dx \quad \text{for all } v \in H_0^1(\Omega),$$

where

$$a(u,v) = \int_\Omega \sum_{i,j} a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}, \tag{14}$$

and where the function spaces $H_g^1(\Omega)$ and $H_0^1(\Omega)$ incorporate the Dirichlet portion of the boundary conditions. Further, the matrix (a_{ij}) is assumed to be symmetric and positive definite. The domain of definition is partitioned in finite elements, such as triangles ($d = 2$), tetrahedrons or prisms ($d = 3$) and on each element we use piecewise linear finite element basis functions. Each finite element is partitioned in 2^d elements of equal volume and the node set is partitioned in two sets, the old (coarse mesh) and the new (added) ones. The finite element matrix is partitioned correspondingly in 2×2 blocks

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{array}{l} \text{(added node points)} \\ \text{(coarse mesh node points),} \end{array}$$

which is the two level matrix.

If we use basis functions for the arising small elements in both the old and new node points, A takes the nodal basis function form while if we keep the basis functions for the old node set corresponding to the whole (unrefined) elements, it takes the form of a hierarchical basis function matrix

$$\widehat{A} = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{pmatrix}.$$

To be specific we illustrate this method for triangular meshes. Two-level methods arise in mesh refinement methods. Given a 'coarse' triangle, it is subdivided in four congruent triangles by joining the mid-edge nodes. In the so arising six node-points we can use either nodal basis functions for the small triangles or hierarchical basis functions, where we keep the previous basis functions in the vertex node points of the coarse triangle and add new basis functions in the mid-edge nodes. The latter can be piecewise linear, with support only on the adjacent three triangles (see Figure 4), called the h-version.

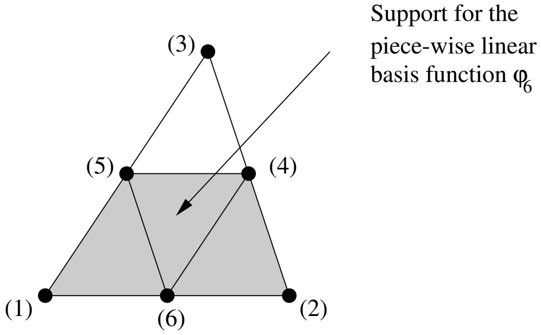


Figure 4: Support for the piece-wise linear basis function

Alternatively, we can use piecewise quadratic basis functions in the added node points with support on the whole triangle, called the p-version ($p = 2$).

The following relation holds between the corresponding nodal A and hierarchical ($\widehat{A} = \widehat{A}_h$) or ($\widehat{A} = \widehat{A}_p$) basis function matrices. Recall that, by assumption, A and $\widehat{A}$ are symmetric and positive definite.

Let $J = \begin{bmatrix} I_1 & J_{12} \\ 0 & I_2 \end{bmatrix}$ where I_1, I_2 are identity matrices corresponding to the coarse vertex node set and the mid-edge node set, respectively, and J_{12} is the interpolation matrix, with two non-zero entries ($= \frac{1}{2}$) in each row, corresponding to linear interpolation on the mid-edge node for each edge. The following relation holds then between function (vectors) represented in the hierarchical and the nodal basis,

$$v_{SB} = Jv_{HB}.$$

From this relation follows

$$\widehat{A} = J^T AJ,$$

and an elementary computation, using this shows the next relations between the corresponding matrix blocks,

$$\widehat{A}_{11} = A_{11}, \qquad \widehat{A}_{12} = A_{12} + A_{11}J_{12},$$
$$\widehat{A}_{21} = A_{21} + J_{12}^T A_{11}, \qquad \widehat{A}_{22} = A_{22} + J_{12}^T A_{12} + A_{21}J_{12} + J_{12}^T A_{11} J_{12}.$$

Further, $\widehat{A}_{22} = A_{2h}$, i.e., the nodal basis function matrix for the coarse (unrefined) mesh and $\widehat{S} = S$, where $\widehat{S} = \widehat{A}_{22} - \widehat{A}_{21}A_{11}^{-1}\widehat{A}_{12}$ and $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

Moreover, the following spectral relations hold:

$$(1-\gamma)\begin{pmatrix} A_{11} & 0 \\ 0 & \widehat{A}_{22} \end{pmatrix} \leq \begin{pmatrix} A_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{pmatrix} \leq (1+\gamma)\begin{pmatrix} A_{11} & 0 \\ 0 & \widehat{A}_{22} \end{pmatrix},$$

$$(1-\gamma^2)\widehat{A}_{22} \leq S \leq \widehat{A}_{22},$$

where $\gamma = \{\rho(\widehat{A}_{22}^{-1/2}A_{21}A_{11}^{-1}A_{12}\widehat{A}_{22}^{-1/2})\}^{1/2}$ and all inequalities are sharp. Further, it is known that the above block diagonal matrix is an optimal preconditioner, i.e. minimizes the spectral condition number, among all block-diagonal preconditioners.

Here $\gamma,\ 0 \leq \gamma < 1$, is identical to the constant in the strengthened C.B.S.-inequality

$$u^TAv \leq \gamma\{u^TAuv^TAv\}^{1/2},$$

which holds for all orthogonal vectors $u, v,\ u = (0,0,0,\alpha_1,\alpha_2,\alpha_3),\ v = (\beta_1,\beta_2,\beta_3,0,0,0)$. We let $\gamma_1,\ \gamma_2$ denote the constants for the h-version (i.e. for $p = 1$) and the p-version (i.e. for $p = 2$) of hierarchical basis functions, respectively. The following relation between γ_1 and γ_2 holds.

Theorem 8 *[37],[11] For any finite element triangular mesh, where each element has been refined into congruent elements, it holds*

$$\gamma_2^2 = \frac{4}{3}\gamma_1^2, \tag{15}$$

where $\gamma_1,\ \gamma_2$ are the C.B.S. constants for the piecewise linear and piecewise quadratic finite elements, respectively.

Corollary 9

$$\gamma_1^2 < \frac{3}{4}.$$

In [34, 6] we consider a multiplicative (=factorized) preconditioner. We use here the notations A_{2h} and A_h for the stiffness matrices corresponding to two consecutive levels. This preconditioner takes the form ([16])

$$M_h = \begin{pmatrix} B_{11} & 0 \\ \widetilde{A}_{21} & S_B \end{pmatrix}\begin{pmatrix} I_1 & B_{11}^{-1}\widetilde{A}_{12} \\ 0 & I_2 \end{pmatrix},$$

where

$$\begin{aligned} \widetilde{A}_{12} &= A_{12} + (A_{11} - B_{11})J_{12}, \\ \widetilde{A}_{21} &= A_{21} + J_{12}^T(A_{11} - B_{11}). \end{aligned} \tag{16}$$

Here J_{12} is the interpolation matrix which transforms the components of the current coarse vector to the new components of the vector on the next finer level. The reason for perturbing the off-diagonal block matrices as done in (16) is that in this way

$$\widehat{M_h} = J^T M_h J, \tag{17}$$

and $\widehat{M_h}$ takes the form

$$\widehat{M_h} = \begin{pmatrix} B_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & S_B + \widehat{A}_{21} B_{11}^{-1} \widehat{A}_{12} \end{pmatrix},$$

which follows from an elementary computation. Here $\widehat{A}_{12} = A_{12} + A_{11} J_{12}$ is the off-diagonal block in the hierarchical basis function matrix

$$\widehat{A}_h = \begin{pmatrix} A_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & A_{2h} \end{pmatrix}.$$

Hence $\widehat{M_h}$ can be considered as a preconditioner to $\widehat{A}_h$ and the extreme eigenvalues of $M_h^{-1} A_h$ equal those of $\widehat{M_h}^{-1} \widehat{A}_h$, since

$$\sup_v \frac{v^T A_h v}{v^T M_h v} = \sup_{\widehat{v}} \frac{\widehat{v}^T \widehat{A}_h \widehat{v}}{\widehat{v}^T \widehat{M_h} \widehat{v}}, \qquad \inf_v \frac{v^T A_h v}{v^T M_h v} = \inf_{\widehat{v}} \frac{\widehat{v}^T \widehat{A}_h \widehat{v}}{\widehat{v}^T \widehat{M_h} \widehat{v}}.$$

Since the off-diagonal blocks in $\widehat{M_h}$ equal those in $\widehat{A}_h$, the estimate of the extreme eigenvalues of $\widehat{M_h}^{-1} \widehat{A}_h$ can be readily done. As shown in [16], if

$$v_1^T A_{11} v_1 \le v_1^T B_{11} v_1 \le (1+b) v_1^T A_{11} v_1, \quad \text{for all } v_1 \in \mathbb{R}^{n_2 - n_1}$$

and

$$v_2^T A_{2h} v_2 \le v_2^T S_B v_2 \le (1+d) v_2^T A_{2h} v_2, \quad \text{for all } v_2 \in \mathbb{R}^{n_1},$$

then

$$\operatorname{cond}\left(M_h^{-1} A_h\right) \le \frac{1+b+d}{1-\gamma^2}. \tag{18}$$

It follows from Corollary 1 that for piecewise linear functions on a triangular mesh it holds $\gamma^2 < 3/4$. For tetrahedrons it holds $\gamma^2 < 9/10$, see [3]. The multiplicative method can be extended recursively replacing S_B with a matrix polynomial approximation

$$S_B^{-1} = [I - P_\nu(M_{2h}^{-1} A_{2h})] A_{2h}^{-1},$$

where $P_\nu(0) = 1$ and P_ν is small on the interval of the eigenvalues of $M_h^{-1} A_h$, where M_{2h} is the preconditioner to A_{2h}. The best approximation is by a shifted and scaled Chebyshev polynomial, see [16]. In this way, the condition number can be stabilized, i.e., bounded by a number which does not depend on the number of levels. The construction is readily extended to multilevels. The polynomial doesn't have to be the same on each level.

As can be seen from the condition number estimates in (18), it is important to control the conditioning of A_{11}. The major task of [6] deals with some recent results in construction of preconditioners B_{11} to A_{11}, with condition number bounds which hold uniformly in problem and mesh parameters. We will consider a multiplicative preconditioner which is applicable to more general systems of partial differential equations.

We consider the matrix corresponding to the mid-edge points in a $2D$-triangulation of the given domain where each macroelement contains four congruent triangles, as previously. Here each element matrix is a 3×3 block matrix where each block has the same order as the differential equation, i.e., 2. It is readily seen that the matrix takes the structure

$$K = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix}, \quad \text{where} \quad K_{11} = K_{22} = K_{33} \quad \text{and} \quad K_{ij} = K_{ji}^T.$$

Further, K is positive definite. K is identical to the finite element matrix for the above element where we have Dirichlet boundary conditions at the vertex nodes and Neumann boundary conditions at the mid-edge node points.

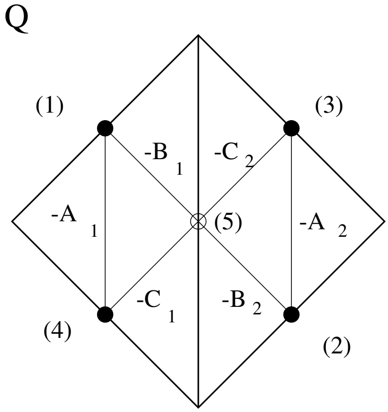


Figure 5: Weakest couplings in the union of two adjacent elements

$$A_{11} = \begin{pmatrix} D_{11} & F_{11} \\ {F_{11}}^T & E_{11} \end{pmatrix} = \begin{pmatrix} D_{11} & 0 \\ {F_{11}}^T & S_{11} \end{pmatrix} \begin{pmatrix} I & {D_{11}}^{-1}F_{11} \\ 0 & I \end{pmatrix}, \tag{19}$$

Consider the factorization (19) where the partitioning of the nodes corresponds to superelements Q consisting of two adjacent triangles chosen such that the coupling along the edges parallel to the interface edge is the weakest in a sense to be defined below, see Figure 5. The union of such superelements is denoted by $\mathcal{T}_h^Q$. The preconditioner B_{11} is defined by block approximation of the related Schur complement, i.e.,

$$B_{11} = \begin{pmatrix} D_{11} & 0 \\ {F_{11}}^T & \widehat{S}_{11} \end{pmatrix} \begin{pmatrix} I & {D_{11}}^{-1}F_{11} \\ 0 & I \end{pmatrix}, \tag{20}$$

where

$$S_{11} = S_{11:I} + \sum_{Q\in\mathcal{T}_h} S_{11:Q}, \qquad \widehat{S}_{11} = S_{11:I} + \sum_{Q\in\mathcal{T}_h} \widehat{S}_{11:Q}, \tag{21}$$

and

$$S_{11:Q} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \qquad \widehat{S}_{11:Q} = \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix}.$$

Here, the blocks S_{ij}, $i, j = 1, 2$, are 4×4 matrices where the partitioning follows the local numbering of the nodes from Figure 5. The first additive term in (21) can be written in the form

$$S_{11:I} = \sum_{E\in\{\mathcal{T}_h - \mathcal{T}_h^Q\}} A_{11:E}, \qquad \mathcal{T}_h^Q = \bigcup_{Q\in\mathcal{T}_h} Q, \tag{22}$$

i.e., $S_{11:I}$ is the part of A_{11} corresponding to the interface between the triangles from $\mathcal{T}_0$, and therefore is unchanged by a static condensation.

Following the local analysis scheme we need the superelement matrices $A_{11:Q}$ and $S_{11:Q}$. The corresponding matrix, ordered as in Figure 5, takes the following structure:

$$A_{11:Q} = \begin{pmatrix} K_{11} & 0 & 0 & K_{14} & K_{15} \\ 0 & K_{22} & K_{23} & 0 & K_{25} \\ 0 & K_{32} & K_{33} & 0 & K_{35} \\ K_{41} & 0 & 0 & K_{44} & K_{45} \\ K_{51} & K_{52} & K_{53} & K_{54} & K_{55} \end{pmatrix},$$

where $K_{ij} = K_{ji}^T$, $K_{23} = K_{14} \equiv A$, $K_{25} = K_{15} \equiv B$, $K_{45} = K_{35} \equiv C$, $K_{11} = K_{22} = K_{33} = K_{44} \equiv D$ and $K_{55} = 2D$.

Here the matrix $A_{11:Q}$ and, in particular, D is symmetric and positive definite. Hence the assembled matrix has the structure:

$$A_{11:Q} = \begin{pmatrix} D & 0 & 0 & A & B \\ 0 & D & A & 0 & B \\ 0 & A^T & D & 0 & C \\ A^T & 0 & 0 & D & C \\ B^T & B^T & C^T & C^T & 2D \end{pmatrix}.$$

Here we will eliminate by static condensation the interior node-point to form a 4×4 block matrix and use its block diagonal part as preconditioner.

Since the bilinear form of the problem (14) is symmetric it follows that all matrices A, B, C are symmetric. Further, it holds

$$-A - B - C = D. \tag{23}$$

"Weakest" coupling

The choice of coupling is such that the matrix A is the weakest in the following sense:

$$\text{(i)} \left\{ \begin{array}{l} -\tau(A+B) \le -(B+C) \\ -\tau(A+C) \le -(B+C) \end{array} \right. \quad \text{for some positive } \tau.$$

(ii) Assume also that $-A-B>0, \quad -A-C>0.$

Further, from (i) and (ii), it can be seen that $-B-C>0$.

As it turns out, the method and results we shall present are equally applicable for the case of a pair of triangles having jumps in the coefficients but otherwise with the same matrices, i.e., where the relations

$$K_{ij}^{(2)} = \nu K_{ij}^{(1)}, \qquad i,j=1,2,3, \quad \text{where} \quad 0<\nu\leq 1, \tag{24}$$

hold. For this case the corresponding stiffness matrix takes the form

$$A_{11:Q} = \begin{pmatrix} D_1 & 0 & 0 & A_1 & B_1 \\ 0 & D_2 & A_2 & 0 & B_2 \\ 0 & A_2^T & D_2 & 0 & C_2 \\ A_1^T & 0 & 0 & D_1 & C_1 \\ B_1^T & B_2^T & C_2^T & C_1^T & D_1+D_2 \end{pmatrix},$$

where $D_2=\nu D_1$, $A_2=\nu A_1$, $B_2=\nu B_1$, $C_2=\nu C_1$. Let $D_1=D$, $A_1=A$, $B_1=B$, $C_1=C$.

Next we eliminate by static condensation the couplings to the interior node point to form a 4×4 block matrix:

$$S_{11:Q} = \begin{pmatrix} D-\alpha E & -\beta E & -\beta F & A-\alpha F \\ -\beta E & \nu(D-\beta E) & \nu(A-\beta F) & -\beta F \\ -\beta F^T & \nu(A-\beta F)^T & \nu(D-\beta G) & -\beta G \\ (A-\alpha F)^T & -\beta F^T & -\beta G & D-\alpha G \end{pmatrix}, \tag{25}$$

where $E=BD^{-1}B^T$, $F=BD^{-1}C^T$, $G=CD^{-1}C^T$ and $\alpha=\dfrac{1}{1+\nu}$, $\beta=\dfrac{\nu}{1+\nu}$. Note that $\alpha+\beta=1$. We precondition $S_{11:Q}$ with $\widehat{S}_{11:Q}$.

Again, we recall that (apart from a scalar factor) this is the optimal preconditioner among all block diagonal preconditioners and to find the condition number $\kappa\left(\widehat{S}_{11:Q}^{-1}S_{11:Q}\right)$ we can compute $\gamma^2=\rho(S_{11}^{-1}S_{12}S_{22}^{-1}S_{21})$ and use the fact that $\kappa\left(\widehat{S}_{11:Q}^{-1}S_{11:Q}\right)=\dfrac{1+\gamma}{1-\gamma}$. The main result is given in the following theorem.

Theorem 10 *The multiplicative preconditioner of A_{11} has an optimal order convergence rate with a relative condition number uniformly bounded by*

$$\kappa\left({B_{11}}^{-1}A_{11}\right)\leq\frac{1+\rho_0}{1-\rho_0}, \tag{26}$$

where $\rho_0=max(\dfrac{1}{1+\tau},\rho^{1/2}(\widetilde{A}^T\widetilde{A}))$. Furthermore, the result holds uniformly in problem parameters and shape of triangles.

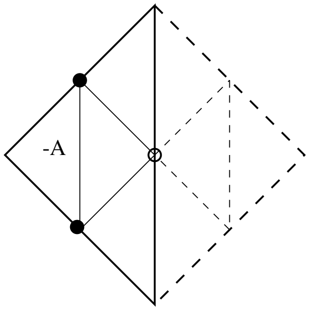


Figure 6: Static condensation in a single element

Remark 2 *The bound in Theorem 10 does not depend on the jump ratio ν. Hence it holds also for $\nu \to 0$ showing the same bound for a single triangle, where the node opposite to the weakest coupling has been eliminated, see figure 6. This property can be of importance for the fictitious domain method, among others.*

Remark 3 *In the case of elasticity problem on a uniform mesh of isosceles triangles, it can be shown that the parameter τ depends on the Poisson ratio $\tilde{\nu}$, $(\tilde{\nu} < 1/2)$ as follows*

$$\tau \leq 2\frac{\sqrt{8(4\tilde{\nu}^2 - 6\tilde{\nu} + 2) + 1} - 1}{\sqrt{8(4\tilde{\nu}^2 - 6\tilde{\nu} + 2) + 1} + 1}.$$

In this case, the behavior of the condition number with respect to the Poisson ratio $(\tilde{\nu})$ is shown in Figure 7.

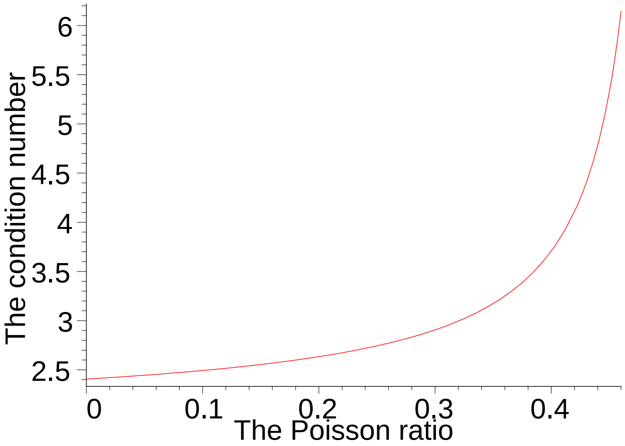


Figure 7: The condition number of elasticity on uniform mesh

Remark 4 *For the scalar partial differential equations the second multiplicative preconditioner gives a bigger bound than the first one and the condition number is bounded up by* 3 *where the coefficients* $-A = a$*,* $-B = b$ *and* $-C = c$ *are weakly connected in the sense of the relation* $|a| \leq b \leq c$*.*

Acknowledgments

This work has been supported partially by CNCPRST, AI 34PRO/00 and LERMA, EMI. The authors would like to thank Pr. A. Agouzal and Dr. B. Polman.

References

[1] B. Achchab, *Estimations d'erreur a posteriori, éléments finis mixtes hiérarchique, méthodes de stabilisation et méthods multiniveaux*, Thèse de Doctorat, Université de Lyon 1, Novembre 1995.

[2] B. Achchab, *Estimations d'erreur a posteriori, éléments finis mixtes hybrides et décomposition de domaines. Applications aux systèmes d'équations aux dérivées partielles*, Thèse d'état, Université Ibn Tofail, Kénitra, Octobre 2000.

[3] B. Achchab, O. Axelsson, L. Laayouni and A. Souissi, *Strengthened Cauchy-Bunyakowski-Schwarz inequality for a three dimensional elasticity system*, Numer. Linear Algebra Appl. Vol. 8(3), pp. 191-205, 2001.

[4] B. Achchab, A. Agouzal, L. Laayouni et A. Souissi, *Estimations d'erreurs en éléments finis pour un problème d'interaction fluide-structure*, Math-Recherche et Applications, Vol 2, T. 1, pp 91-107, 2000.

[5] B. Achchab, B. Polman, L. Laayouni, and A. Souissi, *Anisotropic a posteriori error estimations in convection-diffusion with dominant convection*, to appear in IJASC.

[6] B. Achchab, O. Axelsson, L. Laayouni, and A. Souissi, *A preconditioning method for systems of partial differential equations*, submitted to NLA, 2001.

[7] B. Achchab, A. Agouzal, L. Laayouni, and A. Souissi, *Error analysis for problems satisfying a discrete Gårding inequality in Finite Element Methods*, submitted to ZAMM, 2001.

[8] B. Achchab, A. Agouzal, J. Baranger, J.F. Maître, *Estimations d'erreur a posteriori en éléments finis hiérarchique. Application aux éléments finis mixtes*, Numer Math., 80:159–179, 1998.

[9] B. Achchab, J.F. Maître, *Estimate of the constant in two strengthened C.B.S. Inequalities for F.E.M Systems of 2D Elasticity. Application to*

Multilevel Methods and a Posteriori Error Estimators, Num. Lin. Alg. Appl., 3 (2):147–159, 1996.

[10] O. Axelsson, *On multigrid methods of the two-level type*, In Multigrid methods Proceedings, Köln-Porz, W. Hackbusch and U. Trottenberg, eds, Lecture Notes in Math, 960, pp. 352-367, 1982.

[11] O. Axelsson, *Stabilization of algebraic multilevel iteration methods; additive methods*, Numerical Algorithms, 21:23–47, 1999.

[12] O. Axelsson and I. Gustafsson, *Preconditioning and two-level multigrid methods of arbitrary degree of approximation*, Math. Comp, 40, pp. 219-242, 1983.

[13] O. Axelsson and S. Margenov, *An optimal order multilevel preconditioner with respect to problem and discretization parameters*, Minev, Wong and Lin eds., Advances in Computations, Theory and Practice, Nova Science Publishers, Huntington, New York, Vol. 7, pp. 2-18, 2001.

[14] O. Axelsson and A. Padiy, *On the additive version of the algebraic multilevel iteration method for anisotropic elliptic problems*, SIAM J. Sci. Comput., 20, pp. 1807–1830, 1999.

[15] O. Axelsson and P. Vassilevski *Algebraic multilevel preconditioning methods I*, Numer. Math., 56:157–177, 1989.

[16] O. Axelsson O and P. Vassilevski *Algebraic multilevel preconditioning methods II*, SIAM J. Numer. Anal., 67:1569–1590, 1990.

[17] R.E. Bank and A. Weiser *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44:283–301, 1985.

[18] R. Bank and T. Dupont, *Analysis of a two-level scheme for solving finite element equations*, Tech. Report CNA-159, Center for Numerical Analysis, University of Texas, Austin, TX, 1980.

[19] R. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp., 36, pp. 35-51, 1981.

[20] R.E. Bank and R.K. Smith, *A posteriori error estimates based on hierarchical bases*, SIAM J. Numer. Anal., 30:921–935, 1993.

[21] N.S. Bakhvalov, *The optimization of the methods of solving boundary value problems with a boundary layer*, Zh. Vychist. Mat. Mat. Fiz., 9, pp. 841-859, 1969.

[22] D. Braess, *The condition number of a multigrid method for solving the Poisson equation*, Numer. Math., 37:387–404, 1981.

[23] D. Braess, *The convergence rate of a multigrid method with Gauss-Seidel relaxation for the Poisson equation*, in Multigrid Methods Proceedings, Köln-Porz, (1981), W. Hackbusch and U. Trottenberg, eds, Lecture Notes in Math., 960:368–387, 1982.

[24] H.C. Chen and R.L. Taylor, *Vibration analysis of fluid-solid systems using a finite element displacement formulation*, Int. J. Num. Meth-Eng. 29, pp. 683-698, 1990.

[25] A. Craggs, *The transient response of a coupled plate-acoustic system using plate acoustic finite elements*, J. Sound Vib. 15, pp. 509-528, 1971.

[26] A. Craggs and G. Stead, *Sound transmission between enclosures. A study using plate and acoustic finite elements*, Acustica 35, pp. 89-98, 1976.

[27] E.H. Dowell, G.F. Gorman and D.A. Smith, *Acoustoelasticity general theory, acoustic natural modes and forced response to sinusoidal excitation, including comparison with experiments*, J. of Sound and Vib. 52, pp. 519-542, 1977.

[28] V. Eijkhout and P. Vassilevski *The role of the strenghtened Cauchy-Bunyakowski-Schwarz inequality in multilevel methods*, SIAM Review, 33:405–419, 1991.

[29] L.P. Franca, S.L. Frey and T.J.R. Hughes, *Stabilized finite element methods I: Application to the advective-diffusive model*, Comput. Math. Appl. Mech. Engrg., 95, pp 253-271, 1992.

[30] G.M.L. Gladwell, *A variational formulation of damped acousto-structured vibration problems*, J. of Sound and Vib. 4, pp. 172-186, 1966.

[31] M.A. Hamdi, Y. Ousset and G. Verchery, *A displacement method for the analysis of vibrations of coupled fluid-structure systems* Int. J. Num. Meth. Eng. 13 (1), (1978). (1991).

[32] M. Jung and F.F. Maitre, *Some remarks on the constant in the strengthened C.B.S. inequality: Estimate for hierarchical finite element discretization of elasticity problems*, Numerical Methods for Partial Differential Equations, 15 (4):469–488, 1999.

[33] G. Kunert, *Error estimation for anisotropic tetrahedral and triangular finite element meshes*, Numer. Math., 86(3):283-303, 2000.

[34] L. Laayouni, *Adaptativity in finite elements and multilevel methods*, PhD Thesis, Mohammad V University, Rabat, Morocco, 2001.

[35] J.F. Maitre and F. Musy, *The contraction number of a class of two-level methods; an exact evaluation for some finite element subspaces and model problems*, in Multigrid methods Proceedings, Köln-Porz, 1981, W. Hackbusch and U. Trottenberg, eds, Lecture Notes in Math., 960:535–544, 1982.

[36] S.D. Margenov, *Upper bound of the constant in the strengthened C.B.S. inequality for FEM 2D elasticity equations*, Num. Lin. Alg. Appl., 1 (1):65–74, 1994.

[37] J. F. Maitre and F. Musy, *The contraction number of a class of two-level methods; an exact evaluation for some finite element subspaces and model problems*, Lect. Notes Math., 960, pp. 535–544, 1982.

[38] S. Margenov and P.S. Vassilevski, *Algebraic multilevel preconditioning of anisotropic elliptic problems*, SIAM J. Sci. Comp., 15(5), pp. 1026–1037, 1994.

[39] J.C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Rapport no 49, Centre Math. Appli., Ecole Polytechnique, Mai, 1979.

[40] D.J. Nelske, J. Wolf and L.J. Howell, *Structural-acoustic finite element analysis of the automobile passenger comportement: A review of current practice*, J. of Sound and Vib. 80, pp. 247-266, 1982.

[41] M. Petyt and S.P. Lim, *Finite element analysis of the noise inside a mechanically excited cylinder*, Int. J. Num. Meth. Eng. 13, pp. 109-122, 1978.

[42] J.V. Ramakrishman and L.R. Koral, *A finite element model for sound transmission through laminate composite plates*, J. Sound Vib. 112(3), pp. 433-446, 1987.

[43] O.C. Zienkiewicz and R.E. Newton, *Coupled vibrations of a structure submerged in a compressible fluid*, Int. Symp. Finite Element Techn., Stuttgart, 1969.

[44] O.C. Zienkiewicz and P. Bettess, *Fluid-structure interaction and wave forces. An introduction to numerical treatment*, Int. J. Num. Meth. Eng. 13(1), pp. 1-17, 1978.

[45] O.C. Zienkiewicz and R.L. Taylor, *The finite element method*, Fourth edition. Mc Graw-hill, Vol. 2, 1991.

Bol. Soc. Esp. Mat. Apl. n°23(2003), 35–53

The total variation flow

F. Andreu[1], V. Caselles[2] and J. M. Mazón[1]

[1]Departamento de Análisis Matemático, Universitat de Valencia
[2]Departament de Tecnologia, Universitat Pompeu-Fabra

andreu@uv.es, vicent.caselles@tecn.upf.es, mazon@uv.es

Abstract

We summarize in this paper some of our recent results about the Minimizing Total Variation Flow, which have been mainly motivated by problems arising in Image Processing. First, we recall the role played by the Total Variation in Image Processing, in particular the variational formulation of the restoration problem. Next we state existence and uniqueness results for the Neumann and Dirichlet problems. This treatment also permits us to write the Euler-Lagrange equations of the variational formulation of the restoration problem in terms of a PDE. Finally, we study some qualitative properties and the asymptotic behaviour of the solutions of both, Neumann and Dirichlet, problems.

Key words: *Total Variation Flow, Image Restoration, Bounded Variation Functions, Asymptotic Behaviour, Nonlinear Parabolic Equations*

AMS subject classifications: *35K65, 35K55, 65M06, 47H20*

1 The total variation flow in image processing

We suppose that our image (or data) z is a scalar function defined on a bounded and piecewise smooth open set Ω of $\mathbb{R}^N$ - typically a rectangle in $\mathbb{R}^2$. Generally, the degradation of the original image u occurs during image acquisition and can be modeled by a linear and translation invariant blur and additive noise. The equation relating u to z can be written as

$$z = Ku + n \tag{1}$$

where K is a convolution operator with impulse response k, i.e., $Ku = k * u$, and n is an additive white noise of standard deviation σ. In practice, the noise can be considered as a gaussian.

Fecha de recepción: 19 de Junio de 2002

The problem of recovering u from z is ill-posed. First, the blurring operator need not be invertible. Second, if the inverse operator K^{-1} exists, applying it to both sides of (1) we obtain

$$K^{-1}z = u + K^{-1}n. \tag{2}$$

Writing $K^{-1}n$ in he Fourier domain, we have

$$K^{-1}n = \left(\frac{\hat{n}}{\hat{k}}\right)^{\vee}$$

where $\hat{f}$ denotes the Fourier transform of f and $f^{\vee}$ denotes the inverse Fourier transform.

From this equation, we see that the noise might blow up at the frequencies for which $\hat{k}$ vanishes or it becomes small.

Several methods have been proposed to recover u. Most of them can be classified as regularization methods which may take into account statistical properties (Wiener filters), information theoretic properties ([18]), a priori geometric models ([32]) or the functional analytic behaviour of the image given in terms of its wavelet coefficients ([20],[19]).

When we know nothing about the noise, we can set up the restoration problem as a least squares minimization. In this case we consider u and z to be deterministic. In the discrete case, to obtain the estimate of u from (1) we minimize the criterion

$$J(u) = \| Ku - z \|_2^2$$

which gives an estimate of u in terms of the pseudo-inverse of z, i.e.,

$$u^{+} = (K^tK)^{-1}K^tz,$$

where K^t is the adjoint of K (as far as K^tK is invertible). This is the linear algebraic approach to restoration. As it is well known [9] this estimate of u amplifies the noise due to the ill-conditioning of the operator K.

The typical strategy to solve this ill-conditioning is regularization. Then the solution of (1) is estimated by minimizing the functional

$$J_{\lambda}(u) = \| Ku - z \|_2^2 + \gamma \| Qu \|_2^2 \tag{3}$$

which yields the estimate

$$u_{\gamma} = (K^tK + \gamma Q^tQ)^{-1}K^tz \tag{4}$$

where Q is a linear regularization operator. Observe that to obtain u_{γ} we have to solve a system of linear equations. The role of Q is on one hand to move the small eigenvalues of K^tK away from zero while leaving the large eigenvalues unchanged, and, on the other hand, to incorporate the a priori (smoothness) knowledge that we have on u.

If we treat u and n as random vectors and we select $Q = R_f^{-1/2} R_n^{1/2}$ with R_f and R_n the image and noise covariance matrices, then (4) corresponds to the parametric Wiener filter. When $\lambda = 1$ this corresponds to the Wiener filter that minimizes the mean square error between the original and restored images.

Probably one of the first examples of regularization method [34], [25] consists in choosing between all possible solutions of (2) the one which minimizes the Sobolev (semi) norm of u

$$\int_\Omega |Du|^2 \, dx,$$

which corresponds to the case $Qu = \nabla u$. Then the solution of (3) given by (4) in the Fourier domain is given by

$$\hat{u} = \frac{\overline{\hat{k}}}{|\hat{k}|^2 + 4\gamma\pi^2|\xi|^2}\hat{z}.$$

From the above formula, we see that high frequencies of z (hence, the noise) are attenuated by the smoothness constraint. This was an important step, but the results were not satisfactory, mainly due to the unability of the previous functional to resolve discontinuities (edges) and oscillatory textured patterns. The smoothness constraint is too restrictive. Indeed, functions in $W^{1,2}(\Omega)$ (i.e., functions $u \in L^2(\Omega)$ such that $Du \in L^2(\Omega)$) cannot have discontinuities along rectifiable curves. These observations motivated the introduction of Total Variation in image restoration models by L. Rudin, S. Osher and E. Fatemi in their seminal work [32]. The a priori hypothesis is that functions of bounded variation (the BV model) ([3],[21],[36]) are a reasonable functional model for many problems in image processing, in particular, for restoration problems ([29],[32]). Tipically, functions of bounded variation have discontinuities along rectifiable curves, being continuous in some sense (in the measure theoretic sense) away from discontinuities. The discontinuities could be identified with edges. The ability of this functional to describe textures is less clear: some textures can be recovered, but up to a certain scale of oscillation. An interesting experimental discussion of the adequacy of the BV-model to describe real images can be seen in [2].

On the basis of the BV model, Rudin-Osher-Fatemi [32] proposed to solve the following constrained minimization problem

$$\begin{aligned} &\text{Minimize} \int_\Omega |Du| \, dx \\ &\text{with} \quad \int_\Omega Ku = \int_\Omega z, \quad \int_\Omega |Ku - z|^2 \, dx = \sigma^2|\Omega|. \end{aligned} \tag{5}$$

The first constraint corresponds to the assumption that the noise has zero mean, and the second that its standard deviation is σ. The constraints are a way to incorporate the image acquisition model given in terms on equation (1). Under the assumption $\|z - \int_\Omega z\|_2 \geq \sigma^2$, the constraint

$$\int_\Omega |Ku - z|^2 \, dx = \sigma^2|\Omega| \tag{6}$$

is equivalent to the constraint

$$\int_\Omega |Ku - z|^2 \, dx \leq \sigma^2 |\Omega|. \tag{7}$$

which amounts to say that σ is an upper bound of the standard deviation of n [17]. Moreover, assuming that $K1 = 1$, the constraint $\int_\Omega Ku = \int_\Omega z$ is automatically satisfied [17].

In practice, the above problem is solved via the following unconstrained minimization problem

$$\text{Minimize} \int_\Omega |Du| \, dx + \frac{\lambda}{2} \int_\Omega |Ku - z|^2 \, dx \tag{8}$$

for some Lagrange multiplier λ. The constraint has been introduced as a penalization term. The regularization parameter λ controls the trade-off between the goodness of fit of the constraint and the smoothness term given by the Total Variation. In this formulation, a methodology is required for a correct choice of λ. In [33], Rudin-Osher-Fu used the gradient projection method of Rosen ([30]) which leads to the gradient descent PDE associated to the problem (8) and updated λ so that the constraint (6) is satisfied. The analysis of such algorithm was initiated in [27]. The most succesful analysis of the connections between (5) and (8) were given by A. Chambolle and P.L. Lions in [17]. Indeed, based on the comments given in the above paragraph, they proved that both problems are equivalent for some positive value of the Lagrange multiplier λ.

Motivated by the image restoration problem we initiated in [6] the study of the minimizing total variation flow $u_t = div(\frac{Du}{|Du|})$. Indeed, this PDE is the gradient descent associated to the energy

$$\int_\Omega |Du|.$$

Observe that we are not considering the constraints given by the image acquisition model in this simplified energy. Thus our conclusions will not directly inform us about the complete model (5). Instead, our purpose was to understand how the minimizing total variation flow minimizes the total variation of a function. There are many flows which minimize the total variation of a function. Let us mention in particular the mean curvature motion ([28])

$$\frac{\partial u}{\partial t} = |Du| \text{div} \left(\frac{Du}{|Du|} \right). \tag{9}$$

Indeed, this flow corresponds to the motion of curves in $\mathbb{R}^2$ or hypersurfaces $S(t)$ in $\mathbb{R}^N$ by mean curvature, i.e.,

$$X_t = H\vec{N} \tag{10}$$

where X denotes a parametrization of $S(t)$, H denotes its mean curvature and $\vec{N}$ the outer unit normal. The classical motion given by (10) corresponds to the gradient descent of the area functional $\int_S dS$. Both flows, the classical mean curvature motion (10), and its viscosity solution (9) formulation have been studied by many authors, we refer in particular to the work by L.C. Evans and J. Spruck [22]. They proved, in particular, that the total variation of the (viscosity) solution of (9) decreases during the evolution, as it should happen since the flow decreases the $(N-1)$ Haussdoff measure of the level set surfaces of the solution u and the total variation corresponds to the integral of the (N-1) Hausdorff measure of the boundaries of the level sets. Let us compare the behaviour of the minimizing total variation flow with respect to the mean curvature motion flow. The viscosity solution formulation on the classical mean curvature motion has to be interpreted as follows. If $S(t)$ is a surface moving by mean curvature with initial condition $S(0)$, and $u(0,x)$ is the signed distance to $S(0)$, i.e., if $u(0,x) = d(x, S(0))$ when x is outside $S(0)$, and $u(0,x) = -d(x, C(0))$ if x is inside $S(0)$, then $S(t) = \{x : u(t,x) = 0\}$ for any $t \geq 0$, where $u(t,x)$ is the viscosity solution of (9). This is the level set formulation of the classical motion by mean curvature, initially proposed by S. Osher and J. Sethian in [28] and whose mathematical analysis was given in [22] and was followed by many other works. In particular, as it was shown by G. Barles, H.M. Soner and P. Souganidis [12], if instead of embedding $C(0)$ as the zero level set of a continuous function we just set $u(0,x) = \chi_{C(0)}$ where $C(0)$ is the region inside $S(0)$, and we assume that $S(0)$ is a smooth surface, then $u(t,x) = \chi_{C(t)}$ where $C(t)$ is the region inside $S(t)$. Thus, the mean curvature motion flow decreases the total variation of $\chi_{C(0)}$ by decreasing the $(N-1)$-Haussdorff measure of the boundary $S(t)$ of $C(t)$ [23]. Now, since the total variation of any function $u_0(x) = h\chi_C$ is

$$TV(h\chi_C) = hPer(C)$$

we see that two basic ways of minimizing the total variation of such a function are: either we decrease the heigth of $u_0(x)$ or we decrease the perimeter of its boundary. Our purpose was to explain which strategy was followed by the minimizing total variation flow. As we shall see below, under some geometric conditions for the sets $C(0)$, the strategy of the minimizing total variation flow consists in decreasing the heigth of the function without distortion of its boundary, while a distortion of the boundary will occur when these conditions are not satisfied, in particular, this will happen at points with a strong curvature. Thus the strategy followed by the minimizing total variation flow, compared to the one followed by the mean curvature motion is quite different. This gives an idea of the behavior of (5), at least what are the infinitesimal effects of (5) on the initial datum $u(0,x)$. The methods and results obtained can also be used to produce particular explicit solutions of the denoising problem which correspons to the kernel K in (8) being the identity, i.e., $K = I$.

2 The Neumann problem for the total variation flow

This section is devoted to the Neumann problem for the Total Variation Flow, namely

$$\begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}\left(\dfrac{Du}{|Du|}\right) & \text{in } Q = (0,\infty)\times\Omega \\[2ex] \dfrac{\partial u}{\partial \eta} = 0 & \text{on } S = (0,\infty)\times\partial\Omega \\[2ex] u(0,x) = u_0(x) & \text{in } x \in \Omega \end{cases} \tag{11}$$

where Ω is a bounded set in $\mathbb{R}^N$ with Lipschitz continuous boundary $\partial\Omega$ and $u_0 \in L^1(\Omega)$. As we showed in the first section, this partial differential equation appears when one uses the steepest descent method to minimize the Total Variation, a method introduced by L. Rudin and S. Osher [31], [32] in the context of image denoising and reconstruction. Then solving (11) amounts to regularize or, in other words, to filter the initial datum u_0. This filtering process has less destructive effect on the edges than filtering with a Gaussian, i.e., than solving the heat equation with initial condition u_0. In this context the given *image* u_0 is a function defined on a bounded, smooth or piecewise smooth open subset Ω of $\mathbb{R}^N$ -typically, Ω will be a rectangle in $\mathbb{R}^2$.

As argued in [1], the choice of Neumann boundary conditions is a natural choice in image processing. It corresponds to the reflection of the picture across the boundary and has the advantage of not imposing any value on the boundary and not creating edges on it.

Due to the linear growth of the energy functional associated with problems (11), the natural energy space for these problems is the space of functions of bounded variation. Recall that a function $u \in L^1(\Omega)$ whose partial derivatives in the sense of distributions are measures with finite total variation in Ω is called a *function of bounded variation.* The class of such functions will be denoted by $BV(\Omega)$. Thus $u \in BV(\Omega)$ if and only if there are Radon measures $\mu_1, \ldots, \mu_N$ defined in Ω with finite total mass in Ω and

$$\int_\Omega u D_i \varphi dx = -\int_\Omega \varphi d\mu_i$$

for all $\varphi \in C_0^\infty(\Omega)$, $i = 1, \ldots, N$. Thus the gradient of u is a vector valued measure with finite total variation

$$|Du|(\Omega) = \sup\left\{\int_\Omega u \operatorname{div}\varphi\, dx : \varphi \in C_0^\infty(\Omega,\mathbb{R}^n), |\varphi(x)| \le 1 \text{ for } x \in \Omega\right\}.$$

By $L^1_w(0,T,BV(\Omega))$ we denote the space of functions $w : [0,T] \to BV(\Omega)$ such that $w \in L^1((0,T)\times\Omega)$, the maps $t \in [0,T] \to < Dw(t), \phi >$ are measurable for every $\phi \in C_0^1(\Omega,\mathbb{R}^N)$ and $\displaystyle\int_0^T |Dw(t)|(\Omega)\, dt < \infty$. It is not difficult to see that the conditions on w imply the measurability of the map $t \in [0,T] \to |Dw(t)|(\Omega)$. We shall use the truncature functions defined by $T_k(r) = k \wedge (r \vee (-k))$, $k \geq 0$, $r \in \mathbb{R}$. Our concept of solution is the following

Definition 1 A measurable function $u : (0,T) \times \Omega \to \mathbb{R}$ is a *weak solution* of (11) in $(0,T) \times \Omega$ if $u \in C([0,T], L^1(\Omega)) \cap W^{1,1}_{loc}(0,T; L^1(\Omega))$, $T_k(u) \in L^1_w(0,T; BV(\Omega))$ for all $k > 0$ and there exists $z \in L^\infty((0,T) \times \Omega)$ with $\|z\|_\infty \leq 1$, $u_t = \mathrm{div}(z)$ in $\mathcal{D}'((0,T) \times \Omega)$ such that

$$\int_\Omega (T_k(u(t)) - w) u_t(t)\, dx \leq \int_\Omega z(t) \cdot \nabla w\, dx - |DT_k(u(t))|(\Omega) \qquad (12)$$

for every $w \in W^{1,1}(\Omega) \cap L^\infty(\Omega)$ and a.e. on $[0,T]$.

The main result of [6] is the following:

Theorem 1 *Let $u_0 \in L^1(\Omega)$. Then there exists a unique weak solution of (11) in $(0,T) \times \Omega$ for every $T > 0$ such that $u(0) = u_0$. Moreover, if $u(t), \hat{u}(t)$ are weak solutions corresponding to initial data $u_0, \hat{u}_0$, respectively, then*

$$\|(u(t) - \hat{u}(t))^+\|_1 \leq \|(u_0 - \hat{u}_0)^+\|_1 \quad \text{and} \quad \|u(t) - \hat{u}(t)\|_1 \leq \|u_0 - \hat{u}_0\|_1, \qquad (13)$$

for all $t \geq 0$.

To prove Theorem 1 we use the techniques of completely accretive operators ([14]) and the Crandall-Liggett's semigroup generation Theorem . For that, we associate a completely accretive operator $\mathcal{A}$ to the formal differential expression $-\mathrm{div}(\frac{Du}{|Du|})$ together with Neumann boundary conditions. Then, using Crandall-Liggett's semigroup generation Theorem we conclude that the abstract Cauchy problem in $L^1(\Omega)$

$$\left\{ \begin{array}{l} \dfrac{du}{dt} + \mathcal{A}u \ni 0, \\ \\ u(0) = u_0 \end{array} \right. \qquad (14)$$

has a unique strong solution $u \in C([0,T], L^1(\Omega)) \cap W^{1,1}_{loc}(0,T; L^1(\Omega))$ $(\forall T > 0)$ with initial datum $u(0) = u_0$. Then we shall prove that strong solutions of (14) coincide with weak solutions of (11).

Let us introduce the following operator $\mathcal{A}$ in $L^1(\Omega)$. Following Anzellotti [10], recall first that

$$X(\Omega) := \left\{ z \in L^\infty(\Omega, \mathbb{R}^N) \ : \ \mathrm{div}(z) \in L^1(\Omega) \right\}.$$

Definition 2

$(u,v) \in \mathcal{A}$ if and only if $u, v \in L^1(\Omega)$, $T_k(u) \in BV(\Omega)$ for all $k > 0$ and

there exists $z \in X(\Omega)$ with $\|z\|_\infty \leq 1$, $v = -\mathrm{div}(z)$ in $\mathcal{D}'(\Omega)$ such that

$$\int_\Omega (w - T_k(u)) v\, dx \leq \int_\Omega z \cdot \nabla w dx - |DT_k(u)|(\Omega),$$

$\forall w \in W^{1,1}(\Omega) \cap L^\infty(\Omega)$, $\forall k > 0$.

Theorem 2 *The operator $\mathcal{A}$ is m-completely accretive in $L^1(\Omega)$ with dense domain. For any $u_0 \in L^1(\Omega)$ the semigroup solution $u(t) = e^{-t\mathcal{A}}u_0$ is a strong solution of*

$$\frac{du}{dt} + \mathcal{A}u \ni 0, \qquad u(0) = u_0. \tag{15}$$

The accretivity of the operator $\mathcal{A}$ is proved using the integration by parts formula given in [10]. For that, we need first to prove that we can use test functions in $BV(\Omega) \cap L^\infty(\Omega)$ in the definition of $\mathcal{A}$.

Lemma 3 *We have the following characterization of the operator $\mathcal{A}$,*

$$(u, v) \in \mathcal{A} \quad \textit{if and only if } u, v \in L^1(\Omega),\ T_k(u) \in BV(\Omega) \textit{ for all } k > 0 \textit{ and}$$

$$\textit{there exists } z \in X(\Omega) \textit{ with } \|z\|_\infty \leq 1,\ v = -\mathrm{div}(z) \textit{ in } \mathcal{D}'(\Omega) \textit{ such that}$$

$$\int_\Omega (w - T_k(u))v \leq \int_\Omega (z, Dw) - |DT_k(u)|(\Omega), \quad \forall w \in BV(\Omega) \cap L^\infty(\Omega),\ \forall k > 0. \tag{16}$$

Moreover, we have that i) $\int_\Omega (z, DT_k(u)) = |DT_k(u)|(\Omega)$*, for all* $k > 0$*, and ii)* $\int_\Omega vT_k(u) = |DT_k(u)|(\Omega)$*, for all* $k > 0$*.*

Using this Lemma, we prove the accretivity of $\mathcal{A}$. Then Crandall-Liggett's semigroup generation Theorem proves the existence of a semigroup solution (also called mild solution) of (15). The proof of both existence and uniqueness of weak solutions in the sense of Definition 1 requires the regularity in time of the solution. This is proved using the regularizing effect due to the homogeneity of the operator in the case of Neumann boundary conditions.

The operator $\mathcal{A}$ permits to write the Euler-Lagrange equations of the functional (8) in terms of a partial differential equation. For simplicity, let us define

$$\Psi(u) := \begin{cases} \displaystyle\int_\Omega |Du| & \text{if} \quad u \in L^2(\Omega) \cap BV(\Omega) \\ +\infty & \text{if} \quad u \in L^2(\Omega) \setminus BV(\Omega). \end{cases} \tag{17}$$

The Euler-Lagrange equations of (8) are

$$-\lambda K^t(Ku - z) \in \partial\Psi(u) \tag{18}$$

The subdifferential of Ψ coincides with the operator $\mathcal{A}$ restricted to $L^2(\Omega)$.

Proposition 4 *We have $\partial\Psi = \mathcal{A} \cap (L^2(\Omega) \times L^2(\Omega))$.*

Let us mention an interesting geometric feature of the equation: the H^{N-1} measure of the boundaries of the level sets of the solution decreases with time.

Proposition 5 *Let $u_0 \in L^1(\Omega)$. Let $u(t,x)$ be the weak solution of (11). Then, for almost all $\lambda \in \mathbb{R}$,*

$$|D\chi_{\{u(t)>\lambda\}}| \leq |D\chi_{\{u(s)>\lambda\}}| \tag{19}$$

a.e. in $s,t \in (0,\infty)$, $t > s > 0$.

Note that $|D\chi_{\{u(t)>\lambda\}}| = H^{N-1}(\partial^*\{u(t)>\lambda\})$, where H^{N-1} is the $(N-1)$-dimensional Hausdorff measure and $\partial^*\{u(t)>\lambda\}$ is the reduced boundary of the set $\{x \in \Omega : u(t) > \lambda\}$. Thus, the above proposition says that the length of the boundaries of the level sets decreases with time.

Next, we prove that flat zones which are local maxima (minima) immediately decrease (respectively, increase) with time.

Proposition 6 *Let Ω be a cube in $\mathbb{R}^N$. Let $u_0 \in C(\overline{\Omega})$, $0 \leq u_0 \leq 1$. Suppose that $\{x \in \overline{\Omega} : u_0(x) = 1\} = K \subseteq B \subset\subset \Omega$ for some ball B. Let u be the weak solution of (11). Then $u(t,x) < 1$, for all $t > 0$, $x \in \Omega$.*

This result is proved by comparison with an explicit function satisfying the same property.

We can also compute explicitely the evolution of the characteristic function of a ball $B(p,r)$ when Ω is a ball centered at p. To fix ideas, let $p = 0$, $\Omega = B(0,R)$ and $u_0(x) = k\chi_{B(0,r)}$, where $0 < r < R$ and $k > 0$. We obtain that the solution of (11) in $(0,\infty) \times B(0,R)$ with initial datum u_0 is given by

$$u(t,x) = \left(k - \frac{N}{r}t\right)\chi_{B(0,r)} + \frac{Nr^{N-1}}{R^N - r^N}t\chi_{B(0,R)\setminus B(0,r)}$$

in $(0,T) \times B(0,R)$, where T is given by

$$T\left(\frac{N}{r} + N\frac{r^{N-1}}{R^N - r^N}\right) = k$$

and

$$u(t,x) = \left(k - \frac{N}{r}T\right)\chi_{B(0,R)} = \frac{Nr^{N-1}}{R^N - r^N}T\chi_{B(0,R)}, \quad \text{for } t \geq T.$$

3 The Dirichlet problem for the total variation flow

Suppose that Ω is an open bounded domain with Lipschitz boundary and $\varphi \in L^\infty(\partial\Omega)$. Let $\theta : \Omega \to \mathbb{R}^N$ be a vector field (whose smoothness will be precised below) with $|\theta| \leq 1$. In [11], a variational method was proposed to extend the data φ from $\partial\Omega$ to a function u in Ω along the integral curves of $\theta^\perp$, the vector orthogonal to θ, so that u is constant along the integral curves of $\theta^\perp$. Formally, we think of θ as the vector field made by the normals to the level sets of u, i.e., the sets $\{x \in \Omega : u(x) \geq \lambda\}$, $\lambda \in \mathbb{R}$. In that case we would have that

$\theta \cdot Du = |Du|$. In case that u is a function of bounded variation, almost all levels sets are of finite perimeter and, therefore, one can compute the normal along the boundary of the level sets (modulo a set of H^{N-1} null measure). Moreover, to get φ as a trace of a function u in Ω, the right function space is $BV(\Omega)$, the space of functions of bounded variation in Ω. Thus, to extend φ from $\partial\Omega$ to Ω, it was proposed in [11] to minimize the functional

$$F(u) = \int_\Omega |\nabla u| - \int_\Omega \theta \cdot \nabla u$$

defined in the set of functions of bounded variation $BV(\Omega)$ whose trace at the boundary is given by φ. Formally, if we integrate by parts in the second term of $F(u)$ we obtain

$$F(u) = \int_\Omega |\nabla u| + \int_\Omega \mathrm{div}(\theta) \cdot u - \int_{\partial\Omega} \theta \cdot \vec{n} u,$$

Since u, θ are known at the boundary, minimizing F amounts to minimize

$$E(u) = \int_\Omega |\nabla u| + \int_\Omega \mathrm{div}(\theta) \cdot u.$$

Let us comment on the class of admissible functions where E has to be minimized. We assume that $\mathrm{div}(\theta) \in L^1(\Omega)$ and $\varphi \in L^\infty(\partial\Omega)$. It seems reasonable to impose that the solution u is a bounded function with an L^∞ bound given by $\|\varphi\|_\infty$. Then the second integral in the definition of $E(u)$ is well defined. The first integral requires the use of the space of bounded variation functions. Thus our admissible class is $\mathcal{A} = \{u \in BV(\Omega) : |u(x)| \le \|\varphi\|_\infty \text{ a.e.}, u|_{\partial\Omega} = \varphi\}$. The final model is ([11])

$$\begin{array}{l} \text{Minimize} \displaystyle\int_\Omega |\nabla u| + \int_\Omega \mathrm{div}(\theta) \cdot u \\ u \in \mathcal{A} \end{array} \tag{20}$$

As it is well known ([24]) the solution of this problem has to be understood in a weak sense as the solution of the problem

$$\begin{array}{l} \text{Minimize} \displaystyle\int_\Omega |\nabla u| + \int_\Omega \mathrm{div}(\theta) \cdot u + \int_{\partial\Omega} |u - \varphi| dH^{N-1} \\ u \in BV(\Omega) \\ |u| \le \|\varphi\|_\infty. \end{array} \tag{21}$$

Existence for this variational problem was proved in [24] (Theorem 1.4), when $\theta \in L^1_{loc}(\Omega)^2$, $\mathrm{div}(\theta) \in L^1(\mathbb{R}^2)$, $\varphi \in L^\infty(\partial\Omega)$.

This is one of our motivations to study the Dirichlet problem

$$\left\{ \begin{array}{lll} \dfrac{\partial u}{\partial t} = \mathrm{div}\left(\dfrac{Du}{|Du|}\right) + f(t,x) & \text{in} & Q = (0,\infty) \times \Omega \\ u(t,x) = \varphi(x) & \text{on} & S = (0,\infty) \times \partial\Omega \\ u(0,x) = u_0(x) & \text{in} & x \in \Omega, \end{array} \right. \tag{22}$$

where $u_0 \in L^1(\Omega)$ and $\varphi \in L^\infty(\partial\Omega)$. This evolution equation is related to the gradient descent method used to minimize the functional (21), if we forget about the constraint $|u| \leq \|\varphi\|_\infty$. The constraint would introduce a further term in (22) but will not change the nature of the difficulties related to the solution of the PDE. We shall even make a further simplification, since we shall consider $f(t,x) = 0$. Hence, our aim is to study existence and uniqueness of solutions of the Dirichlet problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}\left(\dfrac{Du}{|Du|}\right) & \text{in} \quad Q = (0,\infty) \times \Omega \\ u(t,x) = \varphi(x) & \text{on} \quad S = (0,\infty) \times \partial\Omega \\ u(0,x) = u_0(x) & \text{in} \quad x \in \Omega \end{cases} \tag{23}$$

where Ω is an open bounded domain with a Lipschitz boundary, $u_0 \in L^1(\Omega)$ and $\varphi \in L^\infty(\partial\Omega)$.

The other motivation for the study of (23) comes from [4] and [15]. The general purpose of these works being the study of elliptic and parabolic problems in divergence form with initial data in L^1. Existence and uniqueness results of entropy solutions when the associated variational energy has a growth at infinity of order p with $p > 1$ are proved in [15] (see also [5], [16]).

In section 2 we consider the equation

$$u_t = \operatorname{div}\left(\frac{Du}{|Du|}\right) \tag{24}$$

in an open bounded Lipschitz domain with Neumann boundary conditions, proving existence and uniqueness of weak solutions. The main point being that, in the case of Neumann boundary conditions, this equation generates a nonlinear contraction semigroup in $L^1(\Omega)$ which is homogeneous of degree 0, a fact related to the regularity in time of the solutions on (24). Indeed, the homogeneity of the operator permits to conclude that $u_t(t) \in L^1(\Omega)$ a.e. for $t > 0$. This was used to prove uniqueness of solutions of (24) in case of Neumann boundary conditions. This property is loosed when we consider the case of Dirichlet boundary conditions. Thus, a different approach is needed.

To make precise our notion of solution we also need to introduce a weak trace on $\partial\Omega$ of the normal component of certain vector fields in Ω. We define the space

$$Z(\Omega) := \left\{ (z,\xi) \in L^\infty(\Omega, \mathbb{R}^N) \times BV(\Omega)^* \ : \ \operatorname{div}(z) = \xi \ \ in \ \ \mathcal{D}'(\Omega) \right\}.$$

We denote $R(\Omega) := W^{1,1}(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)$. For $(z,\xi) \in Z(\Omega)$ and $w \in R(\Omega)$ we define

$$\langle (z,\xi), w \rangle_{\partial\Omega} := \langle \xi, w \rangle_{BV(\Omega)^*, BV(\Omega)} + \int_\Omega z \cdot \nabla w.$$

Then, we obtain that if $w, v \in R(\Omega)$ and $w = v$ on $\partial\Omega$ one has

$$\langle (z,\xi), w\rangle_{\partial\Omega} = \langle (z,\xi), v\rangle_{\partial\Omega} \qquad \forall\ (z,\xi) \in Z(\Omega). \tag{25}$$

As a consequence of (25), we can give the following definition: Given $u \in BV(\Omega) \cap L^\infty(\Omega)$ and $(z,\xi) \in Z(\Omega)$, we define $\langle (z,\xi), u\rangle_{\partial\Omega}$ by setting

$$\langle (z,\xi), u\rangle_{\partial\Omega} := \langle (z,\xi), w\rangle_{\partial\Omega}$$

where w is any function in $R(\Omega)$ such that $w = u$ on $\partial\Omega$. We can prove that for every $(z,\xi) \in Z(\Omega)$ there exists $M_{z,\xi} > 0$ such that

$$|\langle (z,\xi), u\rangle_{\partial\Omega}| \le M_{z,\xi} \|u\|_{L^1(\partial\Omega)} \qquad \forall\ u \in BV(\Omega) \cap L^\infty(\Omega). \tag{26}$$

Now, taking a fixed $(z,\xi) \in Z(\Omega)$, we consider the linear functional $F : L^\infty(\partial\Omega) \to \mathbb{R}$ defined by

$$F(v) := \langle (z,\xi), w\rangle_{\partial\Omega}$$

where $v \in L^\infty(\partial\Omega)$ and $w \in BV(\Omega) \cap L^\infty(\Omega)$ is such that $w_{|\partial\Omega} = v$. By estimate (26), there exists $\gamma_{z,\xi} \in L^\infty(\partial\Omega)$ such that

$$F(v) = \int_{\partial\Omega} \gamma_{z,\xi}(x) v(x)\ dH^{N-1}.$$

Consequently there exists a linear operator $\gamma : Z(\Omega) \to L^\infty(\partial\Omega)$, with $\gamma(z,\xi) := \gamma_{z,\xi}$, satisfying

$$\langle (z,\xi), w\rangle_{\partial\Omega} = \int_{\partial\Omega} \gamma_{z,\xi}(x) w(x)\ dH^{N-1} \qquad \forall\ w \in BV(\Omega) \cap L^\infty(\Omega).$$

In case $z \in C^1(\overline{\Omega}, \mathbb{R}^N)$, we have $\gamma_z(x) = z(x) \cdot \nu(x)$ for all $x \in \partial\Omega$. Hence, the function $\gamma_{z,\xi}(x)$ is the weak trace of the normal component of (z,ξ). For simplicity of the notation, we shall denote $\gamma_{z,\xi}(x)$ by $[z,\nu](x)$.

We need to consider the space $BV(\Omega)_2$, defined as $BV(\Omega) \cap L^2(\Omega)$ endowed with the norm

$$\|w\|_{BV(\Omega)_2} := \|w\|_{L^2(\Omega)} + |Du|(\Omega).$$

It is easy to see that $L^2(\Omega) \subset BV(\Omega)_2^*$ and

$$\|w\|_{BV(\Omega)_2^*} \le \|w\|_{L^2(\Omega)} \qquad \forall\ w \in L^2(\Omega). \tag{27}$$

Now, it is well known (see [35]) that the dual $\left(L^1(0,T; BV(\Omega)_2)\right)^*$ is isometric to the space $L^\infty(0,T; BV(\Omega)_2^*, BV(\Omega)_2)$ of all weakly* measurable functions $f : [0,T] \to BV(\Omega)_2^*$, such that $v(f) \in L^\infty([0,T])$, where $v(f)$ denotes the supremum of the set $\{|\langle w, f\rangle| \ :\ \|w\|_{BV(\Omega)_2} \le 1\}$ in the vector lattice of measurable real functions. Moreover, the dual paring of the isometric is defined by

$$\langle w, f\rangle = \int_0^T \langle w(t), f(t)\rangle\ dt,$$

for $w \in L^1(0,T; BV(\Omega)_2)$ and $f \in L^\infty(0,T; BV(\Omega)_2^*, BV(\Omega)_2)$.

To make precise our notion of solution we need the following definitions.

Definition 3 Let $\Psi \in L^1(0,T;BV(\Omega))$. We say Ψ admits a *weak derivative* in $L^1_w(0,T;BV(\Omega))\cap L^\infty(Q_T)$ if there is a function $\Theta \in L^1_w(0,T;BV(\Omega))\cap L^\infty(Q_T)$ such that $\Psi(t) = \displaystyle\int_0^t \Theta(s)ds$, the integral being taken as a Pettis integral.

Definition 4 Let $\xi \in \big(L^1(0,T;BV(\Omega)_2)\big)^*$. We say that ξ is *the time derivative* in the space $\big(L^1(0,T;BV(\Omega)_2)\big)^*$ of a function $u \in L^1((0,T)\times\Omega)$ if

$$\int_0^T < \xi(t), \Psi(t) > dt = -\int_0^T \int_\Omega u(t,x)\Theta(t,x)dxdt$$

for all test functions $\Psi \in L^1(0,T;BV(\Omega))$ which admit a weak derivative $\Theta \in L^1_w(0,T;BV(\Omega)) \cap L^\infty(Q_T)$ and have compact support in time.

Observe that if $w \in L^1(0,T;BV(\Omega)) \cap L^\infty(Q_T)$ and $z \in L^\infty(Q_T,\mathbb{R}^N)$ is such that there exists $\xi \in \big(L^1(0,T;BV(\Omega))\big)^*$ with $\mathrm{div}(z) = \xi$ in $\mathcal{D}'(Q_T)$, we can define, associated to the pair (z,ξ), the distribution (z,Dw) in Q_T by

$$\langle (z,Dw),\phi\rangle := -\int_0^T \langle \xi(t), w(t)\phi(t)\rangle - \int_0^T \int_\Omega z(t,x)w(t,x)\nabla_x\phi(t,x) \qquad (28)$$

for all $\phi \in \mathcal{D}(Q_T)$.

Definition 5 Let $\xi \in \big(L^1(0,T;BV(\Omega)_2)\big)^*$, $z \in L^\infty(Q_T,\mathbb{R}^N)$. We say that $\xi = \mathrm{div}(z)$ in $\big(L^1(0,T;BV(\Omega)_2)\big)^*$ if (z,Dw) is a Radon measure in Q_T with normal boundary values $[z,\nu] \in L^\infty((0,T)\times\partial\Omega)$, such that

$$\int_{Q_T}(z,Dw) + \int_0^T < \xi(t), w(t) > dt = \int_0^T \int_{\partial\Omega} [z(t,x),\nu]w(t,x)dH^{N-1}dt,$$

for all $w \in L^1(0,T;BV(\Omega)) \cap L^\infty(Q_T)$.

Let $\mathcal{T} = \{T_k, T_k^+, T_k^- \ : \ k > 0\}$. We need to consider a more general set of truncature functions, concretely, the set $\mathcal{P}$ of all nondecreasing continuous fuctions $p : \mathbb{R} \to \mathbb{R}$, such that there exists p' except a finite set and $\mathrm{supp}(p')$ is compact. Obviously, $\mathcal{T} \subset \mathcal{P}$.

Definition 6 A measurable function $u : (0,T)\times\Omega \to \mathbb{R}$ is an *entropy solution* of (23) in $Q_T = (0,T)\times\Omega$ if $u \in C([0,T];L^1(\Omega))$, $p(u(\cdot)) \in L^1_w(0,T;BV(\Omega))$ $\forall\ p \in \mathcal{T}$ and there exist $(z(t),\xi(t)) \in Z(\Omega)$ with $\|z(t)\|_\infty \leq 1$, and $\xi \in \big(L^1(0,T;BV(\Omega)_2)\big)^*$ such that ξ is the time derivative of u in $\big(L^1(0,T;BV(\Omega)_2)\big)^*$, $\xi = \mathrm{div}(z)$ in $\big(L^1(0,T;BV(\Omega)\big)^*$ and $[z(t),\nu] \in \mathrm{sign}\big(p(\varphi) - p(u(t))\big)$ a.e. in $t \in [0,T]$, satisfying

$$-\int_0^T \int_\Omega j(u(t)-l)\eta_t + \int_0^T \int_\Omega \eta(t)|Dp(u(t)-l)|(\Omega) + z(t)\cdot D\eta(t)p(u(t)-l)$$

$$\leq \int_0^T \int_{\partial\Omega} [z(t),\nu]\eta(t)p(u(t)-l),$$

for all $l \in \mathbb{R}$, for all $\eta \in C^\infty(\overline{Q_T})$, with $\eta \geq 0$, $\eta(t,x) = \phi(t)\psi(x)$, being $\phi \in \mathcal{D}(]0,T[)$, $\psi \in C^\infty(\overline{\Omega})$ and $p \in \mathcal{T}$, where $j(r) = \displaystyle\int_0^r p(s)\, ds$.

The main result of [7] is the following.

Theorem 7 *Let $u_0 \in L^1(\Omega)$, and $\varphi \in L^1(\partial\Omega)$. Then there exists a unique entropy solution of (23) in $(0,T)\times\Omega$ for every $T > 0$ such that $u(0) = u_0$. Moreover, if $u(t), \hat{u}(t)$ are the entropy solutions corresponding to initial data $u_0, \hat{u}_0$, respectively, then*

$$\|\big(u(t)-\hat{u}(t)\big)^+\|_1 \leq \|\big(u_0-\hat{u}_0\big)^+\|_1 \quad \text{and} \quad \|u(t)-\hat{u}(t)\|_1 \leq \|u_0-\hat{u}_0\|_1 \tag{29}$$

for all $t \geq 0$.

To prove the existence part of Theorem 7 we use the techniques of completely accretive operators and the Crandall-Liggett's semigroup generation Theorem. So we associate a completely accretive operator $\mathcal{A}_\varphi$ to the formal differential expression $-\mathrm{div}(\frac{Du}{|Du|})$ together with the Dirichlet boundary condition.

Let us introduce the following operator $\mathcal{A}_\varphi$ in $L^1(\Omega)$.

Definition 7 $(u,v) \in \mathcal{A}_\varphi$ if and only if $u, v \in L^1(\Omega)$, $p(u) \in BV(\Omega)$ for all $p \in \mathcal{P}$ and there exists $z \in X(\Omega)$ with $\|z\|_\infty \leq 1$, $v = -\mathrm{div}(z)$ in $\mathcal{D}'(\Omega)$ such that

$$\int_\Omega (w-p(u))v \leq \int_\Omega z\cdot\nabla w - |Dp(u)|(\Omega) + \int_{\partial\Omega} |w-p(\varphi)| - \int_{\partial\Omega} |p(u)-p(\varphi)|,$$

$\forall w \in W^{1,1}(\Omega)\cap L^\infty(\Omega)$ and $\forall p \in \mathcal{P}$.

Theorem 8 *Let $\varphi \in L^1(\partial\Omega)$. The operator $\mathcal{A}_\varphi$ is m-completely accretive in $L^1(\Omega)$ with dense domain.*

To prove uniqueness we shall show that the entropy solutions and semigroup solutions coincide. As consequence of the semigroup theory, (29) is satisfied. Our technique is inspired by a method introduced by Kruzhkov [26] to prove L^1-contraction for entropy solutions for scalar conservation laws: the doubling of variables.

4 Asymptotic behaviour of the solutions

In [8] we study the asymptotic behaviour of the solutions of the Dirichlet problem

$$P_D \begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}\left(\dfrac{Du}{|Du|}\right) & \text{in} \quad Q = (0,\infty) \times \Omega \\ u(t,x) = 0 & \text{on} \quad S = (0,\infty) \times \partial\Omega \\ u(0,x) = u_0(x) & \text{in} \quad x \in \Omega \end{cases}$$

and the Neumann problem

$$P_N \begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}\left(\dfrac{Du}{|Du|}\right) & \text{in} \quad Q = (0,\infty) \times \Omega \\ \dfrac{\partial u}{\partial \eta} = 0 & \text{on} \quad S = (0,\infty) \times \partial\Omega \\ u(0,x) = u_0(x) & \text{in} \quad x \in \Omega \end{cases}$$

for the Total Variational Flow. We describe the behaviour of solutions of the Dirichlet problem (P_D) near the extinction time (we prove that it is finite). This behaviour is described by a function which is a solution of an eigenvalue problem for the operator $-\operatorname{div}\left(\frac{Du}{|Du|}\right)$ and we describe the solutions of this eigenvalue problem in the radial case. Moreover, the explicit solution found for the case in which $u_0 = k\chi_{B(0,r)}$, with $B(0,r) \subset\subset \Omega$, allows us to point out other qualitative properties which are peculiar of this special class of quasilinear equations. For instance, there is an infinite "waiting time", i.e. there is no propagation of the support of the initial datum and, which is more relevant, the solution is discontinuous and has a spatial minimal regularity: $u(t,.) \in BV(\Omega) \setminus W^{1,1}(\Omega)$ for any $t \in [0,+\infty)$ (i.e. the solution does not win any spatial differentiability, in contrast to what happens for the linear heat equation and many other quasilinear parabolic equations).

Respect to the Dirichlet problem our main result is the following.

Theorem 9 *Let $u_0 \in L^\infty(\Omega)$ and let $u(t,x)$ be the unique solution of problem (P_D). Let $d(\Omega)$ be the smallest radius of a ball containing Ω. If $T^*(u_0) = \inf\{t > 0 : \quad u(t) = 0\}$, then*

$$T^*(u_0) \le \frac{d(\Omega)\|u_0\|_\infty}{N}. \tag{30}$$

Let

$$w(t,x) := \begin{cases} \dfrac{u(t,x)}{T^*(u_0) - t} & \text{if} \quad 0 \le t < T^*(u_0), \\ 0 & \text{if} \quad t \ge T^*(u_0). \end{cases}$$

Then, there exists an increasing sequence $t_n \to T^*(u_0)$ and a solution $v^* \neq 0$ of the stationary problem

$$S_D \begin{cases} -\mathrm{div}\left(\dfrac{Dv}{|Dv|}\right) = v & \text{in} \quad \Omega \\ \\ v = 0 & \text{on} \quad \partial\Omega \end{cases}$$

such that

$$\lim_{n\to\infty} w(t_n) = v^* \quad \text{in} \quad L^p(\Omega)$$

for all $1 \leq p < \infty$. Moreover v^* is a minimizer of $\Phi(\cdot) - < \cdot, v^* >$ in $BV(\Omega) \cap L^2(\Omega)$.

For the Neumann problem we obtain the following results.

Theorem 10 *Let $u_0 \in L^1(\Omega)$ and let $u(t,x)$ be the unique solution of problem (P_D). Then*

$$\|u(t) - \overline{u_0}\|_1 \to 0 \quad \text{as} \quad t \to \infty,$$

where

$$\overline{u_0} = \frac{1}{\mathcal{L}^N(\Omega)} \int_\Omega u_0(x)\, dx.$$

Theorem 11 *Suppose $N = 2$. Let $u_0 \in L^\infty(\Omega)$ and let $u(t,x)$ be the unique weak solution of problem (P_N). Let*

$$w(t,x) := \begin{cases} \dfrac{u(t,x) - \overline{u_0}}{T^*(u_0) - t} & \text{if} \quad 0 \leq t < T^*(u_0), \\ \\ 0 & \text{if} \quad t \geq T^*(u_0). \end{cases}$$

Then, there exists an increasing sequence $t_n \to T^*(u_0)$, and a solution $v^* \neq 0$ of the stationary problem

$$S_N \begin{cases} -\mathrm{div}\left(\dfrac{Dv}{|Dv|}\right) = v & \text{in} \quad \Omega \\ \\ \dfrac{\partial v}{\partial \eta} = 0 & \text{on} \quad \partial\Omega \end{cases}$$

such that

$$\lim_{n\to\infty} w(t_n) = v^* \quad \text{in} \quad L^p(\Omega)$$

for all $1 \leq p < \infty$. Moreover v^* is a minimizer of $\Psi(\cdot) - < \cdot, v^* >$ in $BV(\Omega) \cap L^2(\Omega)$.

References

[1] L. Alvarez, P. L. Lions, and J. M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal. **29** (1992), pp. 845-866.

[2] L. Alvarez, Y. Gousseau and J.M. Morel, *The size of objects in natura images,* Preprint CMLA, 1999.

[3] L. Ambrosio, N. Fusco and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Mathematical Monographs, 2000.

[4] F. Andreu, J.M. Mazón, S. Segura and J. Toledo, *Quasilinear Elliptic and Parabolic Equations in L^1 with Nonlinear Boundary Conditions*, Adv. in Math. Sci. and Appl. **7** (1977), 183-213.

[5] F. Andreu, J.M. Mazón, S. Segura and J. Toledo, *Existence and Uniqueness for a degenerate Parabolic Equation with L^1-data*, Trans. Amer. Math. Soc. **315** (1999), 285-306.

[6] F. Andreu, C. Ballester, V. Caselles and J. M. Mazón, *Minimizing Total Variation Flow*, Diff. and Int. Eq. **14** (2001), 321-360.

[7] F. Andreu, C. Ballester, V. Caselles and J. M. Mazón, *The Dirichlet problem for the total variational flow*, J. Funct. Anal. **180** (2001), 347-403.

[8] F. Andreu, V. Caselles, J.I. Diaz, and J.M. Mazón. *Qualitative properties of the total variation flow.* J. Funct. Analysis **188** (2002), 516-547.

[9] H.C. Andrews and B.R. Hunt, *Digital Signal Processing,* Tech. Englewood Cliffs, NJ, Prentice Hall, 1977.

[10] G. Anzellotti, *Pairings Between Measures and Bounded Functions and Compensated Compactness*, Ann. di Matematica Pura ed Appl. IV (135) (1983), 293-318.

[11] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro and J. Verdera, *Filling-In by Joint Interpolation of Vector Fields and Grey Levels*, IEEE Transactions on Image Processing, vol. 10(8), pp. 1200-1211, 2001.

[12] G, Barlet, H. M. Soner and P. Souganidis. *Front propagation and phase field theory.* J. Control Optim **31** (1993), 439-469.

[13] G. Bellettini, V. Caselles and M. Novaga, *The Total Variation Flow in $\mathbb{R}^N$,* To appear in Journal Differential Equations.

[14] Ph. Bénilan and M.G. Crandall, *Completely Accretive Operators*, in Semigroups Theory and Evolution Equations, Ph. Clement et al. editors, Marcel Dekker, 1991, pp. 41-76.

[15] Ph. Bénilan, L. Boccardo, T. Gallouet, R. Gariepy, M. Pierre and J.L Vazquez, *An L^1-Theory of Existence and Uniqueness of Solutions of Nonlinear Elliptic Equations*, Ann. Scuola Normale Superiore di Pisa, IV, Vol. XXII (1995), 241-273.

[16] D. Blanchard and F. Murat, *Renormalised solutions of nonlinear parabolic problems with L^1-data: existence and uniqueness*, Proc. Roy Soc. Edinburgh Sect A. **127** (1997), 1137-1152.

[17] A. Chambolle and P. L. Lions, *Image Recovery via Total Variation Minimization and Related Problems*, Numer. Math. **76** (1997), 167-188.

[18] G. Demoment, *Image reconstruction and restoration: Overview of common estimation structures and problems,* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, 12 (1989), pp. 2024-2036.

[19] D. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage,* Biometrika, 81(3), (1994), pp. 425-455.

[20] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. *Wavelet shrinkage: Asymptopia ?* Journal R. Stat. Soc. B **57** (1995), 301–369.

[21] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Math., CRC Press, 1992.

[22] L. C. Evans and J. Spruck *Motion of level sets by mean curvature I*, J. Diff. Geometry **33** (1991), 635-681.

[23] L. C. Evans and J. Spruck *Motion of level sets by mean curvature II*, Trans. Amer. math. Soc, **330** (1992), 321-332.

[24] M. Giaquinta, G. Modica and J. Soucek, *Functionals with linear growth in the calculus of variations I*, Comment. Math. Univ. Carolinae **20** (1979), 143-156.

[25] C.W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind,* Pitman, Boston, 1984.

[26] S.N. Kruzhkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb. **10** (1970), 217-243.

[27] P. L. Lions, S. Osher and L. Rudin, *Denoising and Deblurring using Constrained Nonlinear Partial Differential Equations*, Tech. Repport, Cognitech Inc., Santa Monica, CA, 1992, submitted to SINUM.

[28] S. Osher and J. A. Sethian *Fronts propagating with curvature-dependent speed: Algorithms base on hamilton-Jacabi formulations* J. of Comp. Phys. **79** (1988), 12-49.

[29] L. Rudin, *Images, Numerical Analysis of Singularities and Shock Filters,* Ph. D. Thesis, Caltech 1987.

[30] J.G. Rosen, *The Gradient Projection Method for Nonlinear Programming. Part II, Nonlinear Constraints,* J. Soc. Indust. Appl. Math. 9 (1961), 514-532.

[31] L. Rudin and S. Osher, *Total Variation based Image Restoration with Free Local Constraints*, Proc. of the IEEE ICIP-94, vol. 1, Austin, TX, 1994, pp. 31-35.

[32] L. Rudin, S. Osher and E. Fatemi, *Nonlinear Total Variation based Noise Removal Algorithms*, Physica D.**60** (1992), 259-268.

[33] L. Rudin, S. Osher and C. Fu, *Total Variation Based Restoration of Noisy, Blurred Images,* SINUM.

[34] A. N. Tikhonov and V. Y. Arsemin, *Solutions of Ill-Posed problems*, John Wiley, New York, 1977.

[35] L. Schwartz, *Analyse IV. Applications a la théorie de la mesure*, Hermann, 1993.

[36] W. P. Ziemer, *Weakly Differentiable Functions*, GTM 120, Springer Verlag, 1989.

Bol. Soc. Esp. Mat. Apl. n°23(2003), 55–90

On two models of orthogonal polynomials and their applications

J. Arvesú and F. Marcellán

Departamento de Matemáticas. Universidad Carlos III de Madrid

jarvesu@math.uc3m.es, pacomarc@ing.uc3m.es

Abstract

This contribution deals with some models of orthogonal polynomials as well as their applications in several areas of mathematics. Some new trends in the theory of orthogonal polynomials are summarized. In particular, we emphasize on two kinds of orthogonality, i.e., the standard orthogonality in the unit circle and a non standard one, which is called multi-orthogonality. Both have attracted the interest of researchers during the past ten years.

Key words: *Linear functionals, Orthogonal polynomials, Stieltjes and Markov functions, rational approximation, Hermite-Padé approximation, multiple orthogonal polynomials, trigonometric Moment problem, Quadratures on the unit circle, linear prediction, electrostatics of zeros.*

AMS subject classifications: *33C45, 42C05*

1 Introduction

The theory of orthogonal polynomials has experienced a relevant growth and an increasing interest during the last twenty years as a consequence, among others, of a substantial revaluation of our perception concerning their nature and applicability. The popularity of orthogonal polynomials is due, in particular, to Louis de Branges's solution of the Beiberbach conjecture which uses an inequality of Askey and Gasper on Jacobi polynomials. But the main reason lies in their wide applications in many areas as approximation theory (Padé approximations [50], continued fractions) numerical analysis, scattering theory, digital signal processing, electrical engineering, theoretical chemistry, solid state physics (Toda lattices) atomic and nuclear physics (eigenfunctions of Hamiltonians) and so forth.

The contemporary general theory of orthogonal polynomials was initiated by G. Szegő in a series of papers starting in 1915. It was Szegő who realized

Fecha de recepción: 19 de Junio de 2002

that real orthogonal polynomials can better be understood by first studying complex orthogonal polynomials on the unit circle. His monograph “Orthogonal polynomials” [55], whose first edition was published in 1939, has been one of the basic and most popular references on the subject. In the decade of the seventies the connection with special functions (mostly hypergeometric and basic hypergeometric functions) provided a creative approach to new ideas coming from mathematical physics (quantum oscillators, Schrödinger equations, and Klein-Gordon equations among others). The monographs by G. Gasper and M. Rahman “Basic hypergeometric series” [25] as well as the books by A. F. Nikiforov and V. B. Uvarov [48] and A. F. Nikiforov, S. K. Suslov and V. B. Uvarov [47] constitute a good sample of this approach. For more information the survey contribution [3] is a nice presentation for the non specialized reader.

On the other hand, the new trends in numerical quadrature [24], spectral methods for boundary value problems [11, 26] and the powerful tools derived of the recent progress in potential theory have consolidated the research not only from a theoretical point of view but by the interactions with other domains as harmonic analysis, operator theory and matrix analysis.

The starting point is the concept of orthogonality with respect to a measure μ supported on an infinite subset Δ of the real line.

If we assume that $\int_\Delta p(x)d\mu(x)$ converges for every polynomial p, then we can introduce an inner product

$$\langle p, q\rangle = \int_\Delta p(x)q(x)d\mu(x),$$

where p, q are polynomials. For such an inner product, which we will say to be standard, a sequence of polynomials $\{p_n\}_{n\in\mathbb{N}}$ is said to be orthogonal with respect to the above inner product if

(i) $p_n(x) = a_n x^n +$ lower degree terms, and $a_n > 0$.

(ii) $\langle p_n, p_m\rangle = \delta_{n,m}$, $m, n \in \mathbb{N}$.

From the definition of the inner product it is straightforward to prove that the sequence $\{p_n\}_{n\in\mathbb{N}}$ satisfies a three-term recurrence relation [19]

$$xp_n(x) = a_{n+1}p_{n+1}(x) + b_n p_n(x) + a_n p_{n-1}(x), \quad p_{-1}(x) = 0, \quad p_0(x) = 1, \tag{1}$$

where $a_n > 0$ and $b_n \in \mathbb{R}$. This recurrence relation contains a basic information both about the behavior of the polynomials as well as the properties of the orthogonality measure.

Non standard examples of orthogonality have been studied intensively during the last decades. For instance:

1. If the measure μ is supported in some domain of $\mathbb{R}^N$ ($N \in \mathbb{N}$), then the orthogonality in the space of polynomials $\mathbb{R}[x_1, \ldots, x_N]$ in N variables constitute an attractive research areas because of their connections with group theory, numerical cubature, etc. Unfortunately, very few monographs have been devoted to this subject and it constitutes one of the most promissing subjects for the research in the next future [21].

2. If μ is a matrix of measures supported on some subset of the real line, the orthogonality in the space of matrix polynomials $\mathbb{R}^{N\times N}[x]$ has been analyzed in a wide set of contributions both from the analytical point of view as from their applications in numerical quadrature (see [22, 23] among others).

3. If $(\mu_0, \mu_1, \ldots, \mu_r)$ is a vector of measures supported on subsets $(\Delta_0, \Delta_1, \ldots, \Delta_r)$ of the real line, then an inner product on the linear space of polynomials

$$\langle p, q\rangle = \sum_{i=0}^{r} \int_{\Delta_i} p^{(i)}(x)q^{(i)}(x)d\mu_i(x),$$

yields an interesting family of polynomials (which are called Sobolev orthogonal polynomials) with potential applications in boundary value problems, smooth least square approximation, etc. [42, 45, 46].

In this contribution we will focus our attention in two kind of non standard families of orthogonal polynomials.

First, we will consider a positive Borel measure μ supported on the unit circle $\mathbb{T}$ and we will introduce an Hermitian inner product

$$\langle p(z), q(z)\rangle = \int_{\mathbb{T}} p(z)\overline{q(z)}d\mu(z).$$

In Section 2 we will discuss some properties of the corresponding sequences of orthogonal polynomials. More precisely, we will analyze quadrature formulas related to knots located on $\mathbb{T}$ as well as some recent results about differential properties of those orthogonal polynomials. As an interesting application we will present the modellization of optimal linear predictors for stationary discrete-time stochastic processes.

Second, we will deal with multiple orthogonal polynomials, i.e., we distribute the orthogonality conditions over a fixed number of intervals. In Section 3 we define two kinds of multiple orthogonality and we explain how they are closely related to simultaneous rational approximation of a system of Markov functions. Special emphasis will be done when the orthogonality conditions are considered with respect to discrete measures (Hahn, Meixner, Kravchuk and Charlier).

2 Orthogonal polynomials on the unit circle

Let μ be a probability measure supported on the unit circle $\mathbb{T} := \{z \ : \ |z| = 1\}$. Associated with μ we can introduce in the linear space $\mathbb{P}$ of polynomials with complex coefficients an inner product

$$\langle p(z), q(z)\rangle = \int_{\mathbb{T}} p(z)\overline{q(z)}d\mu, \quad p, q \in \mathbb{P}. \tag{2}$$

Taking into account the linear operator $\mathcal{H} : \mathbb{P} \mapsto \mathbb{P}$, $(\mathcal{H}p)(z) = zp(z)$ is a unitary operator with respect to the inner product (2), the Gram matrix

$\mathcal{T}$ associated with (2) in terms of the canonical basis $\{z^n\}_{n\in\mathbb{N}}$ has a special structure. The corresponding (j,k) entry is given by $\langle z^j, z^k\rangle = \langle z^{j-k}, 1\rangle$ for $j \geq k$, and $\langle z^j, z^k\rangle = \langle 1, z^{k-j}\rangle$ for $j \leq k$.

In other words, in the first row of the infinite matrix we can concentrate the information about it. Indeed, the entries of every subdiagonal are equal. In the literature such a kind of matrices are called Toeplitz matrices. Let $c_n = \langle z^n, 1\rangle$. We will say c_n is the n-th moment of μ.

If we use the Gram-Schmidt orthogonalization method for the canonical basis, we obtain a unique sequence of monic polynomials $\{\phi_n\}_{n\in\mathbb{N}}$ such that

$$\langle \phi_n(z), z^j\rangle = 0, \quad j = 0, 1, \ldots, n-1.$$

Notice that this orthogonality condition means that the coefficients of the polynomials $\phi_n(z) = z^n + a_{n,n-1}z^{n-1} + \cdots + a_{n,1}z + a_{n,0}$ are the solutions of a system of linear equations

$$\begin{array}{ccccccccc}
a_{n,0}c_0 & + & a_{n,1}c_1 & + & \cdots & + & a_{n,n-1}c_{n-1} & = & -c_n, \\
a_{n,0}\overline{c_1} & + & a_{n,1}c_0 & + & \cdots & + & a_{n,n-1}c_{n-2} & = & -c_{n-1}, \\
\vdots & & \vdots & & \cdots & & \vdots & \vdots & \vdots \\
a_{n,0}\overline{c_{n-1}} & + & a_{n,1}\overline{c_{n-2}} & + & \cdots & + & a_{n,n-1}c_0 & = & -c_1.
\end{array}$$

In other words

$$\mathcal{T}_n(a_{n,0}, a_{n,1}, \ldots, a_{n,n-1})^{\mathrm{T}} = -(c_n, c_{n-1}, \ldots, c_1)^{\mathrm{T}}, \tag{3}$$

where $\mathcal{T}_n$ is the principal submatrix of $\mathcal{T}$ with dimension n. Thus, from the orthogonality condition we get a system of linear equations whose matrix has an underlying Toeplitz structure. The solution of it yields in a natural way to the orthogonal polynomials with respect to the measure μ whose moments are $\{c_n\}_{n\in\mathbb{N}}$. They can also be characterized as the solution of the following extremal problem: *Minimize* $\langle p, p\rangle$ *over all monic polynomials* $p \in \mathbb{P}$.

In the sequel, we will consider the norm induced by the inner product (2) and we will denote it $\|p\| = \sqrt{\langle p, p\rangle}$.

From the orthogonality conditions it follows that the zeros of ϕ_n lie in the unit disk. Indeed, if $\phi_n(\alpha) = 0$, then from $\phi_n(z) = (z-\alpha)q_{n-1}(z)$ we get

$$0 < \|\phi_n\|^2 = (1 - |\alpha|^2)\, \|q_{n-1}\|^2,$$

i.e. $|\alpha| < 1$.

If we introduce the reversed polynomial $\phi_n^*(z) = z^n\overline{\phi_n}(z^{-1})$ we get

$$\langle \phi_n^*(z), z^k\rangle = 0, \quad 1 \leq k \leq n.$$

On the other hand,

$$\langle \phi_{n+1}(z) - z\phi_n(z), z^k\rangle = 0, \quad 1 \leq k \leq n.$$

Consequently, one gets

$$\phi_{n+1}(z) - z\phi_n(z) = \lambda_n\phi_n^*(z),$$

where $\lambda_n = \phi_{n+1}(0)$. The above expression leads to

$$\phi_{n+1}(z) = z\phi_n(z) + \phi_{n+1}(0)\phi_n^*(z), \tag{4}$$

which is a forward recurrence relation for the sequence $\{\phi_n\}_{n\in\mathbb{N}}$. Apparently, in order to obtain ϕ_{n+1} we only need the value $\phi_{n+1}(0)$ but it can be deduced from

$$\langle z\phi_n(z), 1\rangle = -\phi_{n+1}(0)\,\langle \phi_n^*(z), 1\rangle\,.$$

Taking into account

$$\langle \phi_n^*(z), 1\rangle = \|\phi_n\|^2\,,$$

we get

$$\phi_{n+1}(0) = -\frac{c_{n+1} + \sum_{j=1}^{n} a_{n,j-1}c_j}{\|\phi_n\|^2},$$

together with

$$\|\phi_n\|^2 = c_0 + \sum_{j=1}^{n} \overline{a}_{n,n-j}c_j.$$

The recurrence relation (4) was introduced by G. Szegő and constitutes the polynomial counterpart of the Levinson algorithm for the solution of (3) in a finite number of steps [16]. The values $\{\phi_n(0)\}_{n\in\mathbb{N}}$ are called reflection parameters. Notice that $|\phi_n(0)| < 1$ for every $n \in \mathbb{N}$, which is clearly perceptible taking into account the fact that the zeros of ϕ_n lie in the unit disk.

On the other hand, the polynomial $\phi_{n+1}(z) - \phi_{n+1}(0)\phi_{n+1}^*(z)$ satisfies

$$\left\langle \phi_{n+1}(z) - \phi_{n+1}(0)\phi_{n+1}^*(z), z^k \right\rangle = 0, \quad 1 \le k \le n+1,$$

as well as it vanishes for $z = 0$. This means

$$\phi_{n+1}(z) - \phi_{n+1}(0)\phi_{n+1}^*(z) = s_n z\phi_n(z),$$

with $s_n = 1 - |\phi_{n+1}(0)|^2$. Hence,

$$\phi_{n+1}(z) = (1 - |\phi_{n+1}(0)|^2)z\phi_n(z) + \phi_{n+1}(0)\phi_{n+1}^*(z), \tag{5}$$

holds. The above expression (5) is a backward recurrence relation for the sequence $\{\phi_n\}_{n\in\mathbb{N}}$. It was introduced by G. Szegő and it is intimately related with the Schur-Cohn algorithm, a standard way to characterize the polynomials whose zeros lie in the unit disk [16]. This algorithm is very well known in the theory of discrete linear systems and provides a tool for the study of their stability.

Notice that both recurrence relations (4) and (5) are substantially different with respect to the three-term recurrence relation associated with orthogonal polynomials in the real case. Here we need only one parameter in order to generate the sequence of monic polynomials but the reversed polynomial appears as a counterpart.

As a first conclusion, given a probability measure μ we can associate a sequence of moments $\{c_n\}_{n\in\mathbb{N}}$. From them, we obtain the reflection parameters $\{a_n\}_{n\in\mathbb{N}}$ where $a_n = \phi_n(0)$, and thus we get the sequence of monic polynomials orthogonal with respect to μ. What about the converse problem, i.e., how get the measure μ from the moments $\{c_n\}_{n\in\mathbb{N}}$ or from the reflection parameters $\{a_n\}_{n\in\mathbb{N}}$?

2.1 Trigonometric moment problem and quadrature formulas

Given a sequence of complex numbers $\{c_n\}_{n\in\mathbb{N}}$ the trigonometric moment problem consists in finding necessary and sufficient conditions for the existence of a probability measure μ supported on the unit circle such that

$$c_n = \int_{\mathbb{T}} z^n d\mu, \quad n \in \mathbb{N}.$$

Theorem 1 *[30] A necessary and sufficient condition for the solution of the trigonometric moment problem is*

$$\sum_{j,k} c_{j-k} x_j \overline{x}_k \geq 0,$$

for every $\mathbf{x} = (x_j)_{j\in\mathbb{N}}$, or, equivalently, the infinite Toeplitz matrix $\mathcal{T}$ associated with the moments $\{c_n\}_{n\in\mathbb{N}}$ is positive-definite. Under this assumption, the measure μ is unique.

The basic question is to describe a constructive method to get the measure μ. There are two approaches to this problem.

The first one is based on the fact that the absolutely continuous measure $d\mu = \frac{1}{|\phi_n(z)|^2}\frac{dz}{2\pi i z}$ supported on the unit circle induces in $\mathbb{P}_n$ (the linear space of polynomials with complex coefficients and degree less than or equal to n) the same inner product as μ. It is not so complicated to prove that μ_n converges to μ in the $\star$-weak topology [20].

The second one is related to quadrature formulas. Unfortunately, the Gaussian quadrature formulas for the real case cannot be considered on the unit circle because the zeros of orthogonal polynomials lie in the unit disk. Thus, it is not natural to recover the measure from mass points which do not live in the support of the measure. We proceed as follows:

Let $a_n(z) = \frac{z\phi_n(z)}{\phi_n^*(z)}$ a Blaschke product product with $n+1$ zeros in the unit disk. Given $w \in \mathbb{T}$, the equation $a_n(z) = a_n(w)$ has $n+1$ roots on $\mathbb{T}$. We will denote them $(\zeta_{n,j})_{j=0}^n$. Notice that w is a solution. We will order the roots according to the increasing arguments.

$$\zeta_{n,0} = w, \quad \arg\zeta_{n,j+1} \geq \arg\zeta_{n,j}, \quad j = 0, 1, \ldots, n.$$

On the other hand, if we consider the kernel polynomials $\{K_n(x,y)\}_{n\in\mathbb{N}}$ associated with μ

$$K_n(x,y) = \sum_{j=0}^{n} \frac{\phi_j(z)\phi_j(y)}{\|\phi_j\|^2},$$

then we get a Christoffel-Darboux formula

$$K_n(x,y) = \frac{1}{\|\phi_{n+1}\|^2} \frac{\phi^*_{n+1}(x)\overline{\phi^*_{n+1}(y)} - \phi_{n+1}(x)\overline{\phi_{n+1}(y)}}{1 - x\overline{y}}.$$

This formula is the polynomial counterpart of the Gohberg-Semencul algorithm for the inversion of Toeplitz matrices [16].

Notice that $K_n(\zeta_{n,j}, w) = 0$ for $j = 1, 2, \ldots, n$. Furthermore, the polynomial

$$B_{n+1}(z; w) = \phi_{n+1}(z) - \overline{\left(\frac{\phi^*_{n+1}(w)}{\phi_{n+1}(w)}\right)} \phi^*_{n+1}(z),$$

has as the set of zeros $\{\zeta_{n,j}\}_{j=0}^{n+1}$ and satisfies the orthogonality condition

$$\left\langle B_{n+1}(z; w), z^k \right\rangle = 0, \quad 1 \le k \le n.$$

They are called para-orthogonal polynomials.

Theorem 2 *The discrete measure $d\tilde{\mu}_n = \sum_{j=0}^{n}(\delta(z - \zeta_{n,j})K_n(\zeta_{n,j}, \zeta_{n,j}))^{-1}$ induces in $\mathbb{P}_n$ the same inner product as μ. Furthermore, $\tilde{\mu}_n$ converges to μ in the $\star$-weak topology.*

The quadrature formula associated to $d\tilde{\mu}_n$ is called a Szegő quadrature formula [34]. It has been extensively studied in the last decade both form the numerical point of view as well as from an analytical point of view. From this perspective, L. B. Golinskii [29] proved:

Theorem 3 *(i) The sets $(\zeta_{n-1,j})_{j=0}^{n-1}$ and $(\zeta_{n,j})_{j=1}^{n}$ interlace, i.e., between two consecutive points of one of them there is exactly one point of the other.*

(ii) If $\ln \mu' \in L^1(\mathbb{T})$, then

$$|\zeta_{n,k+1} - \zeta_{n,k}| \le \frac{C(\mu)}{\sqrt{n}},$$

for $k = 0, 1, \ldots, n$ and $\zeta_{n,n+1} = \zeta_{n,0}$ as a convention.

(iii) If $(\mu')^{-r} \in L^1(\mathbb{T})$ for some $r > 0$, then

$$|\zeta_{n,k+1} - \zeta_{n,k}| \le C(\mu, r)\frac{\ln n}{n}, \quad k \le n, \quad \zeta_{n,n+1} = \zeta_{n,0},$$

where $C(\mu)$ and $C(\mu, r)$ are universal constants.

(iv) If μ is absolutely continuous and $0 < A \le \mu' < B$, a.e. then

$$\frac{4}{n}\left(\frac{A}{B}\right)^{\frac{1}{2}} \le |\zeta_{n,k+1} - \zeta_{n,k}| \le \frac{4\pi}{n+1}\frac{B}{A}, \quad 1 \le k \le n, \quad \zeta_{n,n+1} = \zeta_{n,0}.$$

Thus, in terms of the properties of the measures we have upper estimates for the distance between two consecutive zeros of para-orthogonal polynomials.

Finally, a necessary and sufficient condition for the uniform distribution of zeros of para-orthogonal polynomials is given in the following:

Theorem 4 *For a fixed $w \in \mathbb{T}$, the sequence $(\zeta_{n,j})_{j=0}^{n}$ $n \in \mathbb{N}$ is uniformly distributed in $\mathbb{T}$ if and only if*

$$\frac{1}{n+1} K_n(e^{i\theta}, e^{i\theta}) d\mu \xrightarrow{\star} \frac{d\theta}{2\pi}.$$

2.2 Differential equations

From the perspective of differential operators, polynomials orthogonal on the unit circle behave very differently with respect to the real case. Here, classical orthogonal polynomials (Jacobi, Laguerre, Hermite and Bessel) can be analyzed as eigenfunctions of second order differential operators with polynomial coefficients. S. Bochner [14] described all the polynomial solutions of a second order differential equation

$$a_2 y''(x) + a_1(x) y'(x) + a_0 y(x) = \lambda_n y(x),$$

where $(a_k)_{k=0}^{2}$ are polynomials of degree at most k. H. L. Krall [40] obtained the polynomial solutions of a fourth-order differential equation

$$\sum_{k=0}^{4} a_k(x) y^{(k)}(x) = \lambda_n y(x), \quad \deg(a_k)_{k=0}^{4} \leq k.$$

Among them three-families of polynomials orthogonal with respect to non absolutely continuous measures appear. He called them Legendre-type, Laguerre-type and Jacobi-type orthogonal polynomials appear. On the other hand, there is a characterization of classical orthogonal polynomials due to Sonin (and independently obtained by W. Hahn) in the sense that they are the families of orthogonal polynomials such that the sequence of their first derivatives constitutes a sequence of orthogonal polynomials.

During the last decade an interesting work was done in the study of differential properties of orthogonal polynomials on the unit circle. Unfortunately, as it was proved in [43], the analog of classical orthogonal polynomials in the Sonin-Hahn sense is reduced to the canonical family of polynomials $\{z^n\}_{n\in\mathbb{N}}$ which is orthogonal with respect to the Lebesgue measure supported on the unit circle. More recently, some works are done about the analysis of second order differential equations associated with polynomials orthogonal with respect to measures supported on the unit circle [1]. The goal was to give an electrostatic interpretation of their zeros. In [33] an absolutely continuous measure $d\mu = w d\theta$ is considered. Let w be positive and differentiable function in a neighborhood of the unit circle, evenmore, a function v is associated with w as follows: $w = \exp(-v)$. The function v is said to be an external field. If

$$\int_{\mathbb{T}} \frac{v'(z) - v'(y)}{z - y} y^n \frac{w(y)dy}{iy},$$

exists for every $n \in \mathbb{Z}$, then the corresponding monic orthogonal polynomials satisfy a differential relation

$$\phi'_n(z) = n(1 - |\phi_n(0)|^2)\phi_{n-1}(z) + M(z;n)\phi_n(z) + N(z;n)\phi_n^*(z), \tag{6}$$

where

$$\begin{aligned} M(z;n) &= i \int_{\mathbb{T}} \frac{v'(z) - v'(y)}{z - y} \frac{\phi_n(y)\overline{\phi_n(y)}}{\|\phi_n\|^2} w(y)dy \\ N(z;n) &= -i \int_{\mathbb{T}} \frac{v'(z) - v'(y)}{z - y} \frac{\phi_n(y)\overline{\phi_n^*(y)}}{\|\phi_n\|^2} w(y)dy. \end{aligned}$$

If we assume $\phi_n(0) \neq 0$, then taking into account the forward recurrence relation we can rewrite (6) in the form

$$\phi'_n(z) = -A(z;n)\phi_n(z) + B(z;n)\phi_{n-1}(z). \tag{7}$$

Next we can define first order linear differential operators $\mathcal{L}_{n,1}$ and $\mathcal{L}_{n,2}$ as follows

$$\begin{aligned} \mathcal{L}_{n,1} &= \frac{d}{dz} + A(z;n) \\ \mathcal{L}_{n,2} &= -\frac{d}{dz} + C(z;n), \end{aligned}$$

where

$$C(n;z) = -A(z;n) + \frac{\|\phi_{n-1}\|^2}{\|\phi_n\|^2} \left(\frac{\phi_n(0)}{\phi_{n+1}(0)} + \frac{1}{z} \right) B(z;n).$$

Thus, the operators $\mathcal{L}_{n,1}$ and $\mathcal{L}_{n,2}$ are lowering and raising operators, respectively, i.e.

$$\begin{aligned} \mathcal{L}_{n,1}\phi_n(z) &= B(z;n)\phi_{n-1}(z) \\ \mathcal{L}_{n,2}\phi_n(z) &= \frac{B(z;n)}{z} \frac{\phi_n(0)}{\phi_{n+1}(0)} \left(1 - |\phi_n(0)|^2\right)^{-1} \phi_{n+1}(z). \end{aligned}$$

From both relations we deduce a second order linear differential equation for the polynomials ϕ_n [18, 33]

$$D(z;n)\phi''_n(z) + E(z;n)\phi'_n(z) + F(z;n)\phi_n(z) = 0, \tag{8}$$

where $D(z;n)$, $E(z;n)$ and $F(z;n)$ can be explicitly given in terms of $A(z;n)$ and $B(z;n)$. In particular, we get

$$\phi''_n(z) - \left[v'(z) + (n-1)z^{-1} + \frac{B'(z;n)}{B(z;n)} \right] \phi'_n(z) + G(z;n)\phi_n(z) = 0.$$

Notice that if $z_{n,k}$ is a zero of ϕ_n, then

$$\phi''_n(z_{n,j}) = \left[v'(z_{n,j}) + (n-1)z_{n,j}^{-1} + \frac{B'(z_{n,j};n)}{B(z_{n,j};n)} \right] \phi'_n(z_{n,j}).$$

This yields

$$\frac{\phi_n''(z_{n,j})}{\phi_n'(z_{n,j})} = v'(z_{n,j}) + (n-1)z_{n,j}^{-1} + \frac{B'(z_{n,j};n)}{B(z_{n,j};n)}.$$

Taking into account that

$$\frac{\phi_n''(z_{n,j})}{\phi_n'(z_{n,j})} = 2\sum_{\substack{1\le k\le n\\ k\ne j}} \frac{1}{z_{n,j}-z_{n,k}},$$

we get

$$v'(z_{n,j}) + (n-1)z_{n,j}^{-1} + \frac{B'(z_{n,j};n)}{B(z_{n,j};n)} + 2\sum_{\substack{1\le k\le n\\ k\ne j}} \frac{1}{z_{n,k}-z_{n,j}} = 0.$$

The left hand side of the above expression is the derivative of the function

$$H(z) = v(z) + (n-1)\ln z + \ln B(z;n) + \ln \prod_{\substack{1\le k\le n\\ k\ne j}} (z-z_{n,k})^2,$$

evaluated at $z = z_{n,j}$.

One can construct a real function

$$T(y_1, y_2, \dots, y_n) = \left| \prod_{j=1}^{n} y_j^{(-n+1)} \left[\frac{\exp(-v(z_j))}{B(y_j,n)}\right] \prod_{1\le j<k\le n} (y_j - y_k)^2 \right|,$$

such that the zeros of ϕ_n are stationary points of this function. One can interpret this function as the total energy function for n unit charges in the unit disk interacting with a one-body confining potential $v(z) + \ln B(z;n)$, an attractive logarithmic potential with a charge $(n-1)$ at the origin, and repulsive logarithmic two-body potentials between pairs of charges. However, all the stationary points are saddle-points.

When v is a rational function, the external field is said to be semiclassical. In such a case, the functions $A(z;n)$ and $B(z;n)$ are rational functions.

Semiclassical orthogonal polynomials on the unit circle have attracted the interest of researchers during the last decade because they provide a constructive method of sequences of orthogonal polynomials [1, 2, 17, 27, 28, 32, 41].

Very few examples were known until such a moment in despite the development of an analytic theory for some very general families of measures supported on the unit circle. Among these families, we have [39]

1. The Szegő class of measures μ such that $\ln \mu' \in L^1(\mathbb{T})$ or, equivalently, $\phi_n(0) \in l_2$.

2. The Nevai class of measures μ such that $\phi_n(0) \to 0$. An element of such a class is $\mu' > 0$ a.e.

3. The Césaro-Nevai class of measures μ such that $\frac{1}{n+1}\sum_{k=0}^{n}|\phi_k(0)| \to 0$. In particular, the Rakhmanov measures μ defined by the condition $\left|\phi_n(e^{i\theta})\right|^2 d\mu \xrightarrow{\star} \frac{\theta}{2\pi}$ belongs to the Césaro-Nevai class.

As a first example of semiclassical external field consider the weight function [56]

$$w(z) = |z-1|^{2\alpha}\,|z+1|^{2\beta}\,\frac{dz}{iz}.$$

In such a case, the reflection parameters are

$$\phi_n(0) = \frac{\alpha + (-1)^n\beta}{n+\alpha+\beta},$$

and the corresponding polynomials orthogonal with respect to $w(z)$ satisfy a second order linear differential equation (8) with

$$\begin{aligned}
D(z;n) = {} & z(z^2-1)\left[(\alpha+(-1)^{n+1}\beta)z + (\alpha+(-1)^n\beta)\right]\\
E(z;n) = {} & \left[(\alpha+(-1)^{n+1}\beta)z + (\alpha+(-1)^n\beta)\right]\\
& \left[(\alpha+\beta+3-n)z^2 - 2(\beta-\alpha)z + (n-1) + \alpha+\beta\right]\\
& -z(z^2-1)(\alpha+(-1)^{n+1}\beta)\\
F(z;n) = {} & \left[(\alpha+(-1)^{n+1}\beta)z + (\alpha+(-1)^n\beta)\right]\left[-(\alpha+\beta+2)nz + (n-1)(\beta-\alpha)\right]\\
& +(\alpha+(-1)^{n+1}\beta)(nz^2 + (\beta-\alpha)z - (n+\alpha+\beta)).
\end{aligned}$$

These polynomials are related to the Jacobi polynomials via the projective mapping of the unit circle onto the interval $[-1,1]$, $z \mapsto \frac{1}{2}(z+z^{-1})$.

As a particular case ($\beta = 0$) we get the circular Jacobi orthogonal polynomials. They arise in a class fo random unitary matrix ensembles where the parameter α is related to the charge of an impurity fixed at $z=1$ in a system of unit charges located on the unit circle at the complex values given by the eigenvalues of this matrix ensemble. In such a situation the monic orthogonal polynomials are hypergeometric polynomials $\phi_n(z) = \left(\frac{\alpha}{\alpha+n}\right){}_2\mathrm{F}_1(-n,\alpha+1;-n+1-\alpha;z)$.

Other example of semiclassical orthogonal polynomials is related to the weight function

$$w(z) = \frac{1}{2\pi I_0(t)}\exp(\frac{t}{2}(z+z^{-1})),$$

where I_ν is the modified Bessel function. The corresponding system of orthogonal polynomials arises from studies of the lenght of longest increasing subsequences of random permutations and unitary matrix models.

Theorem 5 *[33, 60]*

(i) The reflection parameters $a_n(t)$ for the above system of orthogonal polynomials satisfy a discrete Painlevé II equation

$$a_{n+1} + \frac{2na_n}{t(1-a_n^2)} + a_{n-1} = 0,\ \textit{for}\ n \geq 1,\ a_0(t) = 1,\ a_1(t) = -\frac{I_1(t)}{I_0(t)}.$$

(ii) As an alternative to this algebraic equation we get

$$\begin{aligned} a_n'' &= \frac{1}{2}\left[\frac{1}{a_n+1}-\frac{1}{a_n-1}\right](a_n')^2-\frac{1}{t}a_n' \\ &-a_n(1-a_n^2)+\frac{n^2}{t^2}\frac{a_n}{1-a_n^2}, \end{aligned}$$

with the boundary conditions determined by the expansion

$$a_n(t)\sim\frac{(-\frac{1}{2}t)^n}{n!}\left[1+\left(\frac{n}{n+1}-\delta_{n,1}\right)\frac{1}{4}t^2+\mathcal{O}(t^4)\right],\ t\to 0,$$

for $n\geq 1$.

Finally, the modified Bessel orthogonal polynomials (ϕ_n) satisfy the differential equation

$$\frac{d\phi_n}{dt}=\frac{1}{2}\left[\frac{I_1(t)}{I_0(t)}+\frac{\phi_{n+1}(0)}{\phi_n(0)}\right]\phi_n(z)-\frac{1}{2}\left[1+\frac{\phi_{n+1}(0)}{\phi_n(0)}z\right](1-|\phi_n(0)|^2)\phi_{n-1}(z),$$

for $n\geq 1$, where we have considered the derivative with respect to the dynamical parameter t.

On the other hand, taking into account

$$\frac{v'(z)-v'(t)}{z-t}=-\frac{t}{2}\left(\frac{1}{zt^2}+\frac{1}{z^2t}\right),$$

the relation (7) becomes

$$\begin{aligned} \frac{d\phi_n(z)}{dz} &= (1-|\phi_n(0)|^2)\left[\left(n+\frac{t}{2z}+\frac{t}{2}\frac{\phi_{n-1}(0)}{\phi_n(0)}-\right.\right. \\ &\left.\left.-\frac{t}{2}(1-|\phi_{n+1}(0)|^2)\overline{\phi_{n+1}(0)}\phi_n(0)\right)\phi_{n-1}(z)-\frac{t}{2z}\frac{\phi_{n-1}(0)}{\phi_n(0)}\phi_n(z)\right]. \end{aligned}$$

2.3 An application to the prediction of time series

Suppose $\{x_n\}_{n\in\mathbb{N}}$ is a sequence of complex random variables. This sequence is a discrete-time stochastic process, usually referred to as a time series in various applications [16]. The simplest case of such a process contains independent and identically distributed random variables with mean $E(z_n)=0$ and finite variance $E(|z_n|^2)=\sigma^2$. Such a time series is said to be a white noise and has minimal prediction value: the knowledge of the past z_{n-k} $(k\in\mathbb{N})$ does not help in predicting the value z_n. In most applications the time series $\{x_n\}_{n\in\mathbb{Z}}$ contains random variables which are dependent. The knowledge of their dependence structure will be useful in the prediction of values from past observations. A special interest are the stationary time series for which

$$E(x_n)=0,\quad E(\overline{x_n}x_{n+k})=\operatorname{cov}(x_n,x_{n+k})=\gamma(k).$$

This means that the mean and the covariance are independent of the time n. In particular, the autocovariance $\gamma(k)$ depends only on the time lag k. The autocovariance of white noise is

$$\gamma(k) = \begin{cases} \sigma^2, & \text{if } k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

In such a case, if we only deal with the second order behavior of the time series then white noise can be generalized to uncorrelated random variables with zero mean and variance σ^2, respectively, hence we write $z_n \sim WN(0, \sigma^2)$.

For the second order behavior of time series it is quite useful to consider the generating function

$$G(z) = \sum_{k\in\mathbb{Z}} \gamma(k) z^k. \tag{9}$$

This formal Laurent series will contain most information of the time series. We will assume that this Laurent series converges in some annulus, i.e. there exists a nonnegative real number $r < 1$ such that the series converges absolutely for $r < |z| < r^{-1}$. In particular, the series converges on the unit circle and thus

$$G(e^{i\theta}) = \sum_{k\in\mathbb{Z}} \gamma(k) e^{ik\theta},$$

defines a function $S(\theta) = \frac{G(e^{i\theta})}{2\pi}$. Notice that $\gamma(k)$ can be written as the n-th Fourier coefficient of $S(\theta)$:

$$\gamma(k) = \int_0^{2\pi} S(\theta) e^{-ik\theta}.$$

The function $S(\theta)$ is said to be the spectral density of the time series.

A general linear process is a stationary time series of the form

$$x_n = \sum_{k\in\mathbb{Z}} \alpha_k z_{n-k}, \quad z_n \sim WN(0, \sigma^2).$$

It is straightforward to prove that $E(x_n) = 0$ and the autocovariance is $E(\overline{x_n} x_{n+k}) = \sigma^2 \sum_{k\in\mathbb{Z}} \alpha_j \overline{\alpha}_{j+k}$.

General linear processes are composed by means of a white noise but they have a certain covariance structure which makes possible to predict x_n using information of the time series at times different from n. One speaks of coloured noise in such a case. In practice, when one wants to predict the time series x_n at the present n, one only has observations x_{n-k} from the past $(k > 0)$.

A time series of the form $x_n = \sum_{k=0}^{\infty} \alpha_k z_{n-k}$ containing only the white noise sequence of the past observations is known as a causal linear process. They are the most relevant in practice.

If we denote $\alpha(z) = \sum_{k\in\mathbb{Z}} \alpha_k z^k$, the generating function of the parameters for the general linear process, then the generating function $G(z)$ introduced in (9) becomes

$$G(z) = \sigma^2 \overline{\alpha}(z^{-1}) \alpha(z).$$

In particular, the spectral density of a general linear process is

$$S(\theta) = \frac{\sigma^2}{2\pi} \left|\alpha(e^{i\theta})\right|^2,$$

which shows that the spectral density is a positive function on the unit circle. When α is a polynomial of degree m with $\alpha(0) = 1$, then the general linear process is causal and it is known a a moving average process $MA(m)$.

If $\alpha(z) = \frac{1}{\beta(z)}$, where β is a polynomial of degree ν with $\beta(0) = 1$ and the zeros of β lie outside the unit disk, then the process is again causal and is called an autoregressive process $AR(\nu)$. Such a process is given by

$$x_n + \sum_{k=1}^{\nu} \beta_k x_{n-k} = z_n, \quad z_n \sim WN(0, \sigma^2).$$

One of the main purposes of studying time series is to predict as good as possible the future from past observations. This means that we want to predict x_n by means of x_{n-k} $(k > 0)$. For practical purposes, we will use only the last N observations from the past and thus we would like to predict x_n by means of a function $f(x_{n-1}, x_{n-2}, \ldots, x_{n-N})$. Furthermore, we will restrict our analysis to linear predictors, i.e. linear functions of the N variables $x_{n-1}, x_{n-2}, \ldots, x_{n-N}$. Our predictor $\hat{x}_n(N)$ can be expressed by

$$\hat{x}_n(N) = -\sum_{j=1}^{N} a_{N,j} x_{n-j}.$$

The coefficients $a_{N,j}$ can be determined by the condition about the prediction error

$$E\left(\left|\hat{x}_n(N) - x_n\right|^2\right),$$

be minimal. Thus, we get the extremal problem

$$\min_{b_{N,0}=1} E\left(\left|\sum_{j=0}^{N} b_{N,j} x_{n-j}\right|^2\right). \tag{10}$$

This problem can be transformed into an extremal problem in $L^2(\mu)$ where $\mu' = S(\theta)$.

Consider the inner product in $L^2(\mu)$ such that

$$\int_0^{2\pi} e^{-ik\theta} S(\theta) d\theta = \gamma(k).$$

Thus, the mapping $x_n \mapsto z^n$ gives an isometry between the inner product space for span $\{x_n\}$, $n \in \mathbb{Z}$ and $\mathbb{P}$ with the $L^2(\mu)$-norm. The expression (10) is equivalent to minimize

$$\int_0^{2\pi} \left|q_N(e^{i\theta})\right|^2 S(\theta) d\theta,$$

with the constraint $q_N(0) = 1$. But taking into account some previous result, we get $q_N(z) = \phi_N^*(z)$ where $\phi_N(z)$ is the monic polynomial of degree N which is orthogonal on the unit circle with respect to the measure μ with $\mu'(\theta) = S(\theta)$. The minimum is $\|\phi_N\|^2$.

If μ belongs to the Szegő class we have a useful interpretation in terms of prediction of stationary time series. Indeed, if $\ln \mu' \in L^1(\mathbb{T})$, then $\|\phi_N\| \to \alpha < \infty$ an thus it is not possible to find a linear predictor with a variance smaller that $\alpha^{-1} > 0$ even by allowing the complete past.

If $\ln \mu' \notin L^1(\mathbb{T})$, then in this case the predictor $\hat{x}_n(N)$ will converge to the actual random variable x_n if the number of observations in the past N increases. Such processes are called deterministic since we can predict the values in the future exactly form the information of the past.

A more general linear prediction problem can be formulated as follows: Given an integer number ν, let M_i be an increasing sequence of nested finite sets of integers such that $\nu \notin M_i$ and consider the construction of the best linear predictors of x_ν by elements x_n where $n \in M_i$. Let $e(M_i)$ denote the prediction error

$$e(M_i) = \min E\left(\left|x_\nu - \sum_{n\in M_i} c_n x_n\right|^2\right).$$

This problem was first considered by Kolmogorov who proved that if the stochastic process $\{x_n\}_{n\in\mathbb{Z}}$ is real, $\nu = 0$ and $M_i = \{n \ : \ -i \leq n \leq i\}$ then

$$\lim_{i\to\infty} e(M_i) = \left(\frac{1}{\pi}\int_0^\pi \frac{d\theta}{S(\theta)}\right)^{-1}.$$

It follows that if $S(\theta)^{-1}$ is not integrable then the prediction error converges to zero.

The study of best linear predictors which special emphasis on the case when the reciprocal of the absolutely continuous part of the spectral function is not integrable, as well as the rate of convergence of $e(k_i)$ remains an open problem.

On the other hand, if $\{x_n\}$ and $\{y_n\}$ are two stationary discrete time stochastic processes with spectral densities $S(\theta)$ and $T(\theta)$, respectively, and we assume S and T are comparable in the sense that $S(\theta) = g(\theta)T(\theta)$, then a natural question is to analyze the ratio

$$\lim_{i\to\infty} \frac{e(x_n; M_i)}{e(y_n; M_i)},$$

and find it explicitly for several examples.

We have analyzed this problem and given partial answers in [44] when g is a trigonometric positive rational function.

3 Multiple orthogonal polynomials

In this section we present the basic notions, definitions, and notations related with Multiple Orthogonal Polynomials (MOP). Special attention will be paid to the type II multiple discrete orthogonal polynomials. Furthermore, we will sketch the recent progress of the subject for the past few years.

During the past fifteen years there has been increased interest in multiple orthogonal polynomials, particularly promoted by the Soviet mathematical school. There are several survey papers about this topic (see, e.g., [5, 6, 15, 57] and Chapter 4 in the Nikishin and Sorokin book [50]), where the connection with the Hermite-Padé approximants is shown.

Recently, in [7, 8, 9] multiple orthogonal polynomials with respect to discrete measures have been considered. Other efforts in this direction have been considered (see [35, 36, 37, 50, 51, 52, 53, 54]). A quite complete collection of them are surveyed in [58].

For multiple orthogonal polynomials we will need multi-indices consisting in a vector of dimension r of positive integers, for which we use the notation $\vec{n} = (n_1, n_2, \ldots, n_r) \in \mathbb{N}^r$.

When the orthogonality conditions are distributed over r real intervals $\Delta_1, \ldots, \Delta_r$ with respect to r different measures $\mu_1, \mu_2, \ldots, \mu_r$, in order to extend the standard orthogonal polynomials two different ways appear: The so-called type I and type II multiple orthogonal polynomials. Let us start with the last ones:

3.1 Type II multiple orthogonal polynomials

Definition 1 *A polynomial $q_{\vec{n}}(x)$ is said to be a multiple orthogonal polynomial of a multi-index $\vec{n}$ with respect to positive Borel measures*

$$\mu_1, \mu_2, \ldots, \mu_r \quad \textit{such that} \quad \textit{supp}\, \mu_i = \Delta_i \subset \mathbb{R}, \quad i = 1, 2, \ldots, r,$$

if it satisfies the following conditions:

$$\begin{array}{ll} (i) & \deg q_{\vec{n}} \le |\vec{n}| := n_1 + n_2 + \cdots + n_r, \\ (ii) & \displaystyle\int_{\Delta_i} q_{\vec{n}}(x) x^k d\mu_i(x) = 0, \quad k = 0, 1, \ldots, n_i - 1, \quad i = 1, 2, \ldots, r. \end{array} \tag{11}$$

For $r = 1$ multiple orthogonal polynomial becomes standard orthogonal polynomial.

The *existence* of $q_{\vec{n}}(x) = \sum_{k=0}^{|\vec{n}|} a_{k,\vec{n}} x^k$ is always guaranteed, because for its $|\vec{n}| + 1$ unknown coefficients the orthogonality conditions (11) give a system of $|\vec{n}|$ linear algebraic homogeneous equations, which always has a nontrivial solution. However, the matrix of coefficients for such a linear system can be singular. Therefore the *uniqueness* is in general not guaranteed. A simple *counterexample* is when the measures $\mu_1, \mu_2, \ldots, \mu_r$ are all them identical on the same interval. Hence, we need some extra conditions on the r vector of measures in order that the above multiple orthogonal polynomial is unique.

Remark 1 *If one deals with the non-Hermitian complex orthogonality with respect to a set of complex valued functions*

$$m_1(z) = \sum_{k=0}^{\infty} \frac{m_{1,k}}{z^{k+1}}, \ldots, m_r(z) = \sum_{k=0}^{\infty} \frac{m_{r,k}}{z^{k+1}}, \tag{12}$$

over the contours $\Gamma_i \subset \mathbb{C}$ $(i = 1, 2, \ldots, r)$ *then, we can generalize the notion of multiple orthogonal polynomials as follows.*

A multiple orthogonal polynomial $q_{\vec{n}}(z)$ with respect to complex weights (12) verifies

$$\begin{aligned} &(i) && \deg q_{\vec{n}} \leq |\vec{n}|, \\ &(ii) && \oint_{\Gamma_i} q_{\vec{n}}(z) z^k m_i(z) dz = 0, \quad k = 0, 1, \ldots, n_i - 1, \quad i = 1, 2, \ldots, r. \end{aligned}$$

Two different sequences of indices are usually considered. The so-called diagonal and step-line sequences, respectively (see [5, 36, 57]).

3.1.1 Type I multiple orthogonal polynomials

Definition 2 *A vector of polynomials* $(v_{\vec{n},1} v_{\vec{n},2}, \ldots, v_{\vec{n},r})$ *is said to be a multiple orthogonal polynomial vector of type I if each polynomial* $v_{\vec{n},i}$*, where* $i = 1, 2, \ldots, r$*, satisfies the conditions*

$$\begin{gathered} \deg v_{\vec{n},i} \leq n_i - 1, \\ \sum_{i=1}^{r} \int_{\Delta_i} v_{\vec{n},i}(x) x^k d\mu_i(x) = 0, \quad k = 0, 1, \ldots, |\vec{n}| - 2, \quad i = 1, 2, \ldots, r. \end{gathered} \tag{13}$$

When $r = 1$ one recovers the standard orthogonal polynomials.

Again the *existence* of all the polynomials $v_{\vec{n},i}$ $(i = 1, 2, \ldots, r)$ determined by Definition 2 is guaranteed, because there are $|\vec{n}| - 1$ orthogonality conditions which give $|\vec{n}| - 1$ linear algebraic homogeneous equations for the $|\vec{n}|$ unknown coefficients. Therefore the type I multiple orthogonal polynomial vector is determined up to a constant factor.

3.1.2 Connection with Hermite-Padé simultaneous rational approximants

Multiple orthogonal polynomials are intimately related to simultaneous Padé approximation, which is often known as Hermite-Padé approximation.

Let $\Delta_i = (a_i, b_i)$, $i = 1, 2, \ldots, r$, be intervals on the real line, and $\mu_1, \mu_2, \ldots, \mu_r$ be Borel measures on $\mathbb{R}$ with infinitely many points of increase such that $\operatorname{supp} \mu_i \subset \Delta_i$, $i = 1, 2, \ldots, r$. The Markov functions (or Stieltjes functions)

$$m_i(z) = \int_{\Delta_i} \frac{d\mu_i}{z - x}, \quad z \notin \Delta_i \quad i = 1, 2, \ldots, r,$$

can be simultaneously approximated by rational functions with prescribed order near infinity. Two different ways are considered to study such a kind of problems.

The first one: For the multi-index $\vec{n} = (n_1, n_2, \ldots, n_r)$ of nonnegative integers it is well known [50] how to find a polynomial $q_{\vec{n}}(z) \not\equiv 0$ of degree at most $|\vec{n}|$, such that the expressions

$$q_{\vec{n}}(z)m_i(z) = p_{\vec{n},i}(z) + \frac{\zeta_i}{z^{n_i+1}} + \cdots, \quad i = 1, 2, \ldots, r, \tag{14}$$

hold, being $p_{\vec{n},i}(z)$ complex polynomials. Notice that $q_{\vec{n}}(z)$ always exists, because the relations (14) lead to a system of $|\vec{n}| + 1$ homogeneous linear equations. This approximation procedure, where one needs to find the polynomials $q_{\vec{n}}(z)$ and $p_{\vec{n},i}(z)$, is called type II Hermite-Padé approximation.

Through the rational function

$$\pi_i(\vec{n}, z) = \frac{p_{\vec{n},i}(z)}{q_{\vec{n}}(z)}, \quad i = 1, 2, \ldots, r. \tag{15}$$

we denote the simultaneous Hermite-Padé approximants of the r Markov (or Stieltjes) functions $m_1(z), m_2(z), \ldots, m_r(z)$, being $q_{\vec{n}}(z)$ the common denominator of the simultaneous approximants. This polynomial is precisely the multiple orthogonal polynomial due to the connection between the relations (14) and (11).

Thus, type II Hermite-Padé approximation is a rational approximation of the functions $m_i(z)$ $(i = 1, 2, \ldots, r)$ with the same denominator.

In [49], are considered simultaneous Hermite-Padé approximants of several Markov (or Stieltjes) functions, as well as the connection of this kind of approximation with the construction of linear forms on Markov (or Stieltjes) functions with polynomial coefficients.

For a fixed index-vector $\vec{n}$, although the existence of $q_{\vec{n}}(z)$ is guaranteed, the uniqueness is not determined by equalities (11) up to a normalizing constant. Even the rational functions $\pi_1(\vec{n}, z), \pi_2(\vec{n}, z), \ldots, \pi_r(\vec{n}, z)$ are not constructed in a unique way, using the vector-index $\vec{n}$ and the measures $\mu_1, \mu_2, \ldots, \mu_r$. There are two cases where $q_{\vec{n}}(z)$ is unique determined up to a constant factor: The so-called Angelesco and Nikishin systems [50].

Chebyshev system of functions.

Definition 3 *A set of continuous real functions $\{f_k(x)\}_0^n$ in the interval Δ is said to be a Chebyshev system (or T-system) of order n, if the determinant*

$$V(x_0, \ldots, x_n) = \det \begin{pmatrix} f_0(x_0) & f_1(x_0) & \cdots & f_n(x_0) \\ f_0(x_1) & f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(x_n) & f_1(x_n) & \cdots & f_n(x_n) \end{pmatrix}, \tag{16}$$

does not vanish for arbitrary different values $x_0, \ldots, x_n$ of Δ.

$$\sum_{k=0}^{n} |\lambda_k| > 0, \quad \lambda_k \in \mathbb{R}. \tag{17}$$

Thus, the definition 3 is equivalent to the following statement [50]: A set of functions $\{f_k(x)\}_0^n$ constitutes a T-system of order n if every linear combination

$$F(x) = \sum_{k=0}^{n} \lambda_k f_k(x),$$

where $\{\lambda_k\}_{k=0}^n$ verify (17), has at most n zeros in Δ.

For type II multiple orthogonal polynomials there is an useful system of functions.

Definition 4 *An AT-system consists of r weights $\{\rho_k(x)\}_1^r$ supported on the same interval Δ such that*

$$\rho_1(x), x\rho_1(x), \ldots, x^{n_1-1}\rho_1(x), \ldots, \rho_r(x), x\rho_r(x), \ldots, x^{n_r-1}\rho_r(x), \tag{18}$$

is a T-system on Δ for every multi-index $\vec{n} = (n_1, \ldots, n_r)$.

In this contribution we will focus our attention on the set of weights (the extension to measures is straightforward) which is an AT-system. See [7] for more information about the normality of the system of weights).

3.2 Brief description on the classical discrete polynomials.

The classical discrete orthogonal polynomials are those named after Hahn, Meixner, Kravchuk and Charlier. There are several approaches to the study (or characterization) of these polynomials. The more standard one is based on the fact that these discrete orthogonal polynomials are special cases of the basic hypergeometric series [25]. Other usual approaches are the group-theoretical approach [59] and the difference-equation approach on a lattice with a constant mesh [47, 48]. In the present contribution all the classical discrete orthogonal polynomials are considered as special cases of the type-II Hermite-Padé polynomials.

The Hahn polynomials $h_n^{(\alpha_0,\alpha_1)}(x, N)$ are polynomials of degree n which are orthogonal to all lower degree polynomials with respect to the weight function $\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_1+1)/(\Gamma(x+1)\Gamma(N-x))$ on the set of points $x \in [0, N-1]$ (mass points), where $\alpha_0, \alpha_1 > -1$. This means that the orthogonality conditions

$$\sum_{x=0}^{N-1} h_n^{(\alpha_0,\alpha_1)}(x, N) \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_1+1)}{\Gamma(x+1)\Gamma(N-x)} x^k = 0, \quad k = 0, \ldots, n-1, \tag{19}$$

holds.

The Meixner polynomials $m_n^{(\gamma,\upsilon)}(x)$ (being $\gamma > 0$ and $0 < \upsilon < 1$) are orthogonal on the set of points $x \in [0, \infty)$ with respect to the *Pascal distribution* $\upsilon^x(\gamma)_x/\Gamma(x+1)$, where $(\gamma)_x := \Gamma(\gamma+x)/\Gamma(\gamma)$, so that,

$$\sum_{x=0}^{\infty} m_n^{(\gamma,\upsilon)}(x) \frac{\upsilon^x \Gamma(\gamma+x)}{\Gamma(x+1)} x^k = 0, \quad k = 0, \ldots, n-1. \tag{20}$$

The Kravchuk polynomials $k_n^{(p)}(x,N)$ are orthogonal on the set of points $x \in [0,N]$ with respect to the *binomial distribution* $N!p^x(1-p)^{N-x}/(\Gamma(x+1)\Gamma(N+1-x))$, where $p,(1-p)>0$. Therefore, the orthogonality conditions are

$$\sum_{x=0}^{N} k_n^{(p)}(x,N)\frac{N!p^x(1-p)^{N-x}}{\Gamma(x+1)\Gamma(N+1-x)}x^k = 0, \quad k=0,\dots,n-1. \tag{21}$$

Finally, the Charlier polynomials $c_n^{(a)}(x)$ are polynomials of degree n which are orthogonal with respect to the *Poisson distribution* $a^x/\Gamma(x+1)$ $(a>0)$ on the mass points $x \in [0,\infty)$. They satisfy the orthogonality conditions

$$\sum_{x=0}^{\infty} c_n^{(a)}(x)\frac{a^x}{\Gamma(x+1)}x^k = 0, \quad k=0,\dots,n-1. \tag{22}$$

The above four families of discrete orthogonal polynomials satisfy several properties which also allow to characterize them in several ways. Before proceed to comment these properties of classical discrete orthogonal polynomials, let define the forward and backward difference operators

$$\begin{aligned} \Delta y(x(s)) &:= \frac{\triangle}{\triangle x(s)}y(x(s)) = \frac{y(s+h)-y(s)}{x(s+h)-x(s)}, \\ \nabla y(x(s)) &:= \frac{\bigtriangledown}{\bigtriangledown x(s)}y(x(s)) = \Delta y(s-h), \end{aligned} \tag{23}$$

respectively, for a functions y in terms of an arbitrary partition $x(s)$ with mesh h. For more simplicity, let us choose $x(s)=s$ and $\triangle x(s)=h=1$. Such a situation is said to be canonical in the sense that, by a canonical variable we will use x instead of s. Thus, the formula

$$\bigtriangledown^n y(x) = \sum_{i=0}^{n} \frac{(-1)^i n!}{i!(n-i)!}y(x-i) = \sum_{i=0}^{n} \frac{(-n)_i}{i!}y(x-i), \tag{24}$$

holds. It can be proved easily by induction.

Other important property of the operator ∇ (and equivalently of Δ) is the formula of summation by parts

$$\sum_{x=a}^{b} y(x)\bigtriangledown z(x) = y(x)z(x)|_{a-1}^{b} - \sum_{x=a}^{b} z(x-1)\bigtriangledown y(x), \tag{25}$$

whose proof is straightforward taking into account

$$\bigtriangledown[y(x)z(x)] = y(x)\bigtriangledown z(x) + z(x-1)\bigtriangledown y(x). \tag{26}$$

One of the most useful properties of the classical discrete orthogonal polynomials is that they verify the hypergeometric-type difference equation

$$\sigma(x)\triangle\bigtriangledown y(x) + \tau(x)\triangle y(x) + \lambda_n y(x) = 0, \tag{27}$$

where σ and τ are polynomials independent of the degree n, with, $\deg\sigma \leq 2$, $\deg\tau = 1$, and λ_n is a constant depending on n.

This equation can be written in the self-adjoint form

$$\triangle[\sigma(x)\rho(x) \bigtriangledown y(x)] + \lambda_n \rho(x)y(x) = 0,$$

when the function $\rho(x)$ satisfies the *Pearson-type* equation

$$\triangle[\sigma(x)\rho(x)] = \tau(x)\rho(x). \tag{28}$$

As a simple consequence of the second order linear difference equation (27) we get their polynomial solutions satisfy a finite-difference analog of the Rodrigues formula, i.e.,

$$p_n(x) = \frac{c_n}{\rho(x)} \bigtriangledown^n \left[\rho(x+n)\prod_{k=1}^{n}\sigma(x+k)\right], \quad \bigtriangledown^n := \underbrace{\bigtriangledown\cdots\bigtriangledown}_{n\text{-times}}, \tag{29}$$

where c_n is a normalizing factor depending on n.

From (29) we can deduce the relation between $\triangle p_n(x)$ and the polynomials themselves. Hence, the first finite differences of the discrete orthogonal polynomials are again orthogonal polynomials of the same family, but with different parameters, i.e.,

$$\triangle p_n(x) = -\lambda_n \frac{c_n}{\tilde{c}_{n-1}} p_{n-1}(x),$$

where λ_n and c_n are the corresponding eigenvalue of (27) and normalizing factor of (29), respectively. The coefficient $\tilde{c}_{n-1}$ is the normalizing constant in the Rodrigues formula (29) for the polynomial $p_{n-1}(x)$ obtained by replacing $\rho(x)$ by $\sigma(x+1)\rho(x+1)$. Indeed

$$\begin{cases} \triangle h_n^{(\alpha_0,\alpha_1)}(x,N) = \dfrac{h_{n-1}^{(\alpha_0+1,\alpha_1+1)}(x,N-1)}{(n+\alpha_0+\alpha_1+1)^{-1}}, \quad c_n = \dfrac{(-1)^n}{n!}, \\ \triangle m^{(\gamma,\upsilon)}(x) = \dfrac{n(\upsilon-1)}{\upsilon} m_{n-1}^{(\gamma+1,\upsilon)}(x), \quad c_n = \upsilon^{-n}, \\ \triangle k_n^{(p)}(x,N) = k_{n-1}^{(p)}(x,N-1), \quad c_n = \dfrac{(p-1)^n}{n!}, \\ \triangle c_n^{(a)}(x) = -\dfrac{n}{a} c_{n-1}^{(a)}(x) \quad c_n = a^{-n}. \end{cases} \tag{30}$$

From the orthogonality conditions all these families verify the three-term recurrence relation

$$xp_n(x) = \frac{a_n}{a_{n+1}} p_{n+1}(x) + \left[\frac{b_n}{a_n} - \frac{b_{n+1}}{a_{n+1}}\right] p_n(x) + \frac{a_{n-1}||p_n||^2}{a_n||p_{n-1}||^2} p_{n-1}(x). \tag{31}$$

From the hypergeometric property, we get

$$\sigma(x) \bigtriangledown p_n(x) = \frac{\lambda_n}{n\tau_n'} \left[\tau_n(x)p_n(x) - \frac{c_n}{c_{n+1}} p_{n+1}(x)\right], \tag{32}$$

being $p_n(x) = a_n x^n + b_n x^{n-1} +$ *lower terms*, and $\tau_n(x) = \tau(x+n) + \sigma(x+n) - \sigma(x)$.

For the classical discrete orthogonal polynomials starting from any of the properties (27)-(32), or from one of the orthogonality conditions (19)-(22), can be deduced the other properties.

3.3 Discrete polynomials with simultaneous orthogonality

Now we will present five families of multiple discrete orthogonal polynomials which constitute an AT system (see [7]). Here we give their Rodrigues-type formulas [9].

3.3.1 Examples

Multiple Hahn polynomials. The multiple Hahn polynomials are orthogonal polynomials associated with an AT system consisting of Hahn weights on $[0, N-1]$. These polynomials verify simultaneous orthogonality conditions with respect to r measures over the same mass points belonging to the interval $[0, N-1]$. This system has different singularities at 0 and the same singularity at 1, when the set of mass points tends to infinity and one substitutes the variable $x \in [0, N-1]$ by $(N-1)s$. Let us discuss this affirmation in more detail. It is natural to expect that the Hahn polynomials $h_n^{(\alpha_0,\alpha_1)}(x, N)$, after the linear change of variable $x = (N-1)s$, which transforms the interval $[0, N-1]$ into $[0,1]$, will tend to the Jacobi polynomials $P_n^{(\alpha_0,\alpha_1)}(s)$ when N tends to infinity (i.e., when the mesh $h = \triangle s = 1/(N-1)$ in the new variable s tends to 0), and that the weight function $\rho(x)$ will tend, up to a constant factor, to the weight function $x^{\alpha_0}(1-x)^{\alpha_1}$, where α_0, $\alpha_1 > -1$ for the Jacobi polynomials, orthogonal on $[0,1]$.

More precisely, replacing x by $(N-1)s$ in the Hahn weight one has

$$\rho(x) = \frac{\Gamma(N - Ns + \alpha_1)\Gamma(Ns + 1 + \alpha_0)}{\Gamma(N - Ns)\Gamma(Ns + 1)}.$$

Using the well known relation [48]

$$\frac{\Gamma(z+a)}{\Gamma(z)} = z^a \left[1 + \mathcal{O}\left(\frac{1}{z^2}\right)\right], \quad |\arg z| \leq \pi - \delta, \quad \delta > 0, \tag{33}$$

or, equivalently

$$\lim_{z\to\infty} \frac{\Gamma(z+a)}{z^a \Gamma(z)} = 1, \tag{34}$$

one gets that $\rho(x)$ behaves as $N^{\alpha_0} N^{\alpha_1} s^{\alpha_0} (1-s)^{\alpha_1}$ when $N \to \infty$.

Let $\alpha_0 > -1$ and $\alpha_1, \ldots, \alpha_r$ be such that each $\alpha_i > -1$, $i = 1, 2, \ldots, r$, and $\alpha_i - \alpha_j \notin \mathbb{Z}$ whenever $i \neq j$. The function $\hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x, N)$ denotes the monic multiple Hahn polynomial of degree $|\vec{n}| < N-1$ for the multi-index $\vec{n} \in \mathbb{N}^r$ and

$\vec{\alpha} = (\alpha_1, \ldots, \alpha_r)$ that satisfies the orthogonality conditions

$$\sum_{x=0}^{N-1} \hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N) \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i+1)}{\Gamma(x+1)\Gamma(N-x)} x^k = 0, \\ k = 0, \ldots, n_i - 1, \quad i = 1, \ldots, r. \tag{35}$$

Proposition 6 *The following finite-difference analog of the Rodrigues formula*

$$\begin{aligned} \hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N) &= \frac{(-1)^{|\vec{n}|}}{\prod_{i=1}^{r} (|\vec{n} + n_i \vec{e}_i| + \alpha_0 + \alpha_i)_{n_i}} \\ &\times \frac{\Gamma(N-x)}{\Gamma(N+\alpha_0-x)} \mathcal{D} \frac{\Gamma(N+\alpha_0-x)}{\Gamma(N-|\vec{n}|-x)}, \end{aligned} \tag{36}$$

where

$$\mathcal{D} := \prod_{i=1}^{r} \mathcal{D}_{i,n_i}, \quad \mathcal{D}_{i,n_i} = \frac{\Gamma(x+1)}{\Gamma(x+\alpha_i+1)} \bigtriangledown^{n_i} \frac{\Gamma(x+\alpha_i+n_i+1)}{\Gamma(x+1)},$$

holds [9].

Remark 2 *The product of the r difference operators $\mathcal{D}_{i,n_i}$ can be taken in any order since these operators commute.*

From the orthogonality relations (35) we can write

$$\sum_{x=0}^{N-1} \hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N) \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i+1)}{\Gamma(x+1)\Gamma(N-x)} \bigtriangledown \pi_{k+1}(x+1) = 0, \\ k = 0, 1, \ldots, n_i, \quad i = 1, 2, \ldots, r,$$

where

$$\pi_k(x) = x(x-1)\cdots(x-k+1) = \frac{x!}{(x-k)!} = \frac{\bigtriangledown \pi_{k+1}(x+1)}{k+1}. \tag{37}$$

Hence, using the summation by parts (25), as well as the fact that $\pi_{k+1}(0) = 0$ and $\Gamma^{-1}(0) = 0$, one obtains

$$\sum_{x=0}^{N} \bigtriangledown \left[\hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N) \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i+1)}{\Gamma(x+1)\Gamma(N-x)} \right] \pi_{k+1}(x) = 0, \\ k = 0, 1, \ldots, n_i - 1, \quad i = 1, 2, \ldots, r.$$

Thus, we get the raising operators

$$\begin{aligned} &\bigtriangledown \left[\hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N) \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i+1)}{\Gamma(x+1)\Gamma(N-x)} \right] \\ &= c_{n,i} \frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i)}{\Gamma(x+1)\Gamma(N-x+1)} \hat{h}_{\vec{n}+\vec{e}_i}^{(\alpha_0-1,\vec{\alpha}-\vec{e}_i)}(x,N+1), \quad i = 1, 2, \ldots, r, \end{aligned} \tag{38}$$

where the constant factors $c_{n,i}$ are found comparing the coefficients of the power $|\vec{n}|+1$ of x on the two sides of (38), i.e.,

$$c_{i,n} = -(|\vec{n}| + \alpha_0 + \alpha_i).$$

If we apply the above operator $(l_i - 1)$ times, where $l_i \in \mathbb{N}$, $(l_i > 1)$ on the expression (38), we get

$$\begin{gathered}\bigtriangledown^{l_i}\left[\hat{h}_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N)\frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i+1)}{\Gamma(x+1)\Gamma(N-x)}\right] = (-1)^{l_i} \\ (|\vec{n}|+\alpha_0+\alpha_i)_{l_i}\frac{\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i)}{\Gamma(x+1)\Gamma(N-x+1)}\hat{h}_{\vec{n}+\vec{e}_i}^{(\alpha_0-l_i,\vec{\alpha}-\vec{e}_{l_i})}(x,N+l_i), \\ i=1,2,\dots,r.\end{gathered} \tag{39}$$

The multiplication by the ratios $\Gamma(x+1)\Gamma(N-x+1)/\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_i)$ and $\Gamma(N+\alpha_0-x)\Gamma(x+\alpha_j)/\Gamma(x+1)\Gamma(N-x+1)$ on both sides of expressions (39), being $(j=1,\dots,i-1,i+1\dots,r)$, and the successive application of $\bigtriangledown^{l_j}$ on both sides of the equalities leads to (36).

In an analogous way to the multiple Hahn polynomials we can consider the discrete orthogonal polynomials of simultaneous orthogonality over the mass points $x=0,1,2,\dots$, with respect to r different *Pascal distributions.* Here can be distinguished two different cases that we will show in more detail below.

Multiple Meixner polynomials (first kind)

The multiple Meixner polynomials of the first kind $\hat{m}_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x)$ are given by the weights $\upsilon^x\Gamma(x+\gamma_i)/\Gamma(x+1)\Gamma(\gamma_i)$, where $\upsilon\in(0,1)$ and $\gamma_i>0$, for $i=1,2,\dots,r$. The assumption $\gamma_i-\gamma_j\notin\mathbb{Z}$ guarantees the AT property of the system of Meixner weights (or Pascal distributions).

So, the orthogonality conditions which determine the polynomial $\hat{m}_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x)$ are

$$\sum_{x=0}^{\infty}\hat{m}_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x)x^k\frac{\upsilon^x\Gamma(\gamma_i+x)}{\Gamma(x+1)} = 0, \quad k=0,\dots,n_i-1, \quad i=1,\dots,r. \tag{40}$$

From (40) it is no so difficult to deduce the raising operators

$$\mathcal{D}_i\hat{m}_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x) = \frac{\upsilon-1}{\upsilon(\gamma_i-1)}m_{\vec{n}+\vec{e}_i}^{(\vec{\gamma}-\vec{e}_i,\upsilon)}(x), \quad \mathcal{D}_i := \frac{\Gamma(x+1)}{\upsilon^x(\gamma_i-1)_x}\bigtriangledown\frac{\upsilon^x(\gamma_i)_x}{\Gamma(x+1)}, \tag{41}$$

where $(a)_x := \Gamma(a+x)/\Gamma(a)$ is the Pochhammer symbol.

A repeated use of the raising operator (41) gives the Rodrigues-type formula

$$\hat{m}_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x) = \left(\frac{\upsilon}{\upsilon-1}\right)^{|\vec{n}|}\left[\prod_{i=1}^{r}\frac{(\gamma_i+n_i-1)!}{(\gamma_i-1)!}\right]\Gamma(x+1)\mathcal{D}\,\Gamma^{-1}(x+1), \tag{42}$$

where $\mathcal{D} := \prod_{i=1}^{r}\mathcal{D}_{i,n_i}$, and $\mathcal{D}_{i,n_i} := \upsilon^{-x}\left(\gamma_i\right)_x^{-1}\bigtriangledown^{n_i}\upsilon^x\left(\gamma_i+n_i\right)_x$.

Multiple Meixner polynomials (second kind)

The other family of multiple Meixner polynomials appears when the simultaneous orthogonality conditions are distributed over the same set of discrete points $\mathbb{N}_0$ with respect to the weights $\upsilon_i^x(\gamma)_x/x!$, $(i=1,\ldots,r)$. To avoid the system of measures will be identical system we assume $\upsilon_i \neq \upsilon_j$ whenever $i \neq j$. The restrictions $\gamma > 0$, and $\upsilon_i \in (0,1)$, $(i=1,\ldots,r)$ are essentially inherited from the classical case (20). Thus, the AT property is again guaranteed. Hence, the polynomial $\hat{m}_{\vec{n}}^{(\gamma,\vec{\upsilon})}(x)$ determined by the following orthogonality conditions

$$\sum_{x=0}^{\infty} \hat{m}_{\vec{n}}^{(\gamma,\vec{\upsilon})}(x) x^k \frac{\upsilon_i^x \Gamma(\gamma+x)}{\Gamma(x+1)} = 0, \quad k=0,\ldots,n_i-1, \quad i=1,\ldots,r, \tag{43}$$

is called multiple Meixner polynomial of the second kind.

An analogous procedure as that carried out for the multiple Meixner polynomial of the first kind allows to obtain a finite-difference analog of the Rodrigues formula

$$\begin{aligned} \hat{m}_{\vec{n}}^{(\gamma,\vec{\upsilon})}(x) &= (\gamma+|\vec{n}|-1)!^2 \left[\prod_{i=1}^{r}\left(\frac{\upsilon_i}{\upsilon_i-1}\right)^{n_i}\frac{1}{(\gamma+n_i-1)!}\right] \\ &\times\Gamma(x+1)\mathcal{D}\,\Gamma^{-1}(x+1), \end{aligned} \tag{44}$$

where $\mathcal{D} := \prod_{i=1}^{r}\mathcal{D}_{i,n_i}$, and $\mathcal{D}_{i,n_i} = \upsilon_i^{-x}\left(\gamma+|\vec{n}|-n_i\right)_x^{-1}\nabla^{n_i}\,\upsilon_i^x\left(\gamma+|\vec{n}|\right)_x$.

Multiple Kravchuk polynomials

The multiple Kravchuk polynomials are orthogonal polynomials of degree $|\vec{n}| < N$, associated with an AT system of Kravchuk weights (*binomial distributions).* The function $\hat{k}_{\vec{n}}^{(\vec{p})}(x,N)$ denotes the monic multiple Kravchuk polynomials for which the r orthogonality conditions are given on the same set of finite number of points $x=1,2\ldots,N$ with respect to different binomial distributions. Thus, the orthogonality conditions become

$$\begin{gathered} \sum_{x=0}^{N} \hat{k}_{\vec{n}}^{(\vec{p})}(x,N)x^k \frac{N!p_i^x(1-p_i)^{N-x}}{\Gamma(x+1)\Gamma(N+1-x)} = 0, \quad 0<p_i<1, \\ k=0,1,\ldots,n_i-1, \quad i=1,2,\ldots,r. \end{gathered} \tag{45}$$

From (45), using (37) and (25) one obtains the raising operators

$$\begin{aligned} \mathcal{D}_i\hat{k}_{\vec{n}}^{(\vec{p})}(x,N) &= -\frac{1}{p_i(1-p_i)}\hat{k}_{\vec{n}+\vec{e}_i}^{(\vec{p})}(x,N+1), \quad i=1,2,\ldots,r \\ \mathcal{D}_i &:= \frac{N!p_i^x(1-p_i)^{N-x}}{\Gamma(x+1)\Gamma(N+1-x)}\,\nabla\,\frac{(N+1)!p_i^x(1-p_i)^{N+1-x}}{\Gamma(x+1)\Gamma(N+2-x)}. \end{aligned} \tag{46}$$

An appropriate combination of the raising operators (46) leads to the Rodrigues-type formula

$$\hat{k}_{\vec{n}}^{(\vec{p})}(x,N) = (-1)^{|\vec{n}|}\left[\prod_{i=1}^{r} p_i^{n_i}\right]\Gamma(x+1)\mathcal{D}\,\Gamma^{-1}(x+1), \tag{47}$$

where $\mathcal{D} := \prod_{i=1}^{r}\mathcal{D}_{i,n_i}$, and $\mathcal{D}_{i,n_i} = \frac{\Gamma^{-1}(N+1)p_i^{-x}(1-p_i)^x}{\Gamma^{-1}(N+1-x)}\nabla^{n_i}\frac{\Gamma(N-n_i+1)p_i^x(1-p_i)^{-x}}{\Gamma(N-n_i+1-x)}$.

Multiple Charlier polynomials

Finally, the Poisson discrete measures

$$\mu_i = \sum_{x=0}^{\infty}\frac{a_i^x}{\Gamma(x+1)}, \quad a_i > 0, \quad i = 1,2,\dots,r,$$

determine the associated system of simultaneous orthogonal polynomials $\hat{c}_{\vec{n}}^{(\vec{a})}(x)$, with $\vec{a} = (a_1, a_2, \dots, a_r)$ such that $a_i \neq a_j$, under the orthogonality conditions

$$\sum_{x=0}^{\infty}\hat{c}_{\vec{n}}^{(\vec{a})}(x)\frac{a_i}{\Gamma(x+1)}x^k = 0, \quad k = 0,1,\dots,n_i-1, \quad i = 1,2,\dots,r. \tag{48}$$

Thus, $\hat{c}_{\vec{n}}^{(\vec{a})}(x)$ is said to be the multiple Charlier polynomial (see also [10]).

In the same sense as we have proceeded in the above cases, the raising operators

$$\nabla\left[\hat{c}_{\vec{n}}^{(\vec{a})}(x)\frac{a_i^x}{\Gamma(x+1)}\right] = -\frac{1}{a_i}\frac{a_i^x}{\Gamma(x+1)}\hat{c}_{\vec{n}+\vec{e}_i}^{(\vec{a})}(x), \quad i = 1,2,\dots,r, \tag{49}$$

can be obtained from (48).

A repeated use of the raising operators (49) gives the Rodrigues-type formula

$$\hat{c}_{\vec{n}}^{(\vec{a})}(x) = (-1)^{|\vec{n}|}a_1^{n_1}a_2^{n_2}\cdots a_r^{n_r}\Gamma(x+1)\underbrace{\left[\prod_{i=1}^{r} a_i^{-x}\,\nabla^{n_i}\,a_i^x\right]}_{\mathcal{D}}\Gamma^{-1}(x+1). \tag{50}$$

3.4 Recurrence relation for multiple discrete orthogonal polynomials

Let $\mathcal{D}_{n_i}$, where n_i is the i-th coordinate of the vector index $\vec{n}$, be a difference operator defined by

$$\mathcal{D}_{n_i} := g_{i,1}(x;\alpha_1,\dots,\alpha_n)\,\nabla^{n_i}\,g_{i,2}(x;\beta_1,\dots,\beta_n), \quad \alpha_k,\beta_k \in \mathbb{R}, \quad k = 1,\dots,n,$$

being $g_{i,1}$ and $g_{i,2}$ certain functions depending, in general, on the i-th orthogonality measure.

The Rodrigues-type formulas allow us to introduce the following notation for the MDOP

$$q_{\vec{n}}(x) = c_{\vec{n},r} f_1(x) \mathcal{D}_{\vec{n}} f_2(x), \quad \mathcal{D}_{\vec{n}} := \prod_{i=1}^{r} \mathcal{D}_{n_i}, \tag{51}$$

where the coefficient $c_{\vec{n},r}$ depends on the parameters of the measures $\mu_1, \ldots, \mu_r$, and $f_1(x)$, $f_2(x)$ can also depend, in general, on $|\vec{n}|$ and $\mu_1, \ldots, \mu_r$.

Lemma 7 *Let $n_k \in \mathbb{N}$. Then, the product of n_k backward difference operators over the function $xf(x)$ can be written as*

$$\bigtriangledown^{n_k} xf(x) = n_k \bigtriangledown^{n_k-1} f(x) + (x - n_k) \bigtriangledown^{n_k} f(x), \tag{52}$$

where $\bigtriangledown^{n_k} := \underbrace{\bigtriangledown \cdots \bigtriangledown}_{n_k \text{times}}$.

From (24) and (26), the proof is straightforward using induction, i.e.,

$$\begin{aligned} \bigtriangledown^{n_k} xf(x) &= \bigtriangledown^{n_k-1} [\bigtriangledown xf(x)] = \bigtriangledown^{n_k-1} f(x) + \bigtriangledown^{n_k-1} [(x-1) \bigtriangledown f(x)] \\ &= 2 \bigtriangledown^{n_k-1} f(x) + \bigtriangledown^{n_k-2} \left[(x-2) \bigtriangledown^2 f(x)\right] = \cdots = \\ &= n_k \bigtriangledown^{n_k-1} f(x) + (x - n_k) \bigtriangledown^{n_k} f(x). \end{aligned}$$

Corollary 8 *The relation*

$$\mathcal{D}_{n_i} xf(x) = n_i \mathcal{D}_{n_i-1} f(x) + (x - n_i) \mathcal{D}_{n_i} f(x),$$

holds.

Lemma 9 *Let $\mathcal{D}_{n_i}$, where n_i is the i-th coordinate of the vector-index $\vec{n}$, be a difference operator defined in* (51). *Then,*

$$\mathcal{D}_{\vec{n}-\vec{e}} xf(x) = \left[\sum_{i=1}^{r} (n_i - 1) \prod_{j=1}^{r} \mathcal{D}_{n_j - \delta_{j,i} - 1} + (x - |\vec{n}| + r) \mathcal{D}_{\vec{n}-\vec{e}} \right] f(x), \tag{53}$$

where $\mathcal{D}_{\vec{n}-\vec{e}} := \mathcal{D}_{n_1-1} \cdots \mathcal{D}_{n_r-1} = \prod_{j=1}^{r} \mathcal{D}_{n_j-1}$ and $\delta_{i,j}$ Kronecker's delta.

Applying r times Corollary 8 we get (53).

If the condition

$$\bigtriangledown [g_{n_i,2}(x) f_2(x)] = g_{n_i,2}(x) [ax + b] f_2(x),$$

is verified by the set of functions $g_{n_i,2}(x)$ $(i = 1, \ldots, r)$, then the MDOP $q_{\vec{n}}(x)$ satisfies a $r + 2$ recurrence relation.

Thus, we have proved the following theorem, which generalizes the results presented in [36] for any vector index $\vec{n}$ (see theorem 13 below).

Theorem 10 *The multiple discrete orthogonal polynomials satisfy a $(r + 1)$-order recurrence relation (for every row and column in the Padé Table), where r is the number of orthogonality conditions.*

3.5 Application of MOP and open problems

This section deals with the application and open problems in which are involved the MOP and various fields of mathematics. Among them we can emphasize on number theory, special functions, and the spectral analysis of non-symmetric banded Hessenberg operators.

Number theory

Number theory is perhaps the more natural field of application of MOP. Indeed, the roots of these mathematical objects go back to the nineteenth century. More precisely, in 1878 Hermite used the Hermite-Padé approximants to prove the transcendence of the number "e" [31]. Traditionally, the Hermite method has been considered the main tool in order to investigate the arithmetic properties of real numbers. Related with the transcendence of π we remind the solution of the famous old problem about the quadrature of the circle given by Lendemann in 1882.

The multiple orthogonal polynomials seem to be an useful tool to prove the irrationality and transcendence of certain numbers.

Let us comment the problem about the arithmetic nature of the values of Riemann zeta-function

$$\zeta(s)=\sum_{n=1}^{\infty}\frac{1}{n^s},$$

at the odd positive integer numbers, because the first result in this direction has been obtained just in 1979 by Apéry.

The proof of Apéry is based on the following elementary lemma:

Lemma 11 *Let x be a real number, and p_n, q_n two sequences of integers ($n\in\mathbb{N}$). If p_n and q_n are such that*

i) $q_nx-p_n\neq 0$, *for all* $n\in\mathbb{N}$,

ii) $\lim\limits_{n\to\infty}(q_nx-p_n)=0$,

then x is irrational.

Theorem 12 (Apéry [4]) *$\zeta(3)$ is irrational.*

Apéry found the sequence of numbers

$$\begin{aligned} p_n &= \sum_{j=0}^{n}\binom{n}{j}^2\binom{n+j}{n}^2\left[\sum_{m=1}^{n}\frac{1}{m^3}+\sum_{m=1}^{j}\frac{(-1)^{m-1}}{2m^3\binom{n}{m}\binom{n+m}{m}}\right]\\ q_n &= \sum_{j=0}^{n}\binom{n}{j}^2\binom{n+j}{n}^2, \end{aligned} \tag{54}$$

which after some normalization give a sequence of integers $\bar{p}_n$ and $\bar{q}_n$ such that

$$0\neq r_n=\bar{p}_n-\bar{q}_n\zeta(3)\to 0,\quad n\to\infty.$$

Thus, using the previous Lemma 11 the statement holds.

The Beukers' contributions [12, 13] helped to understand where the sequences of integers (54) come from. On the other hand, the proofs given by Sorokin [53] and Van Assche [57] based on multiple orthogonal polynomials about the irrationaly of $\zeta(3)$ are very constructive and interesting by themselves.

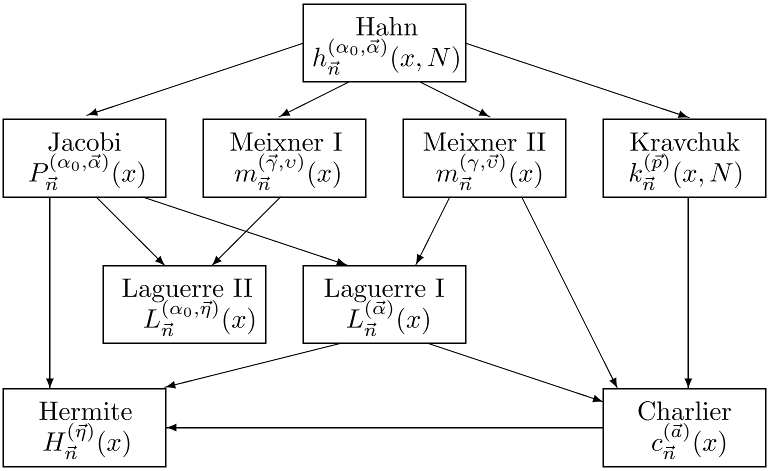
Concerning to the arithmetic nature of the values of Riemann zeta-function there is a challenging open problem which consists in proving the irrationality of $\zeta(5), \zeta(7), \ldots$, as well as the transcendency of $\zeta(3), \ldots$, etc.

Special functions and limit relations

The Rodrigues-type formula for the MOP suggests that the MOP could be expressed in terms of hypergeometric series. So, it would be a very good contribution from the point of view of special functions to clarify this question.

Other interesting problem is to classify all the classical multiple orthogonal polynomials and to establish the limit relations between them. It would be a nice approach to do this starting from the multiple Askey-Wilson polynomials (see [38] for classical cases) because all the other cases like multiple q-Hahn, q-Meixner, q-Kravchuk and q-Charlier can be obtained by limit transitions (see [8] for the q examples of MOP on the linear q lattice).

In [58] some families of classical MOP are deduced via the connection by limit transitions. In [9] the limit relations between discrete MOP and continuous MOP for AT systems are obtained (see the table below).


Hahn $h_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x,N)$
Jacobi $P_{\vec{n}}^{(\alpha_0,\vec{\alpha})}(x)$
Meixner I $m_{\vec{n}}^{(\vec{\gamma},\upsilon)}(x)$
Meixner II $m_{\vec{n}}^{(\gamma,\vec{\upsilon})}(x)$
Kravchuk $k_{\vec{n}}^{(\vec{p})}(x,N)$
Laguerre II $L_{\vec{n}}^{(\alpha_0,\vec{\eta})}(x)$
Laguerre I $L_{\vec{n}}^{(\vec{\alpha})}(x)$
Hermite $H_{\vec{n}}^{(\vec{\eta})}(x)$
Charlier $c_{\vec{n}}^{(\vec{a})}(x)$

Non-symmetric band operators

Here we will show the relationship between MOP and the spectral theory of non-symmetric operators. Let us start mentioning few classical results. Assuming that the higher order difference operator is represented by a band matrix, i.e.,

$$H = (a_{i,j})_{i,j=0}^{\infty},\ a_{i,j} = 0,\ (i > j + k,\ j > i + m),\ \text{and}\ a_{n,n-k} \neq 0,\ a_{n,n+m} \neq 0, \tag{55}$$

in the standard basis of the Hilbert space $l_2(\mathbb{N}_0)$

$$e_k = (\underbrace{0,\ldots,0}_{k-\text{times}},1,0,\ldots), \quad k \in \mathbb{N}_0,$$

one concludes that if H is a Jacobi matrix (i.e., symmetric tridiagonal matrix with real coefficients and positive extreme diagonals) then, the moment problem associated with H is determined. Hence by the Stone theorem, the class of closed operators

$$He_n = a_n e_{n-1} + b_n e_n + a_{n+1} e_{n+1}, \quad \text{where} \quad \left\{ \begin{array}{l} a_n = a_{n,n-1} \\ b_n = a_{n,n} \end{array} \right. ,$$

coincides with the class of simple spectrum self-adjoint operators (also known as Lebesgue operators). Therefore, by the Von Neumann spectral theorem there exists a unique operator valued measure $\mathcal{F}_t$, for which H admits the representation

$$H = \int_{\mathbb{R}} t\mathcal{F}_t.$$

The spectral measure for the operator H is the positive Borel measure

$$\mu(t) = \langle \mathcal{F}_t e_0, e_0 \rangle .$$

Thus, the Weyl function is

$$\mathcal{S}(z) = \langle \mathcal{R}_z e_0, e_0 \rangle , \tag{56}$$

where $\mathcal{R}_z$ is the resolvent operator defined as

$$\mathcal{R}_t = (zI - H)^{-1} = \int_{\mathbb{R}} \frac{d\mathcal{F}_t}{z-t}.$$

For the self-adjoint operator the Weyl function becomes the Markov (or Stieltjes) function

$$m(z) = \int_{\mathbb{R}} \frac{d\mu(t)}{z-t}. \tag{57}$$

The theory of orthogonal polynomial enjoys a very important result, known as Favard's theorem, which connects the spectral theory of self-adjoint operators and the theory of orthogonal polynomials. This theorem says that a sequence of polynomials $q_n(t)$ which verifies the recurrence relation (1) is always the orthogonal polynomial sequence with respect to the spectral measure

$$\int_{\mathbb{R}} q_n(t) t^k d\mu(t) = 0, \quad k = 0,\ldots,n-1. \tag{58}$$

Consecuently, the orthogonality (58) and the recurrence relation (1) are equivalent ways to describe orthogonal polynomials.

Remark 3 *The three-term recurrence relation* (1) *can be written as*

$$\underbrace{\begin{pmatrix} b_0 & a_1 & 0 & \cdots & 0 \\ a_1 & b_1 & a_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_{n-1} \\ 0 & \cdots & 0 & a_{n-1} & b_{n-1} \end{pmatrix}}_{H_n} \begin{pmatrix} q_0(t) \\ q_1(t) \\ \vdots \\ \vdots \\ q_{n-1}(t) \end{pmatrix} = t \begin{pmatrix} q_0(t) \\ q_1(t) \\ \vdots \\ \vdots \\ q_{n-1}(t) \end{pmatrix} - a_n q_n(t) \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Since the zeros of orthogonal polynomials are simple, one concludes that the eigenvalues of the Jacobi matrix H_n are the zeros of $q_n(t)$. Hence, the connection with the spectral theory of self-adjoint operators is clearly established when one considers the infinite matrix H (instead of H_n) acting as an operator $H : l_2 \to l_2$ on appropriate domains.

On the other hand, the rational function $\pi_n(z) = \frac{p_n(z)}{q_n(z)}$ (see (15) for the multiple case) being $p_n(z)$ the other linearly independent solution of the difference equation (1), i.e.,

$$\begin{gathered} t y_n(t) = a_{n+1} y_{n+1}(t) + b_n y_n(t) + a_n y_{n-1}(t), \quad n = 0, 1, \ldots, \\ p_1(t) = \frac{1}{a_1}, \quad p_{-1}(t) = 0, \end{gathered} \tag{59}$$

is the diagonal Padé approximant for the Markov (or Stieltjes) function

$$m(z) - \pi_n(z) = \frac{\zeta_n}{z^{2n+1}} + \ldots .$$

Despite the non-symmetric character of certain operators H,a proper choice of the rational approximants for the Weyl functions (56) (see also (57)) of the operator guarantees the connection with the entries of the matrix H.

In [36] an example of an operator H associated to non-symmetric $p+2$ diagonal matrix (55) is analyzed. This example shows how the three-term recurrence relation and the Jacobi matrix have a natural extension for multiple orthogonal polynomials.

Let $\vec{k}(n)$ ($n \in \mathbb{N}$) be a sequence of multi-indices such that $n = kr + j$, where $0 \le j < r$, and then set

$$\vec{k}(n) = (\underbrace{k+1, \ldots, k+1}_{j-times}, k, \ldots, k). \tag{60}$$

If all these indices are normal, then we have a weakly complete system. This condition is guaranteed if we consider the spectral problem for H i.e., if $q_n(z)$ and $p_n^{(j)}(z)$ ($j = 1, \ldots, r$) are the $r+1$ linearly independent solutions of $(r+1)$-order difference equation

$$\begin{aligned} z y_n &= a_{n,n-r} y_{n-r} + \cdots + a_{n,n} y_n + a_{n,n+1} y_{n+1}, \\ p_j^{(j)} &= \frac{1}{a_{j-1,j}}, \quad p_n^{(j)} = 0, \quad n < j, \quad j = 1, \ldots, r. \end{aligned}$$

Then, the connection between the Hermite-Padé approximants for the system of functions

$$m_j(z) = \langle \mathcal{R}_z e_{j-1}, e_0 \rangle \,, \quad j = 1, \ldots, r, \tag{61}$$

and the spectral problem is given in the following

Theorem 13 (Kalyagin [36]) *For $n = kr + j$ the vector of rational functions*

$$(\pi_1(\vec{n}, z), \ldots, (\pi_r(\vec{n}, z)) \,,$$

(see (15)*) is the Hermite-Padé approximant of index* (60) *for the system* (61)

Notice that in general for non-symmetric operators the notion of spectral positive measure has no sense. However, for non-symmetric operators the multiple orthogonal polynomials (Hermite-Padé polynomials) can be used, instead of the notion of standard orthogonal polynomials with respect to the positive spectral measure supported on the real line.

Acknowledgments

This work has been supported by Dirección General de Investigación (MCyT) of Spain under grant BFM2000-0206-C04-01 and INTAS project 2000-272. J. Arvesú was partially supported by the Dirección General de Investigación (Comunidad Autónoma de Madrid).

References

[1] M. Alfaro, F. Marcellán, *Recent trends in orthogonal polynomials on the unit circle*, in: Orthogonal Polynomials and their Applications, C. Brezinski et al. (eds.), IMACS Annals on Computational and Applied Mathematics Vol. 9. J. C. Baltzer, Basel, 1991, pp. 3–14.

[2] M. Alfaro, F. Marcellán, *Carathéodory functions on the unit circle*, Methods of Complex Analysis in Approximation Theory, A. Martínez-Finkelshtein et al. (eds.), Publicaciones Universidad de Almería, 1997, pp. 1–22.

[3] R. Álvarez-Nodarse, *Polinomios ortogonales: historia y aplicaciones.* Bol. Soc. Esp. Mat. Apl. 18 (2001) 19–45.

[4] R. Apéry, *Irrationalité de $\zeta(2)$ et $\zeta(3)$*, Astérisque 61 (1979), 11–13.

[5] A. I. Aptekarev, *Multiple orthogonal polynomials*, J. Comp. App. Math. 99 (1998), 423–447.

[6] A. I. Aptekarev, H. Stahl, *Asymptotics of Hermite-Padé polynomials*, in: A. Gonchar, E. B. Saff (Eds.), Progress in Approx. Theory, vol. 19, Springer Ser. Compt. Math. Springer, Berlin, 1992, pp. 127–167.

[7] J. Arvesú, *On the normality of the system of functions*, submitted.

[8] J. Arvesú, *q-Multiple orthogonality: Examples*, submitted.

[9] J. Arvesú, J. Coussement and W. Van Assche, *Discrete Multiple Orthogonal Polynomials*, J. Comp. Appl. Math. (accepted).

[10] J. Arvesú, W. Van Assche, *Discrete multiple orthogonal polynomials*, oral communication to the International INTAS Workshops on Constructive Complex Analysis, Leuven February 3–5, 1999.

[11] C. Bernardi, Y. Maday, *Approximations Spectrales de Problèmes aux Limites Elliptiques*, Springer Verlag, Paris 1991.

[12] F. Beukers, *A note on the irrationality of $\zeta(2)$ and $\zeta(3)$*, Bull. London Math. Soc. 11 (1979), 268–272.

[13] F. Beukers, *Padé approximations in number theory*, in Padé Approximation and its Applications, Lecture Notes in Math., vol. 888, Springer-Verlag, Berlin, 1981, pp. 90–99.

[14] S. Bochner, *Über Sturm-Liouvillesche Polynomsysteme*, Math. Z., 29 (1929), 730–736.

[15] M. G. Bruin, *Simultaneous Padé approximants and orthogonality*, Lecture Notes in Math., vol. 1171, Springer, Berlin, 1985, pp. 74–83.

[16] A. Bultheel, M. Van Barel, *Linear prediction: Mathematics and Engineering*, Bull. Belg. Math. Soc. Simon Stevin, 1 (1994), 1–58.

[17] M. J. Cantero, F. Marcellán, L. Moral, *A class of nonsymmetric orthogonal polynomials on the unit circle*, J. Approx. Theory 109 (2001) 30–47.

[18] Y. Chen, M. E. H. Ismail, *Ladder operators and differential equations for orthogonal polynomials*, J. Phys. A: Math. Gen. 30 (1997) 7818–7829.

[19] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach Science Publishers, New York, 1978.

[20] L. Daruis, P. González-Vera, F. Marcellán, *Gaussian quadrature formulas on the unit circle*, J. Comp. Appl. Math. 140 (2002) 159–183.

[21] C. Dunkl and Y. Xu, *Orthogonal polynomials of several variables*, Encyclopedia of Math., Cambridge University Press, 2001.

[22] A. J. Durán, *Ratio asymptotics for orthogonal matrix polynomials*, J. Approx. Theory 100 (1999) 304–344.

[23] A. J. Durán, E. Defez, *Orthogonal matrix polynomials and quadrature formulas*, Lin. Alg. Appl. 345 (2002) 71–84.

[24] D. Funaro, *Polynomial Approximation of Differential Equations*, Lecture Notes in Physics, Volume 8. Springer Verlag, Berlin 1992.

[25] G. Gasper, M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, 1990.

[26] W. Gautschi, F. Marcellán, L. Reichel (eds.), *Quadrature and Orthogonal Polynomials*, in: Numerical Analysis 2000, Volume 5. North Holland, Elsevier, Amsterdam 2001.

[27] E. Godoy, F. Marcellán, *An analog of the Christoffel formula for polynomial modification of a measure on the unit circle*, Boll. Un. Mat. Ital. (7) 5-A (1991) 1–12.

[28] E. Godoy, F. Marcellán, *Orthogonal polynomials and rational modification of measures*, Canad. J. Math. 45 (1993) 930–943.

[29] L. Golinskii, *Quadrature formula and zeros of para-orthogonal polynomials on the unit circle*, Preprint.

[30] U. Grenander, G. Szegő, *Toeplitz forms and their applications*, Chelsea, New York, 1984, Second Edition.

[31] C. Hermite, *Sur la généralisation des fractions continues algébriques*, Ann. Math. 2 (21) (1893) 289–308.

[32] M. E. H. Ismail, X. Li, *On sieved orthogonal polynomials IX. Orthogonality on the unit circle*, Pacific J. Math. 153 (1992) 289–297.

[33] M. E. H. Ismail, N. S. Witte, *Discriminants and functional equations for polynomials orthogonal on the unit circle*, J. Approx. Theory 110 (2001) 200–228.

[34] W. B. Jones, O. Njastad, W. J. Thron, *Moment theory, orthogonal polynomials, quadrature, and continued fractions associated with the unit circle*, Bull. London Math. Soc. 21 (1989) 113–152.

[35] V. A. Kalyagin, *On a class of polynomials defined by two orthogonality relations*, Mat. Sb., 110, (1979), 609-627; English transl. in Math. USSR Sb., 38 (1981), 563–580.

[36] V. A. Kalyagin, *The operator moment problem, vector continued fractions and an explicit form of the Favard theorem for vector orthogonal polynomials*, J. Comput. Appl. Math., 65 (1995) 181–193.

[37] V. A. Kaliaguine [Kalyagin], A. Ronveaux, *On a system of classical polynomials of simultaneous orthogonality*, J. Comput. Appl. Math., 67 (1996), 207–217.

[38] R. Koekoek, R. F. Swarttouw, *The Askey-scheme of hypergeometric orthogonal polynomials and its q-analogue*, Reports of the Faculty of Technical Mathematics and Informatics, 94-05, Delft University of Technology, Delft, 1994.

[39] S. V. Khrushchev, *Schur's Algorithm, orthogonal polynomials, and convergence of Wall's continued fractions in* $L^2(\mathbb{T})$, J. Approx. Th., 108 (2001), 161–248.

[40] H. L. Krall, *On orthogonal polynomials satisfying a certain fourth order differential equation*, The Pennsylvania State College Bulletin, 6 (1940), 1–24.

[41] A. P. Magnus, MAPA 3072A, *Special Topics in Approximation Theory: Semi-classical Orthogonal Polynomials on the Unit Circle*, Université Catholique de Louvain-La-Neuve 2000 (manuscript).

[42] F. Marcellán, M. Alfaro, M. L. Rezola, *Orthogonal polynomials on Sobolev spaces: Old and new directions*, J. Comput. Appl. Math. 48 (1993) 113–131.

[43] F. Marcellán, P. Maroni, *Orthogonal polynomials on the unit circle and their derivatives*, Constr. Approx. 7 (1991) 341–348.

[44] F. Marcellán, F. Peherstorfer, R. Steinbauer, *Orthogonal properties of linear combinations of orthogonal polynomials II*, Adv. Comp. Math., 7 (1997) 401–428.

[45] A. Martínez Finkelshtein, *Analytic aspects of Sobolev orthogonal polynomials revisited*, J. Comput. Appl. Math. 127 (2001) 255–266.

[46] H. G. Meijer, *A short history of orthogonal polynomials in a Sobolev space I. The non-discrete case*, Niew Archief voor Wiskunde 14 (1996) 93–112.

[47] A. F. Nikiforov, S. K. Suslov, and V. B. Uvarov, *Classical Orthogonal Polynomials of a Discrete Variable*, Springer Series in Computational Physics, Springer-Verlag, Berlin, 1991.

[48] A. F. Nikiforov, V. B. Uvarov, *Special Functions of Mathematical Physics*, Birkhäuser Verlag, Basel, 1988.

[49] E. M. Nikishin, *On the system of Markov functions*, Vestnik Moskv. Gos. Univ. Ser. I Mat. Meh., 4 (1979) 60–63: English transl. in Moscow Univ. Math. Bull., 34 (1979).

[50] E. M. Nikishin, V. N. Sorokin, *Rational Approximations and Orthogonality*, Translation of Mathematical Monographs, vol. 92, Amer. Math. Soc., Providence, RI, 1991.

[51] L. R. Piñeiro, *On simultaneous approximations for a collection of Markov functions*, Vestnik Moskov. Univ., Ser. I (1987), 67–70; English transl. in Moscow Univ. Math. Bull. 42 (1987), 52–55.

[52] V. N. Sorokin, *Simultaneous Padé approximants for finite and infinite intervals*, Izv. Vyssh. Uchebn. Zaved. Mat., 108 (1984), 45-52; English transl. in J. Soviet Math., 28 (1984), 56–64.

[53] V. N. Sorokin, *A generalization of classical orthogonal polynomials and the convergence of simultaneous Padé approximants*, Trudy Sem. Im. I. G. Petrovsk., 11 (1986), 125–165; English transl. in J. Soviet Math., 45 (1989), 1461–1499.

[54] V. N. Sorokin, *Simultaneous Padé approximations for functions of Stieltjes type*, Sibirsk. Mat. Zh., 31 (1990), 128–137; English transl. in Siber. Math. J., 31 (1990), 809–817.

[55] G. Szegő, *Orthogonal polynomials*, Coll. Publ. 23, Amer. Math. Soc., Providence RI 1975 (Fourth Edition).

[56] C. Tasis, *Propiedades diferenciales de los polinomios ortogonales relativos a la circunferencia unidad*, Tesis Doctoral, Universidad de Cantabria, 1989.

[57] W. Van Assche, *Multiple orthogonal polynomials, irrationality and transcendence*, in: B. C. Berndt et al. (Eds.), Contemporary Mathematics, vol. 236, Amer. Math. Soc., Providence, RI, 1999, pp. 325–342.

[58] W. Van Assche, Els Coussement, *Some classical multiple orthogonal polynomials*, J. Comp. Appl. Math. 127 (2001), 317-347.

[59] N. Ja. Vilenkin, A. U. Klimyk, *Representations of Lie Groups and Special Functions*, Kluwer Academic Publishers, Dordrecht, 1992.

[60] C. B. Wang, *Orthonormal polynomials on the unit circle and spatially discrete Painlevé II equation*, J. Phys. A: Math. Gen. 32 (1999) 7207–7217.

Bol. Soc. Esp. Mat. Apl. n°23(2003), 91–100

Matrices de Hadamard

T. Domínguez

Departamento de Análisis Matemático,
Facultad de Matemáticas, Universidad de Sevilla

tomasd@us.es

Resumen

Se definen las matrices de Hadamard y se indican algunos problemas referentes a su existencia, construcción y unicidad. Se mostrarán algunos ejemplos de aplicaciones de las matrices de Hadamard para resolver problemas de muy diferentes áreas de las matemáticas, concretamente: obtención de determinantes maximales, diseño de pesadas, detección de errores y corrección de códigos y finalmente problemas más modernos enmarcados en la Teoría Geométrica de los espacios de Banach.

Palabras clave: *Determinantes maximales, residuos cuadráticos, diseño de pesadas, estructura normal, corrección de códigos.*

Clasificación por materias AMS: *15A15, 11A07, 94B05, 46B20*

1 Introducción

El objeto de este artículo es hacer una recapitulación sobre ciertos resultados conocidos que muestran la interacción entre las diversas partes de las matemáticas. En concreto mostramos el interés que tiene un tipo especial de matrices, definidas por Hadamard en 1893 [4], para la resolución de problemas que surgen en áreas tan alejadas (aparentemente) como la teoría de códigos, diseños de pesadas o propiedades geométricas de los espacios de Banach. Al mismo tiempo y siguiendo la motivación que inspiró a Hadamard la definición de estas matrices, estudiamos el problema de buscar matrices con determinantes maximales. Por otra parte, la existencia de matrices de Hadamard de cualquier orden múltiplo de 4 es un problema aún abierto. Comentaremos este problema de existencia y mostraremos un método de construcción debido a Pailey, basado en el uso de residuos cuadráticos correspondientes a números primos.

Los resultados referidos a determinantes maximales pueden ser ampliados en [2]y [8] y los de teoría de códigos en [6].

Fecha de recepción: 20 de Mayo de 2002

2 Matrices de Hadamard: definición y ejemplos

Una matriz de Hadamard H de orden n es una matriz $n \times n$ formada por 1's y -1's tal que sus filas son ortogonales, esto es, $HH^t = nI$. Puesto que al multiplicar cualquier fila o columna por -1 en una matriz de Hadamard se obtiene otra matriz de Hadamard, podemos cambiar la primera fila y columna para que estén formadas sólo por 1's. Una matriz de este tipo se llama normalizada. Mostramos a continuación algunos ejemplos de matrices de Hadamard normalizadas:

$$H_1 = \left(1\right), \qquad H_2 = \left(\begin{array}{rr} 1 & 1 \\ 1 & -1 \end{array}\right), \qquad H_4 = \left(\begin{array}{rrrr} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{array}\right).$$

La primera cuestión que surge al considerar matrices de Hadamard es la siguiente: ¿existen matrices de Hadamard de cualquier orden? Veamos que si existe H_n y $n > 2$, entonces n tiene que ser múltiplo de 4. En efecto cambiando las columnas de orden podemos suponer que las tres primeras filas de H son como sigue (donde $-$ representa -1):

$$\begin{array}{cccc} 11\ldots1 & 11\ldots1 & 11\ldots1 & 11\ldots1 \\ 11\ldots1 & 11\ldots1 & --\ldots- & --\ldots- \\ \underbrace{11\ldots1}_{i} & \underbrace{--\ldots-}_{j} & \underbrace{11\ldots1}_{n/2-i} & \underbrace{--\ldots-}_{n/2-j} \end{array}$$

Como la segunda y la tercera fila son ortogonales, se tiene

$$i - j + n/2 - j - n/2 + i = 0,$$

$$2i - 2j = 0 \Longrightarrow i = j.$$

Así $n = 4i$ tiene que ser múltiplo de 4.

Ahora la pregunta sería: ¿existen matrices de Hadamard de cualquier orden que sea múltiplo de 4? Si repasamos los ejemplos anteriores vemos que cada uno procede del anterior mediante la construcción

$$H_{2n} = \left(\begin{array}{cc} H_n & H_n \\ H_n & -H_n \end{array}\right).$$

Así se obtienen matrices de orden $1, 2, 4, 8, \ldots, 2^n, \ldots$ que son llamadas matrices de Sylvester. ¿Qué sucede para otros órdenes, por ejemplo, 12? Veamos otro método de construcción, llamado de Payley. Para ello necesitamos considerar residuos cuadráticos.

3 Residuos cuadráticos. Construcción de Payley

Definición 1 *Sea p un número primo mayor que 2. Los números $1^2, 2^2, 3^2, \ldots, (p-1)^2$ módulo p son llamados residuos cuadráticos módulo p.*

Para encontrarlos basta considerar

$$1^2, 2^2, \ldots, \left(\frac{p-1}{2}\right)^2.$$

En efecto

$$(p-a)^2 = p^2 - 2ap + a^2 \equiv a^2 \pmod{p},$$

por lo cual $(p-a)^2$ y a^2 dan lugar al mismo residuo. Además son todos distintos. En efecto, si $i^2 \equiv j^2 \pmod{p}$, se tiene que p es divisor de $i^2 - j^2 = (i+j)(i-j)$. Como p es primo, tiene que dividir a $i-j$ o a $i+j$, lo cual es imposible si $i - j \neq 0$, pues $0 < i+j < p$. Por consiguiente hay $(p-1)/2$ residuos. Los restantes $(p-1)/2$ números $(\operatorname{mod} p)$ son llamados no residuos. El cero no es considerado. Por ejemplo para $p = 7$ tenemos $1^2 \equiv 1, 2^2 \equiv 4, 3^2 \equiv 2$, con lo cual los residuos son $1, 2, 4$ y los no residuos $3, 5, 6$.

A partir de los residuos podemos definir la función χ de Legendre definida sobre los enteros por:

$$\begin{aligned} \chi(i) &= 0, && \text{si } i \equiv 0 \pmod{p}, \\ \chi(i) &= 1, && \text{si } i \pmod{p} \text{ es residuo,} \\ \chi(i) &= -1, && \text{si } i \pmod{p} \text{ es no residuo.} \end{aligned}$$

Supongamos que p es un primo mayor que 2. Formamos la matriz de Jacobsthal $Q = (q_{ij})$ de orden p donde $q_{ij} = \chi(j - i)$. Por ejemplo, para $p = 7$ se obtiene

$$Q = \begin{pmatrix} 0 & 1 & 1 & - & 1 & - & - \\ - & 0 & 1 & 1 & - & 1 & - \\ - & - & 0 & 1 & 1 & - & 1 \\ 1 & - & - & 0 & 1 & 1 & - \\ - & 1 & - & - & 0 & 1 & 1 \\ 1 & - & 1 & - & - & 0 & 1 \\ 1 & 1 & - & 1 & - & - & 0 \end{pmatrix}.$$

Si ponemos ahora

$$\begin{pmatrix} 1 & \mathbf{1} \\ \mathbf{1} & Q - I \end{pmatrix},$$

se puede probar que se obtiene una matriz de Hadamard de orden $p + 1$. Todo esto es también cierto para cualquier potencia de p siendo p un número primo. Así podemos obtener matrices de orden $4, 8, 12, 20, 24, 28$. Este método de construcción fue dado por Pailey [7]. Por otros métodos de construcción (existen muchos diferentes) se han encontrado matrices de Hadamard de orden n si n es múltiplo de 4 hasta 264. No se conoce si existe una matriz de Hadamard de orden 268.

Otro problema interesante es el número de matrices de orden n que pueden ser construidas. Decimos que dos matrices son equivalentes si se puede pasar de una a otra por intercambio de filas o columnas, o multiplicando filas y columnas por -1. Se puede probar que sólo hay una clase de equivalencia para orden $1, 2, 4, 8, 12$. Sin embargo hay cinco clases de orden 16 y tres de orden 20. No se conoce el número de clases de equivalencia de orden 24.

4 Determinantes maximales

El origen de las matrices de Hadamard se debe al siguiente problema: sea $A = (a_{ij})$ una matriz de orden n tal que $|a_{ij}| \leq 1$. ¿Cuál es el valor máximo posible de $\det(A)$? Puesto que $\det(\lambda A) = \lambda^n \det(A)$, la restricción impuesta $\|A\|_\infty \leq 1$ es sólo una condición de normalización. La primera fácil observación es la siguiente: si $M = \text{máx}\det(A)$ con A como antes, existe A formada por 1's y -1's tal que $\det(A) = M$. En efecto, sacando cualquier a_{ij} como factor común resulta

$$\det(A) = a_{ij}c + d,$$

donde d es independiente de a_{ij}. Si $|a_{ij}| \neq 1$, reemplazando a_{ij} por 1 si $c \geq 0$, o por -1 si $c < 0$, se obtiene A_0 con $\det(A_0) \geq \det(A)$. Por otra parte, Hadamard probó que para matrices A con $\|A\|_\infty \leq 1$ se tiene $\det(A) \leq n^{n/2}$. Nótese que si A es una matriz de Hadamard, se tiene

$$AA^t = nI \Rightarrow (\det(A))^2 = n^n,$$

por lo que el determinante es maximal. Además Hadamard probó que $\det(A) = n^{n/2}$ si y sólo si A es una matriz de Hadamard. Por lo tanto, si no existe matriz de Hadamard de orden n, el máximo anterior es estrictamente menor que $n^{n/2}$. De hecho, no se conoce una fórmula general en este caso. Los primeros valores son

1	2	3	4	5	6	7	8	9	...
1	2	4	16	48	160	576	4096	14336	...

5 Teoría de Códigos

Una aplicación interesante de las matrices de Hadamard se refiere a la Teoría de Códigos. Esta es una rama de las Matemáticas e Ingeniería surgida a mediados del siglo pasado. Aunque tiene su origen en problemas de Ingeniería, el tema se ha desarrollado usando métodos matemáticos cada vez más sofisticados.

Los códigos se inventaron para corregir errores en canales de comunicación debidos al ruido. Supongamos que se envía un mensaje digitalizado (esto es, formado por 0's y 1's). Normalmente al enviar un 0 se recibe un 0, pero en algún caso (digamos con probabilidad 1/100) por problemas de ruido el 0 se recibe como un 1. Es importante para el receptor advertir la existencia del error y corregirlo. ¿Cómo puede hacerse esto? Podemos "alargar"el mensaje introduciendo signos de control. El ejemplo más usual para nosotros es el NIF con su letra de control al final de los números del DNI. Si se produce un error al introducir estos números (salvo que haya dos errores que se compensen, lo cual es altamente improbable) la letra de control lo marca. Los códigos más sencillos son los lineales. Veamos un ejemplo. Suponemos que estamos mandando una "palabra"digital de tres letras $u_1u_2u_3$ donde $u_i = 0$ ó 1. Tomamos una matriz de dimensión $n \times 3$ formada por 0's y 1's. Por ejemplo, para $n = 3$, consideramos

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Alargamos el código poniendo

$$(u_1 u_2 u_3) \to \begin{pmatrix} I_{3\times 3} \\ A \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}$$

en base 2, donde $x_i = u_i$ para $i = 1, 2, 3$. Así se obtiene la siguiente correspondencia entre los códigos que queremos enviar y los códigos alargados con tres cifras de control

$$\begin{array}{rcl} 000 & \rightarrow & 000000, \\ 100 & \rightarrow & 100011, \\ 010 & \rightarrow & 010101, \\ 001 & \rightarrow & 001110, \\ 110 & \rightarrow & 110110, \\ 101 & \rightarrow & 101101, \\ 011 & \rightarrow & 011011, \\ 111 & \rightarrow & 111000. \end{array}$$

Vemos que los números de control pueden coincidir, pero lo hacen para palabras que se diferencian en los tres dígitos. De esta forma, si suponemos que sólo se produce un error (o dos), la secuencia de control lo advertirá. Pero también es importante que en caso de error el receptor sea capaz de regenerar la palabra enviada. En efecto, la distancia mínima entre dos palabras de este código (esto es, la distancia de Hamming = número de cifras diferentes entre dos palabras) es 3. De esta manera, si hay un error, sólo una de las palabras correctas es posible. Los dos problemas —detección de errores y reconstrucción de la palabra correcta— son la clave de esta teoría. Elegir la matriz (de comprobación de la paridad) más apropiada en este caso o diseñar otros tipos de códigos es un problema de gran interés y dificultad. Para la reconstrucción de palabras es importante que la distancia de Hamming d sea lo mayor posible. Es fácil ver que un código con distancia mínima d puede corregir hasta $[(d-1)/2]$ errores. En el ejemplo anterior se puede corregir un error, pero no hay garantía para corregir dos errores.

En general, un (n, M, d) código es un conjunto de M vectores de dimensión n tal que dos vectores cualesquiera difieren en, al menos, d coordenadas. Lógicamente interesa que n sea lo más pequeño posible y en cambio M y d sean lo más grande posible. La siguiente desigualdad establece la cota mejor posible:

Teorema 1 (Cota de Plotkin) *Para cualquier* (n, M, d) *código* $\mathcal{C}$ *para el cual* $n < 2d$, *se tiene*

$$M \leq 2\left[\frac{d}{2d-n}\right].$$

Demostración . Para M par (la demostración es similar para M impar) vamos a calcular $S = \sum_{u,v\in\mathcal{C}} d(u,v)$ de dos formas diferentes. En primer lugar, como hay $M(M-1)$ parejas y la distancia entre ellas es al menos d, se tiene $S \geq M(M-1)d$. Por otra parte, pongamos las M palabras como filas de una matriz $M \times n$. Supongamos que la columna i contiene x_i 0's y $M - x_i$ 1's. Entonces esta columna contribuye $2x_i(M - x_i)$ a la suma. Por tanto $S = \sum_{i=1}^{n} 2x_i(M - x_i)$. Por cálculo elemental se prueba que el máximo de esta función de n variables se alcanza para $x_i = M/2$, obteniéndose de esta forma $S \leq nM^2/2$. Tenemos entonces

$$M(M-1)d \leq nM^2/2,$$

lo que implica

$$M \leq \frac{2d}{2d-n} \Rightarrow \frac{M}{2} \leq \frac{d}{2d-n} \Rightarrow \frac{M}{2} \leq \left[\frac{d}{2d-n}\right] \Rightarrow M \leq 2\left[\frac{d}{2d-n}\right].$$

Por ejemplo, si queremos con longitud $n = 6$ obtener $d = 4$, el número máximo de palabras es $2[4/2] = 4$. □

A la vista de la cota de Plotkin sería importante conseguir códigos en los que se alcance esta cota. Así para un n y d fijados se obtendría un código con el número máximo de palabras posibles. Para conseguir este código las desigualdades que hemos manejado se tienen que convertir en igualdades. Por tanto, todas las palabras tienen que estar a distancia d y en cada columna debe haber el mismo número de 0's que de 1's. Una herramienta fundamental para conseguirlo son las matrices de Hadamard.

Teorema 2 (Levenshteim [5]) *Si* n *y* d *son pares y existen matrices de Hadamard de orden* $4[d/(2d-n)]$ *y* $4[d/(2d-n)] + 4$, *las cotas de Plotkin son alcanzadas.*

Hay resultados similares para n y/o d impares. La demostración se basa en utilizar adecuadamente matrices de Hadamard, pegando trozos de ellas apropiadamente.

Mostramos como ejemplo un (12,4,8) código. Consideramos la matriz de Hadamard de orden 8 con los -1's convertidos en 1's y los 1's en 0's, esto es,

$$\begin{pmatrix} 0&0&0&0&0&0&0&0\\ 0&1&0&0&1&0&1&1\\ 0&1&1&0&0&1&0&1\\ 0&1&1&1&0&0&1&0\\ 0&0&1&1&1&0&0&1\\ 0&1&0&1&1&1&0&0\\ 0&0&1&0&1&1&1&0\\ 0&0&0&1&0&1&1&1 \end{pmatrix}.$$

Quitamos la primera columna y posteriormente las filas que comienzan por 1. Volvemos a quitar la primera columna y pegamos la matriz resultante consigo misma, obteniéndose el código $(12, 4, 8)$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix},$$

que puede corregir hasta tres errores.

6 Diseño de pesada

Pesando varios objetos juntos, en lugar de hacerlo separadamente, es posible determinar los pesos con mayor exactitud. Las técnicas para hacer esto se llaman diseños de pesada y las mejores están basadas en las matrices de Hadamard. Estas técnicas se pueden aplicar a diferentes problemas de medidas, no sólo pesos, sino también longitudes, voltajes, resistencias, concentraciones químicas, frecuencias, etc. De hecho se puede aplicar a cualquier experimento en los que la medida de varios objetos es la suma de las medidas individuales.

Supongamos que queremos pesar cuatro objetos con una balanza de dos platillos, la cual comete un error ϵ cada vez que se usa. Podemos suponer que ϵ es una variable aleatoria con media cero (puesto que debe estar equilibrada) y varianza σ^2. Si pesamos los cuatro objetos separadamente y los valores obtenidos son y_1, y_2, y_3, y_4 con errores desconocidos $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$, los verdaderos valores serían $a = y_1 + \epsilon_1$, $b = y_2 + \epsilon_2$, $c = y_3 + \epsilon_3$, $d = y_4 + \epsilon_4$. Por otra parte supongamos que hacemos cuatro pesadas de la siguiente forma

$$\left\{ \begin{array}{lcl} a+b+c+d & = & y_1 + \epsilon_1, \\ a-b+c-d & = & y_2 + \epsilon_2, \\ a+b-c-d & = & y_3 + \epsilon_3, \\ a-b-c+d & = & y_4 + \epsilon_4, \end{array} \right.$$

lo cual significa que los objetos con signo + son colocados a un lado de la balanza y los negativos al otro. Como la matriz de coeficientes es una matriz de Hadamard, es muy fácil resolver el sistema, obteniéndose

$$a = \frac{y_1 + y_2 + y_3 + y_4}{4} + \frac{\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4}{4},$$

o sea, el error es ahora

$$\frac{\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4}{4}.$$

Al hacer una media de los cuatro errores, éste se reduce, lo cual en términos estadísticos se expresa en función de la varianza de la siguiente forma: la varianza de $c\epsilon$ es c^2 por la varianza de ϵ y la varianza de la suma de variables aleatorias independientes es la suma de las varianzas. Así la varianza de

$$\frac{\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4}{4}$$

es $\sigma^2/4$. Lo mismo sucede para las demás incógnitas. En general, para n objetos, usando una matriz de Hadamard de orden n reducimos la varianza de σ^2 a σ^2/n. Se puede demostrar que esta es la varianza más pequeña que se puede obtener con un diseño de pesada de este tipo (elección de signos para los objetos a pesar).

7 Geometría de los espacios de Banach

Sea X un espacio de Banach y A un subconjunto convexo acotado de X. Llamamos radio de Chebyshev de A al número

$$r(A) = \inf\{\sup_{y\in A} \|x-y\|\,;\, x \in A\}.$$

Nótese que de forma menos rigurosa podríamos decir que $r(A)$ es "el radio de la bola más pequeña con centro en un punto de A que contiene a A". La constante de Jüng del espacio se define como

$$N(X) = \inf\left\{\frac{\operatorname{diam} A}{r(A)}\right\},$$

donde el ínfimo se toma sobre todos los conjuntos convexos acotados no unitarios de X. Por ejemplo, para el plano euclídeo $N(X) = \operatorname{diam} T/r(T)$, donde T es un triángulo equilátero y su valor es $\sqrt{3}$.

Aunque el coeficiente $N(\ell_2)$ fue calculado por Bynum en 1980, el valor de $N(\ell_p)$ y $N(L_p)$ fue un problema abierto durante diez años. En 1990 obtuvimos (véase [3]) el valor de estos coeficientes usando algunas desigualdades deducidas de teoría de la interpolación. La idea básica para encontrar este valor fue demostrar que bastaba considerar conjuntos que fueran la envolvente convexa de conjuntos finitos de puntos, todos equidistantes del centro de Chebyshev. Por otra parte, el Teorema de la Sección Esférica de Dvoretski nos dice que en cualquier espacio de Banach de dimensión infinita existen copias casi isométricas de un espacio euclídeo de cualquier dimensión finita. Este hecho, junto a las anteriores consideraciones, nos dice que el valor de $N(X)$ es mayor o igual que $\sqrt{2}$ para cualquier espacio Banach de dimensión infinita. Pero, ¿qué sucede para los espacios de dimensión finita? Por ejemplo para ℓ_p^n, esto es, $\mathbb{R}^n$ con la norma p.

Las técnicas usadas para calcular $N(\ell_p)$ pueden también ser aplicadas para obtener una cota superior de $N(\ell_p^n)$. El cálculo de $N(X)$ para espacios euclídeos (de dimensión finita) fue iniciado por Jüng en 1901, obteniendo $N(\ell_2^n)$. Recordando que sólo tenemos que considerar conjuntos finitos de X para calcular $N(X)$ y que en espacios n-dimensionales todo punto que está en la envolvente convexa de m puntos $x_1, \ldots, x_m$ está también en la envolvente convexa de un subconjunto formado por a lo más $n+1$ puntos, podemos concluir que sólo debemos considerar conjuntos finitos de X formados por a lo más $n+1$ puntos y que podemos suponer que estos puntos equidistan del centro de Chebyshev. Así el problema puede ser formulado equivalentemente de la siguiente forma: ¿cuál es el hiperpoliedro de $n+1$ vértices inscribible en la esfera unidad, conteniendo 0 en su interior, con diámetro mínimo? Estudiemos con más detalle este problema.

Supongamos que $A = \{x_1, \ldots, x_N\}$, $N \leq n+1$, es un conjunto en ℓ_p^n. Por traslación podemos suponer que el origen es el centro de Chebyshev, el cual está en la envolvente convexa de A, y por homotecia que $\|x_i\| = 1$, $i = 1, \ldots, N$. Aplicando ciertas desigualdades de convexidad (ver [1], Lemma II.3.8) obtenemos

$$\left(1 - \frac{1}{N}\right)^{\alpha-2} \operatorname{diam} A^{\alpha} \ (1 - \sum_{j=1}^{N} t_j^2) \geq 2,$$

donde $\alpha = p$ si $2 \leq p < +\infty$, y $\alpha = \frac{p}{p-1}$ si $1 < p \leq 2$. Usando el Teorema de los multiplicadores de Lagrange es fácil comprobar que $1 - \sum_{j=1}^{N} t_j^2$, con la ligadura $\sum_{j=1}^{N} t_j = 1$, alcanza un máximo si $t_j = 1/N$, $j = 1, \ldots, N$. Por consiguiente, tenemos

$$\left(1 - \frac{1}{N}\right)^{\alpha-2} \left(1 - \frac{1}{N}\right) \operatorname{diam} A^{\alpha} \geq 2,$$

esto es,

$$\frac{\operatorname{diam} A}{r(A)} \geq 2^{1/\alpha} \left(\frac{n+1}{n}\right)^{1-\frac{1}{\alpha}}.$$

Para $n = p = 2$ este es el valor exacto de $N(\ell_2^2) = \sqrt{3}$, puesto que es la razón entre el diámetro de un triángulo equilátero y el radio de la circunferencia circunscrita. Sucede igual para ℓ_2^n para todo n, pues en este caso

$$2^{1/\alpha} \left(\frac{n+1}{n}\right)^{1-\frac{1}{\alpha}} = \sqrt{\frac{2(n+1)}{n}}$$

es el diámetro del tetraedro (o hipertetraedro) inscrito en la hiperesfera unidad. ¿Qué sucede si $p \neq 2$? Desde luego la cota superior sólo puede ser alcanzada si todas las desigualdades se transforman en igualdades. Por tanto, todos los t_j tienen que ser iguales a $1/(n+1)$ y todas las distancias $\|x_j - x_k\|, j \neq k$, tienen que ser iguales a $\operatorname{diam} A$. Por tanto, el hiperpoliedro debe ser un hipertetraedro inscribible en la superficie esférica y cuyo centro geométrico sea el centro de la esfera. En el caso más simple, para $n = 2$, esto significa que la igualdad sólo puede ser alcanzada para un triángulo equilátero (en norma p) con vértices en la circunferencia unidad (otra vez debemos entender la circunferencia para la norma p) tal que el centro geométrico del triángulo es el origen de coordenadas. Por tanto la cuestión es ahora: ¿existe algún triángulo equilátero en ℓ_p^2 verificando esta condición? Si tal triángulo existe, ¿es $2^{2/\alpha-1}3^{1-1/\alpha}$ la longitud de su lado? Usando matrices de Hadamard veremos la respuesta en algunos casos especiales.

Supongamos que existe una matriz de Hadamard de orden $n+1$ y sea

$$\begin{pmatrix} 1 & v_1 \\ 1 & v_2 \\ \vdots & \vdots \\ 1 & v_{n+1} \end{pmatrix}$$

esta matriz, donde $v_1, \ldots, v_{n+1}$ son vectores en $\mathbb{R}^n$. Consideramos el conjunto $A = \{v_1, \ldots, v_{n+1}\}$ en ℓ_p^n para $p < 2$. Es claro que

$$\|v_i - v_j\| = 2\left(\frac{n+1}{2}\right)^{1/p}$$

si $i \neq j$, porque dos filas distintas de la matriz de Hadamard tienen $(n+1)/2$ 1's o -1's en la misma posición. Además $\|x_i\| = n^{1/p}$ para $i = 1, \ldots, n+1$. Entonces

$$\frac{\operatorname{diam} A}{r(A)} = 2^{1/q}\left(\frac{n+1}{n}\right)^{1/p},$$

que es el valor correspondiente a la cota superior. Por tanto en este caso

$$2^{1/q}\left(\frac{n+1}{n}\right)^{1/p}$$

es el valor exacto de $N(\ell_p^n)$. Es un problema abierto calcular $N(\ell_p^n)$ si $p > 2$ o si no existe matriz de Hadamard de orden $n+1$.

Referencias

[1] J. Ayerbe, T. Domínguez y G. López, *Measures of Noncompactness in Metric Fixed Point Theory*, Operator Theory Advances and Applications, Vol 99, Birkäuser, Berlin 1997.

[2] R. Bellman, *Introduction to Matrix Analysis*, McGraw Hill, New York 1970.

[3] T. Domínguez, *Normal structure coefficients in L_p-spaces*, Proc. Royal Soc. Edinburgh, 117A, pp. 299-303, 1991.

[4] J. Hadamard, *Résolution d'une question relative aux déterminants*, Bull. Sci. Math. 2, 240-248 (1893).

[5] V.I. Levenshtein, *The applications of Hadamard matrices to a problem in coding*, Problems of Cybernetic 5, 166-184 (1964).

[6] F.J. Macwilliams y N.J.A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam 1978.

[7] R.E.A.C. Pailey. *On orthogonal matrices*, J. Math. and Phys. 12, 311-320 (1933).

[8] J. Williamson, *Note on Hadamard's determinant Theorem*, Bull. Amer. Math. Soc. 53, 608-613 (1947).

Bol. Soc. Esp. Mat. Apl. n^o23(2003), 101–122

Simulación numérica de diferentes procesos industriales relacionados con la producción de silicio. Proyectos de colaboración con Ferroatlántica I+D.

A. Bermúdez

Departamento de Matemática Aplicada
Universidade de Santiago de Compostela

mabermud@usc.es

Resumen

El objetivo de este trabajo es la descripción de las líneas de investigación desarrolladas en el marco de colaboración que mantiene la empresa Ferroatlántica I+D con el Departamento de Matemática Aplicada de la Universidade de Santiago de Compostela. En el artículo, se aborda la simulación numérica de diferentes problemas físicos relacionados con la producción del silicio; para ello se describen los modelos matemáticos desarrollados, los métodos numéricos utilizados y los principales resultados obtenidos.

Palabras clave: *Simulación numérica, hornos eléctricos, colada de silicio, purificación de silicio, electromagnetismo, transferencia de calor, elementos finitos.*

1 Introducción

El Grupo Ferroatlántica está integrado por diferentes empresas dedicadas a la producción de ferroaleaciones. Su actividad le convierte en el primer grupo español del sector y el segundo de la Unión Europea. Además, es el primer productor español independiente de energía eléctrica.

Las actividades de investigación y desarrollo del Grupo Ferroatlántica son gestionadas por la empresa Ferroatlántica I+D, que juega un papel fundamental en el desarrollo y comercialización de nuevas tecnologías dentro del grupo. En los últimos años, Ferroatlántica I+D ha mantenido una estrecha colaboración con la Universidad y otros Centros Públicos de Investigación principalmente en Galicia. En concreto, Ferroatlántica I+D ha inicado proyectos de investigación con el Departamento de Matemática Aplicada de la Universidade de Santiago de Compostela en el año 1996. En este artículo expondremos los principales temas de investigación abordados y los resultados obtenidos. La actividad investigadora ha estado orientada al desarrollo de modelos matemáticos y programas de ordenador que permiten simular numéricamente diferentes procesos industriales de interés para la empresa. Así, destacan los siguientes campos:

1. Modelado de electrodos metalúrgicos.
2. Modelado de nuevos sistemas de colada para ferroaleaciones.
3. Modelado de sistemas innovadores para la purificación del silicio.

El equipo de investigación que ha participado en el desarrollo y resolución numérica de los diferentes modelos está integrado por A. Bermúdez, M.C. Muñiz, R. Leira, F. Pena y P. Salgado por parte del Departamento de Matemática Aplicada y por J. Bullón, M. Lage, A. Lorenzo por parte de Ferroatlántica I+D.

El desarrollo del artículo es el siguiente: la Sección 2 se ocupa de la descripción del proceso de producción de silicio y sus aplicaciones. En la Sección 3 se describe la problemática de los electrodos, piezas clave en un horno de producción de silicio, y se describen los modelos matemáticos desarrollados en este tema. En la Sección 4 se aborda el modelado matemático de un innovador sistema de colada sobre placa vibrante de cobre refrigerada por agua, para la solidificación del silicio. La Sección 5 se ocupa del modelado de sistemas relacionados con la purificación del silicio. Finalmente, en la Sección 6 se presenta un avance de la investigación en curso sobre la simulación numérica de un horno de inducción, también destinado a la purificación de silicio.

2 El silicio y sus aplicaciones

El silicio (Si) es el segundo elemento más abundante en la corteza terrestre después del oxígeno. Se obtiene industrialmente por reducción del dióxido de silicio, en forma de cuarzo o cuarcita, con carbón, mediante una reacción química que puede escribirse en forma simplificada (ver [16]):

$$\mathrm{Si\,O_2} + 2\,\mathrm{C} = \mathrm{Si} + 2\,\mathrm{CO}. \tag{1}$$

Este proceso tiene lugar en un horno eléctrico de reducción, más concretamente, en un “horno de arco”que se describe en la siguiente sección. El silicio tiene una gran variedad de aplicaciones dependiendo de la cantidad de impurezas presentes en el producto. Así, se distingue:

- **Silicio electrónico:** es el más puro de los empleados industrialmente, denominado 9N (“9 nueves”≡ 99,9999999 % de pureza). Se utiliza para la fabricación de los semiconductores, en la que se requiere una precisión en la conducción eléctrica muy alta.

- **Silicio solar:** precisa de menor pureza que el anterior, aunque en la actualidad comparte con el silicio electrónico varias fases del proceso de producción. Sirve para fabricar las células fotovoltaicas de los paneles solares.

- **Silicio metal:** es el silicio metalúrgico que contiene un 1-2 % de otros elementos. Se usa principalmente en aleaciones con otros metales no férreos, como el aluminio.

- **Silicio químico:** es el que se utiliza para la fabricación de siliconas, productos de gran consumo por su variedad de aplicaciones.

- **Ferrosicilio:** es el que contiene más de un 2 % de otros elementos, especialmente hierro (de ahí su nombre). Tiene menor precio que el silicio metal y se usa para producir aceros al silicio.

3 Modelado de electrodos metalúrgicos

Un horno eléctrico para la producción de silicio está compuesto, en general, por una cuba cilíndrica que contiene materiales carbonosos y tres electrodos cuyos ejes forman un triángulo equilátero centrado en la cuba (ver Figura 1). Los electrodos son las piezas clave del horno y su propósito es la conducción de la corriente eléctrica, normalmente corriente alterna y trifásica. En la punta de cada electrodo se genera un arco eléctrico, que crea las altas temperaturas necesarias para que tengan lugar las diferentes reacciones químicas del proceso de reducción. El buen funcionamiento del horno depende, en gran medida, de conseguir condiciones de operación adecuadas en los electrodos.

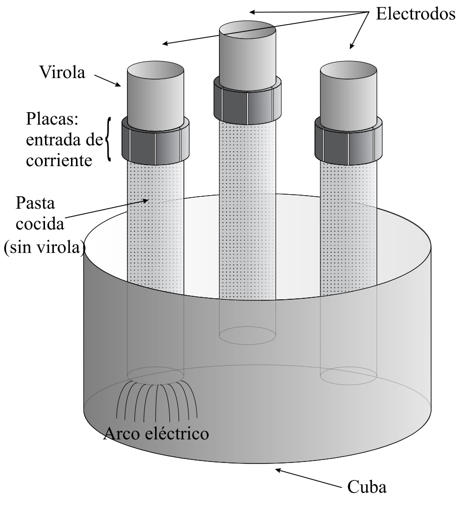


Figura 1: Cuba de un horno de reducción.

Los electrodos clásicos utilizados mayoritariamente en la industria son los electrodos de grafito puro, los precocidos y los Söderberg. Estos últimos, están compuestos de pasta carbonosa que se va cociendo durante el proceso y son los más empleados en la industria del ferrosilicio. Sus ventajas son que se construyen de tamaños mayores y que son más baratos que los de grafito puro o los precocidos. Sin embargo, los electrodos Söderberg no se pueden utilizar para la producción de silicio metal, ya que durante el proceso contaminan el producto. Así, hasta la pasada década, los electrodos precocidos eran la única alternativa para la producción de silicio metal.

A principios de los años 90, Ferroatlántica I+D desarrolló un nuevo tipo de electrodo, el llamado electrodo ELSA, que es apto para la producción de silicio metal. Si bien los electrodos clásicos están formados por un único material, el electrodo ELSA es un electrodo compuesto que está formado por un núcleo de grafito rodeado de pasta (ver Figura 2). El electrodo ELSA es mucho más barato que el electrodo precocido, por lo que se ha convertido en una alternativa muy interesante para la producción de silicio metal al reducir en más del 10 % los costes de producción.

En general, las condiciones de funcionamiento de cualquier tipo de electrodo son complejas, por lo que la simulación numérica constituye una herramienta muy importante para estudiar su comportamiento. Modelar el problema en un ordenador permite estudiar la influencia de los diferentes parámetros que intervienen en el funcionamiento del electrodo, y por lo tanto del horno, sin necesidad de experimentos delicados y costosos. Así, el comportamiento de los electrodos clásicos ha sido estudiado durante décadas (ver por ejemplo [12] y [10]). La mayoría de estos trabajos asumen simetría cilíndrica y utilizan métodos de diferencias y elementos finitos para la resolución de los modelos. Sin embargo, la estructura compuesta del electrodo ELSA hace que su comportamiento termoeléctrico sea diferente del de los electrodos clásicos. En particular, el electrodo ELSA combina el grafito, que es muy buen conductor de la electricidad, con la pasta, que solo es buena conductora a temperaturas altas. Por lo tanto, no sólo el grafito es importante en el movimiento de la columna, sino también en la distribución de la corriente. Así, el llamado *efecto piel* es comparable al de electrodos clásicos sólo en la punta. Por otra parte, la pasta cambia de estado durante el proceso, siendo su cocción un fenómeno fundamental; sin embargo, este proceso de cocción también presenta notables diferencias con respecto al que tiene lugar en los electrodos Söderberg (ver [15] para más detalles). Teniendo en cuenta estas diferencias, la simulación numérica del electrodo ELSA ha constituido el objetivo fundamental de muchos de los proyectos y contratos mantenidos con Ferroatlántica I+D. Cabe señalar sin embargo, que los modelos desarrollados son lo suficientemente generales para simular el comportamiento de cualquier tipo de electrodo.

La naturaleza de los modelos desarrollados para estudiar el comportamiento del electrodo ELSA es muy diversa y ha dado lugar al estudio numérico y matemático de un amplio número de problemas. Los resultados más relevantes se recogen en diferentes publicaciones ([1, 2, 7, 5, 6]). Por ello, en esta sección daremos una descripción bastante general de los principales problemas abordados y de los resultados obtenidos.

La simulación del electrodo ELSA ha tenido dos objetivos fundamentales: conocer y controlar los diferentes parámetros que intervienen en la cocción de la pasta y conocer los factores que pueden provocar roturas en el electrodo. Así, los modelos matemáticos desarrollados tienen como objeto el cálculo de la distribución de la temperatura, de la densidad de corriente y de las tensiones en el electrodo, bajo diferentes condiciones de operación.

El primer paso ha sido el desarrollo de modelos matemáticos basados en la simetría cilíndrica; es decir, despreciando el *efecto proximidad* de los otros dos electrodos y considerando condiciones de contorno axisimétricas. En estas condiciones se puede suponer que los diferentes campos no dependen de la variable angular. Estas hipótesis permiten resolver el problema en un dominio bidimimensional, en concreto, en una sección radial del electrodo (ver Figura 2). Así, se han desarrollado y resuelto los siguientes modelos:

- Un modelo **termoeléctrico estacionario**, que permite conocer la distribución de la temperatura y la densidad de corriente en una sección radial del electrodo en estado estacionario. El modelo electromagnético se obtiene a partir de las ecuaciones de Maxwell en régimen armónico de baja frecuencia y el modelo térmico a partir de la ecuación de transferencia de calor en estado estacionario. Se trata de un problema acoplado, ya que la fuente de calor en el problema térmico es el efecto Joule y, por otra parte, los parámetros termoeléctricos dependen de la temperatura.

- Un modelo **termoeléctrico transitorio**, que resuelve las ecuaciones de Maxwell acopladas con la ecuación de transferencia de calor en estado transitorio. Así, puede obtenerse la evolución de la temperatura con el tiempo y tener en cuenta aspectos de la operación que dependen del tiempo como son los descensos del electrodo, las desconexiones de la red eléctrica, etc.

- Un modelo **termomecánico**, que permite conocer la distribución de esfuerzos en el electrodo debido a su peso y a los gradientes térmicos. Los resultados obtenidos han permitido, por ejemplo, mejorar el diseño de los *nipples*, que son las piezas que unen las columnas de grafito.

Para resolver numéricamente los diferentes modelos se han escrito programas en Fortran 77, que han sido utilizados para simular el comportamiento de electrodos reales. En las referencias [3], [7] y [15] se presentan algunos de los resultados obtenidos por los diferentes modelos. En la Figura 3 se presenta, por ejemplo, la evolución de la temperatura con respecto al tiempo en un punto situado en el eje del electrodo a la altura de placas. Nótese que en la simulación se tienen en cuenta los deslizamientos y desconexiones del electrodo de la red eléctrica varias horas al día. Además, se ha desarrollado un paquete informático llamado ELSATE, que incluye un menú desplegable que permite al usuario la introducción de datos a través de cuadros de diálogo muy simples. El paquete permite la visualización de una amplia gama de resultados. En la Figura 4 se presenta un ejemplo de los cuadros de diálogo.

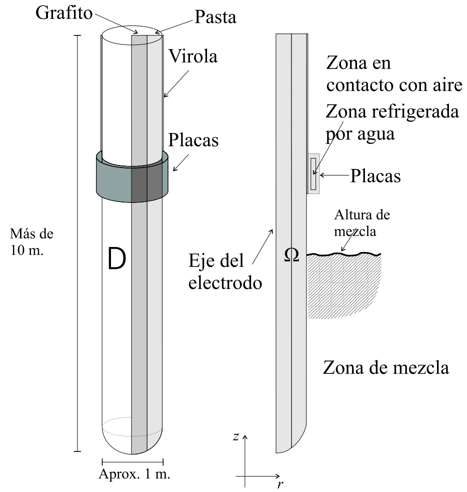
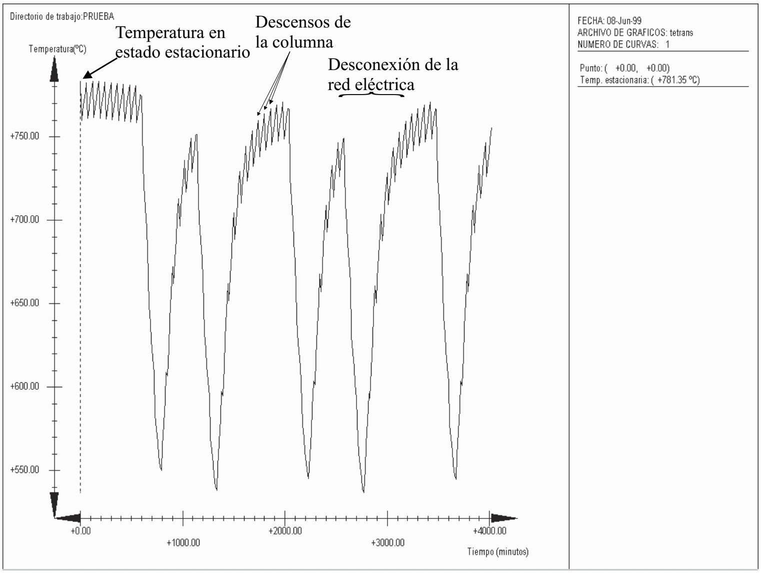


Figura 2: Esquema del electrodo ELSA y dominio bidimensional de los modelos axisimétricos.

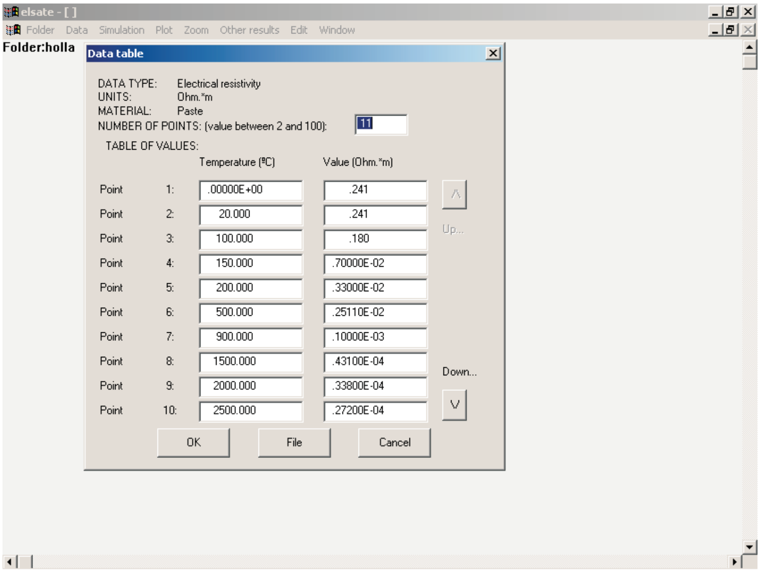


FECHA: 08-Jun-99
ARCHIVO DE GRAFICOS: tetrans
NUMERO DE CURVAS: 1

Punto: (+0.00, +0.00)
Temp. estacionaria: (+781.35 ºC)

Figura 3: Evolución de la temperatura.

Los modelos bidimensionales descritos previamente han proporcionado información importante sobre algunos de los parámetros que intervienen en el funcionamiento del electrodo ELSA y tienen la ventaja de un importante ahorro computacional con respecto a un modelo puramente tridimensional. Sin embargo, como se ha indicado anteriormente, la hipótesis de simetría cilíndrica



Folder:holla

Data table

DATA TYPE: Electrical resistivity
UNITS: Ohm.*m
MATERIAL: Paste
NUMBER OF POINTS: (value between 2 and 100): 11

TABLE OF VALUES:

		Temperature (ºC)	Value (Ohm.*m)
Point	1:	.00000E+00	.241
Point	2:	20.000	.241
Point	3:	100.000	.180
Point	4:	150.000	.70000E-02
Point	5:	200.000	.33000E-02
Point	6:	500.000	.25110E-02
Point	7:	900.000	.10000E-03
Point	8:	1500.000	.43100E-04
Point	9:	2000.000	.33800E-04
Point	10:	2500.000	.27200E-04

Figura 4: Un ejemplo de menú del paquete ELSATE.

obliga a despreciar ciertos hechos como el efecto de los otros dos electrodos (el llamado *efecto de proximidad*) o la no simetría de las condiciones de contorno. Estos hechos solo pueden ser considerados por un modelo puramente tridimensional. Además, los modelos tridimensionales son necesarios para simular los electrodos Soderberg que por su propia geometría no poseen simetría cilíndrica. Por ello, con el objetivo de tener en cuenta los efectos tridimensionales se ha desarrollado un **modelo termoeléctrico tridimensional**, que es lo suficientemente general para simular cualquier tipo de electrodo e incluso el horno completo. La necesidad de resolver el problema electromagnético a partir de datos realistas, es decir, medibles en la práctica, nos ha llevado a investigar en el campo de las ecuaciones de Maxwell en baja frecuencia (ver [15], para detalles).

4 Modelado de nuevos sistemas de coladas para ferroaleaciones

4.1 Descripción del proceso físico

El silicio producido en los hornos de arco eléctrico se extrae en estado líquido a través de un orificio situado en la parte inferior de la cuba. El sólido que se obtiene tras su enfriamiento debe ser desmenuzado para obtener trozos de silicio del tamaño deseado. El silicio solidifica en forma de cristales. En el frente de cristalización, las impurezas que contiene el metal se desplazan a la parte aún líquida, por lo que la mayor concentración de éstas se dará en las últimas zonas que hayan alcanzado la temperatura de solidificación.

El procedimiento más sencillo consiste en almacenar el silicio líquido en unos recipientes cilíndricos llamados *cucharas*, para dejarlo enfriar al aire (ver Figura 5). Pero con este sistema el eje central del bloque de silicio contendrá un mayor porcentaje de impurezas que los bordes y al triturarlo no es posible conseguir que la pureza de cada trozo sea homogénea, como exigen algunos de los usos de este producto. En este contexto, en la Sección 5 se presenta un nuevo mecanismo de enfriamiento del silicio a partir de la cuchara tradicional con el objetivo de obtener un silicio con un reparto de impurezas más homogéneo.

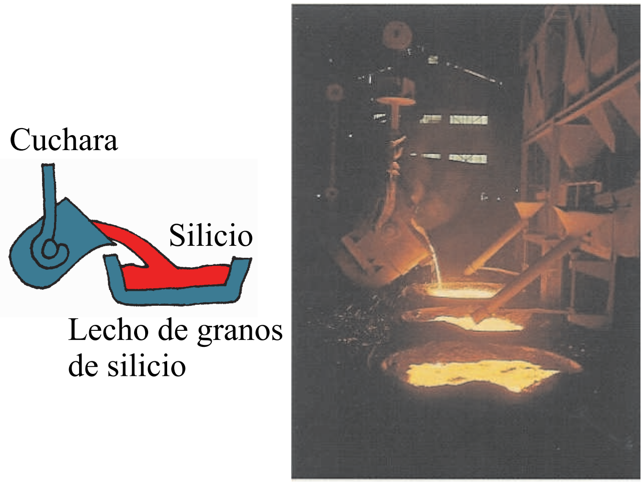


Figura 5: Esquema y foto de la colada clásica.

Ferroatlántica I+D ha desarrollado una alternativa a los sistemas tradicionales de colada que consiste en verter el silicio líquido en una placa refrigerada para producir una delgada lámina de metal en la que la concentración de impurezas sea constante en el sentido longitudinal (ver Figura 7). El sistema consiste en un canal que riega la placa en zig-zag y en una placa que está inclinada y posee un mecanismo que la hace vibrar para permitir que la lámina de silicio se deslice sin pegarse a la placa. Una serie de tubos por los que circula agua la enfrían por su parte inferior.

La placa de enfriamiento se compone de varias secciones. Cada una de ellas mide casi 3 metros de largo, cerca de 2 metros de ancho y 50 mm de alto. Las dos primeras son de cobre y las restantes de hierro, hasta alcanzar una longitud de unos 15 metros. El sistema de refrigeración está compuesto por varios grupos de tubos horadados en la placa (ver Figura 6). Una serie de seis tubos se repite a lo largo de toda la placa. Los tubos de mayor diámetro contienen en su interior otro tubo de diámetro menor, soldado a su parte inferior, para que el agua circule por la parte superior del primero.

La capa de silicio cubre el 80 % de la anchura de la placa. Se estima que la producción ronda las 8 toneladas de silicio por hora.

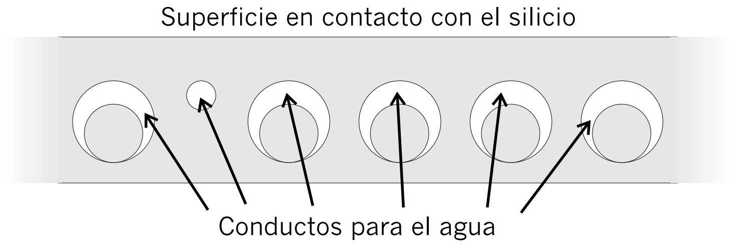


Figura 6: Detalle de un corte longitudinal de la placa.

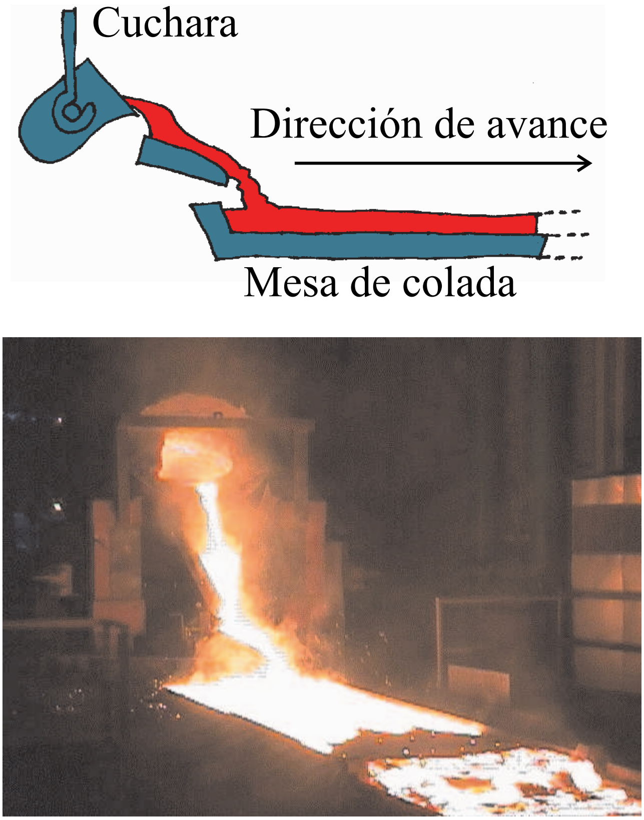


Figura 7: Esquema y foto de una placa de colada.

La simulación numérica de la transferencia de calor que se produce en la lámina de silicio y en la placa se convierte en una herramienta muy interesante para conocer mejor el proceso de enfriamiento. Son varios los objetivos que se persiguen: por un lado se desea aumentar el caudal de silicio colado, pero sin producir daños por fusión o deformación en la placa de cobre (nótese que el silicio funde a una temperatura muy superior a la del cobre); por otro se pretende conocer (e intentar controlar) la región de la lámina de silicio que solidifica en último lugar, es decir, donde se concentran la mayor parte de las impurezas. Para ello, se ha desarrollado un modelo matemático que permite obtener la distribución de temperaturas en la placa de refrigeración cubierta por la lámina de silicio en movimiento continuo.

4.2 El modelo matemático

El problema que se plantea es hallar la temperatura de un cuerpo compuesto por la plancha de cobre refrigerada y la lámina de silicio existente sobre ella.

Dado que la anchura de la placa es mucho mayor que su altura y los gradientes de temperatura son mucho menores en la primera dirección que en la segunda, es factible restringir el problema al estudio de lo que ocurre en una sección longitudinal de la placa (ver Figura 8). Tomaremos como dominio $\Omega \subset \mathbb{R}^2$ la unión de la sección media de la placa de cobre, Ω_c, y de la lámina de silicio, Ω_s.

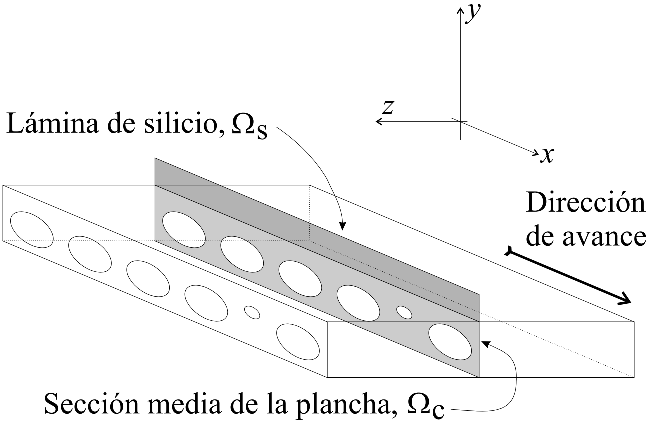

y
z
x
Lámina de silicio, Ω_S
Dirección de avance
Sección media de la plancha, Ω_C

Figura 8: Sección media de la placa.

El modelo se obtiene a partir de la ecuación de transferencia de calor en estado transitorio. Además, como el silicio cambia de estado líquido a sólido debemos tener en cuenta el calor latente liberado al alcanzar la temperatura de solidificación T_S. Por ello, conviene escribir la ecuación de transmisión del calor en términos de la variable *entalpía* e:

$$\frac{\partial e(\mathbf{x})}{\partial t} + \mathbf{v}(\mathbf{x}) \cdot \mathbf{grad}\, e(\mathbf{x}) - \operatorname{div}\left((k(\mathbf{x})\, \mathbf{grad}\, T(\mathbf{x})\right) = 0, \tag{2}$$

donde $\mathbf{x}$ es la coordenada espacial, T es la temperatura, $\mathbf{v}$ es la velocidad y k la conductividad térmica.

La entalpía en el silicio se expresa como función de la temperatura a través de un operador monótono multivaluado

$$e(\mathbf{x}) \in H(\mathbf{x}, T), \tag{3}$$

siendo

$$\mathcal{H}(\mathbf{x},T) = \begin{cases} \int_0^T \rho\, c\, ds, & T < T_S, \\ \left[\int_0^T \rho\, c\, ds, \int_0^T \rho\, c\, ds + \rho(\mathbf{x},T_S)\, L\right], & T = T_S, \\ \int_0^T \rho\, c\, ds + \rho(\mathbf{x},T_S)\, L, & T > T_S, \end{cases} \tag{4}$$

donde L el "calor latente de fusión" o calor por unidad de masa necesario para realizar el cambio de estado, ρ es la densidad del silicio y c su calor especifico.

En el caso de la placa, al no haber cambio de estado, la expresión para la entalpía es, simplemente,

$$H(\mathbf{x},T) = \int_0^T \rho(\mathbf{x},s)\, c(\mathbf{x},s)\, \mathrm{d}s. \tag{5}$$

Cabe señalar que todos los parámetros dependen de la posición (por haber dos materiales, silicio y cobre) y de la temperatura.

La velocidad a la que se mueve el silicio se supone constante y con sólo componente horizontal. La placa de cobre, por su parte, está quieta. La velocidad es, por tanto,

$$\mathbf{v} = \begin{cases} v^*\, \mathbf{e}_x & \text{en } \Omega_s, \\ \mathbf{0} & \text{en } \Omega_c. \end{cases}$$

El modelo se completa definiendo condiciones de contorno adecuadas. En concreto, en la frontera vertical izquierda del dominio correspondiente a la entrada del silicio, se supone conocida e igual a la temperatura a la que éste sale de la cuchara. En el resto de las fronteras, se consideran condiciones de radiación-convección siendo nula la radiación en la zona en contacto con el agua. Los coeficientes de convección han sido obtenidos a partir de fórmulas semiempíricas de la bibliografía (ver [14] para más detalles).

Por otra parte, la placa posee un mecanismo vibratorio que permite el deslizamiento del silicio. Como la amplitud de las vibraciones es pequeña comparada con el movimiento de la lámina de silicio, hemos supuesto que la velocidad sólo tiene componente horizontal. Sin embargo, un aspecto en el que inciden de forma determinante las vibraciones es la transmisión del calor entre el silicio y la placa. Llamemos Γ_{I+} y Γ_{I-} a las superficies de silicio y cobre en las que se produce el contacto, T_+ a la temperatura del silicio y T_- a la del cobre. Debido a la vibración, no podremos asumir que exista un contacto térmico perfecto entre ambos materiales, sino que supondremos que existe una transmisión imperfecta, es decir, que existe una "resistencia de contacto", regulada por una ley del tipo siguiente:

$$-k(T_+)\frac{\partial T_+}{\partial \mathbf{n}} = h_r(T_+ - T_-) \quad \text{en } \Gamma_{I+}, \tag{6}$$

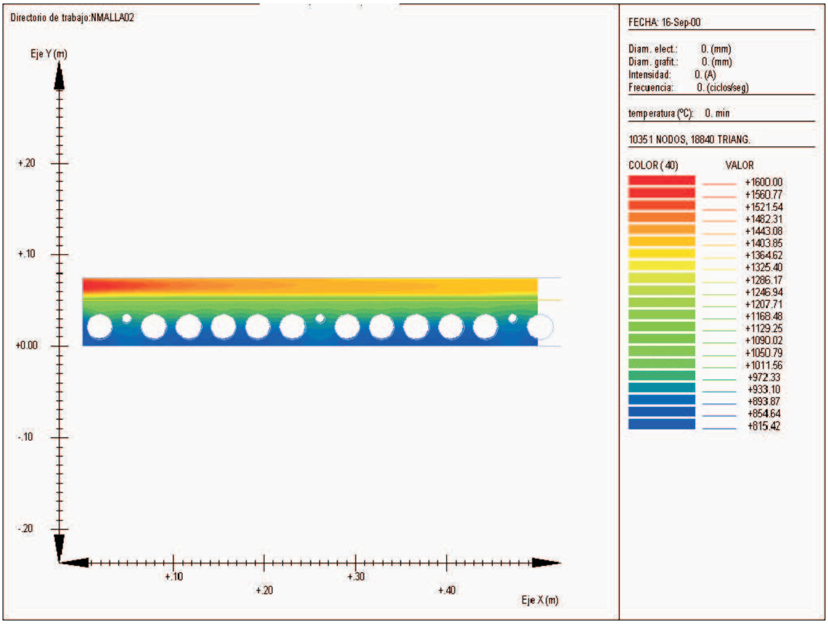
$$-k(T_-)\frac{\partial T_-}{\partial \mathbf{n}} = h_r(T_- - T_+) \quad \text{en } \Gamma_{I-}. \tag{7}$$

El coeficiente h_r es a priori desconocido y muy difícil de medir de forma directa. Por ello, este parámetro ha sido ajustado numéricamente a partir de medidas experimentales de la temperatura de salida del agua de refrigeración.

La ecuación (2) se discretiza en tiempo utilizando un esquema implícito en el que el término convectivo se discretiza mediante el método de características. La ecuación semidiscretizada se resuelve usando elementos finitos continuos y lineales a trozos sobre su formulación variacional. Para tener en cuenta el calor latente de solidificación del silicio y la no linealidad debida a los parámetros, se ha usado un algoritmo iterativo (ver [4] para más detalle). El hecho de no conocer la resistencia térmica entre silicio y placa nos ha obligado utilizar este parámetro como elemento de ajuste para que el modelo reproduzca el incremento observado en la temperatura de salida del agua de refrigeración.

4.3 Resultados numéricos

El algoritmo para resolver el problema de la colada ha sido implementado en un ordenador mediante un programa escrito en lenguaje Fortran. Presentamos aquí los resultados de una simulación sobre un dominio semejante al de la placa de enfriamiento utilizada por Ferroatlántica S.L.


Directorio de trabajo:NMALLA02

FECHA: 16-Sep-00

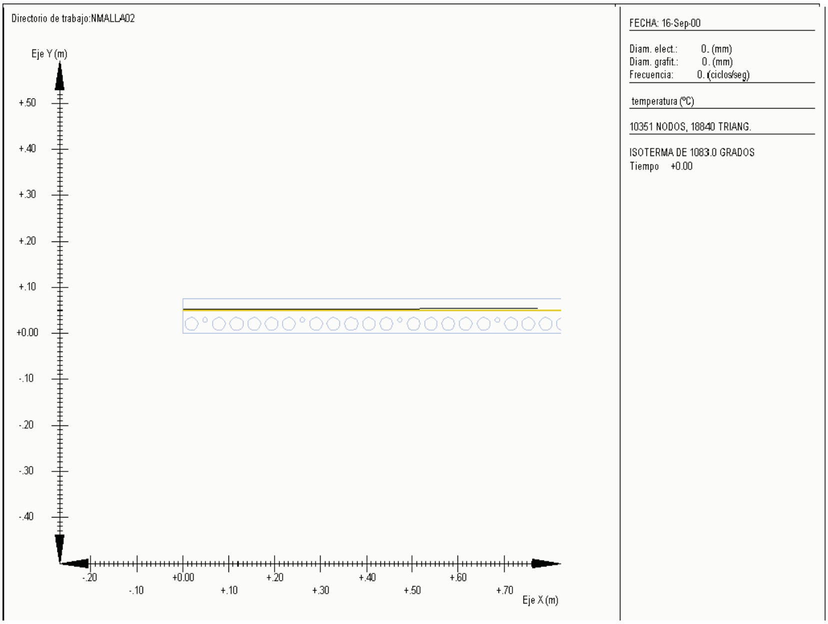
Diam. elect.: 0. (mm)
Diam. grafit.: 0. (mm)
Intensidad: 0. (A)
Frecuencia: 0. (ciclos/seg)

temperatura (ºC): 0. min

10351 NODOS, 18840 TRIANG.

Figura 9: Temperatura a la entrada de la placa

En todas las figuras, el dominio es una sección longitudinal de la placa junto con una sección de la lámina de silicio situado sobre ella. El silicio líquido entra en el dominio por la parte izquierda y el movimiento vibratorio lo desplaza hacia la derecha. Como se ve en la Figura 9, la temperatura en la lámina de silicio es superior en su interior, pues tanto el aire como el agua a través del cobre extraen el calor del silicio líquido.



FECHA: 16-Sep-00

Diam. elect.: 0. (mm)
Diam. grafit.: 0. (mm)
Frecuencia: 0. (ciclos/seg)

temperatura (ºC)

10351 NODOS, 18840 TRIANG.

ISOTERMA DE 1083.0 GRADOS
Tiempo +0.00

Figura 10: Isoterma de fusión del cobre.

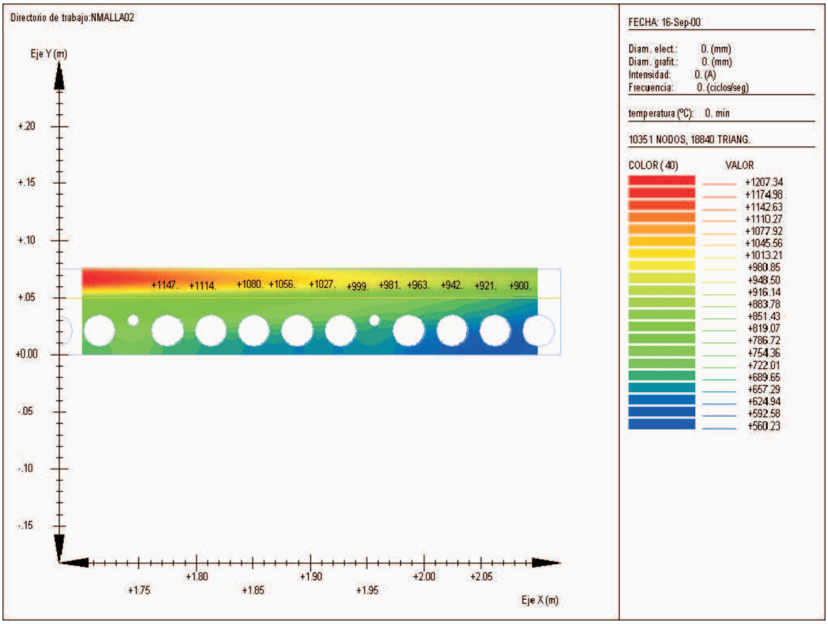
La temperatura de entrada se toma a 1600 °C, casi 200 °C por encima de su temperatura de fusión.

Lógicamente, el comienzo de la placa es el lugar donde existe un mayor riesgo de que el cobre se funda o deforme por alcanzar altas temperaturas. Se han hecho varias pruebas para ajustar cuál debería ser el caudal de agua a circular por los tubos de refrigeración para evitar que el cobre alcance su temperatura de fusión, 1083 °C. La resistencia de contacto provoca que la temperatura deje de ser continua al pasar a través de la interfase entre grafito y cobre. Como se ve en la Figura 10, la isoterma de 1083 °C no interseca la placa de cobre.

5 Modelos de purificación del silicio

5.1 Descripción del proceso físico

En los últimos años se ha producido un importante crecimiento de la demanda de silicio de alta pureza para la fabricación de placas solares. Actualmente el silicio solar se obtiene como un subproducto de la fabricación de silicio electrónico. Sin embargo las estimaciones de la demanda de silicio solar superan a las del electrónico y es por eso por lo que las grandes compañías están llevando a cabo grandes proyectos de investigación para encontrar métodos alternativos y directos para la purificación del silicio metal de manera que se obtenga silicio



Directorio de trabajo:NMALLA02

FECHA: 16-Sep-00

Diam. elect.: 0. (mm)
Diam. grafit.: 0. (mm)
Intensidad: 0. (A)
Frecuencia: 0. (ciclos/seg)

temperatura (ºC): 0. min

10351 NODOS, 18840 TRIANG.

Figura 11: Temperatura en la parte final del dominio.

solar. En este contexto, el diseño de nuevos sistemas de colada como el visto en la sección anterior u otros procedimientos como el que se describe en esta sección persiguen obtener siempre un silicio más puro.

Un sistema utilizado por Ferroatlántica para enfriar el silicio que sale del horno consiste en un recipiente cilíndrico que contiene en su interior un crisol (ver Figura 12). En el crisol se introduce silicio fundido que irá solidificando por la pérdida de calor, creando un frente de solidificación que progresa de la parte inferior a la superior. Es importante que este frente de solidificación sea lo más horizontal posible de modo que las impurezas emigren hacia las capas altas para ser posteriormente eliminadas. Por ello, conocer y controlar la distribución de temperaturas en el silicio bajo diferentes condiciones de operación y con diferentes geometrías son los objetivos que se persiguen con la simulación numérica de este proceso.

5.2 El modelo matemático

El modelo matemático que permite obtener la distribución de temperaturas en el recipiente descrito anteriormente y en el silicio contenido en su interior es la ecuación de transmisión del calor en estado transitorio sin convección, es decir, con velocidad nula. Supondremos que los campos no dependen de la coordenada angular, lo cual permite resolver el problema en un dominio bidimensional. En concreto, el dominio de cálculo es la sección del dominio cilíndrico que se presenta en la Figura 13. Nótese que el interior del cilindro está formado

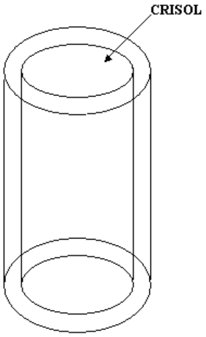


Figura 12: Esquema del dominio.

por diferentes materiales: la capa externa es una manta de alúmina aislante, mientras que la interior está formada por cemento.

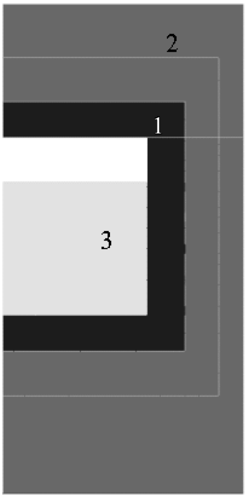


Figura 13: Sección radial de la cuchara: manta de alúmina (2), cemento (1), silicio (3).

Dado que el silicio cambia de estado durante el proceso es necesario utilizar la ecuación de transmisión del calor en términos de la entalpía, que escrita en coordenadas cilíndricas queda de la forma:

$$\dot{e} - \frac{1}{r}\frac{\partial}{\partial r}\left(k\frac{\partial T}{\partial r}\right) - \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) = 0. \tag{8}$$

La expresión de la entalpía en términos de la temperatura es la descrita en la sección anterior dependiendo de si el material experimenta o no un cambio de estado.

Con respecto a las condiciones de contorno, en las fronteras exteriores del recipiente se considera una condición de convección, mientras que en el hueco que queda entre la tapa y el crisol se considera una condición de radiación. En el resto de las fronteras se proporciona el flujo de calor, que es conocido a partir de resistencias internas que calientan el recipiente.

El esquema numérico utilizado es el mismo que el descrito en la sección anterior.

5.3 Resultados numéricos

Se ha desarrollado un programa Fortran que permite resolver la ecuación de transferencia de calor en la cuchara descrita anteriormente. En esta sección se presentan los resultados obtenidos al estudiar el comportamiento de la cuchara bajo determinadas condiciones de operación. En concreto, se realiza un precalentamiento de la cuchara mediante resistencias internas durante 10 horas. A continuación, se introduce el silicio en el crisol a una temperatura de 1480 °C y se conectan resistencias en la parte inferior de la tapa. La potencia máxima de las resistencias internas es de 10 Kw, manteniendo esta potencia mientras las temperaturas en ellas sean inferiores a 1450 °C. A partir de 1450 °C, la potencia de las resistencias disminuye linealmente hasta ser apagadas definitivamente a 1550 °C.

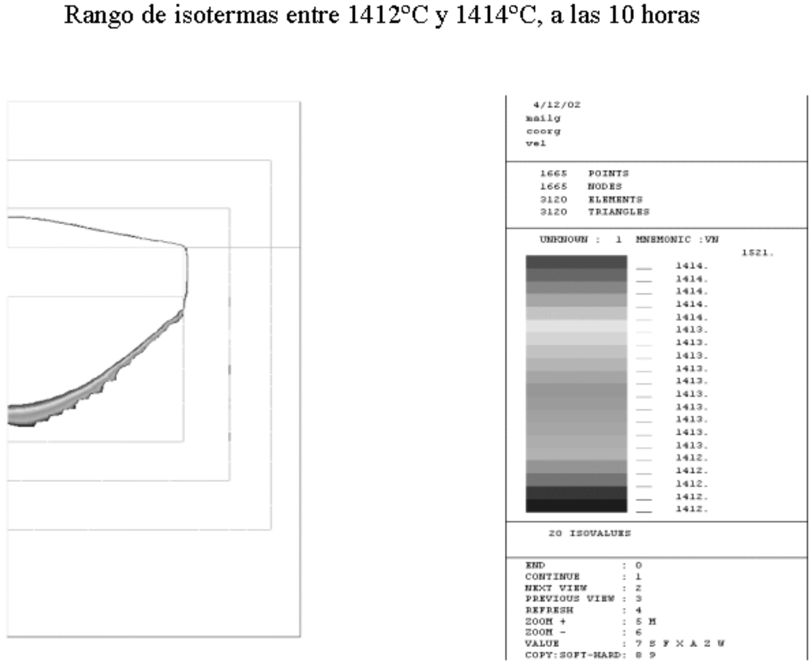
Con la composición de materiales descrita en la sección anterior, la Figura 14 muestra el frente de solidificación obtenido (temperaturas entre 1412 y 1414 °C) a las 10 horas.

Dado que el principal objetivo consistía en obtener un frente de solidificación más horizontal, se abordó la posibilidad de introducir un nuevo material muy conductor tal y como se indica en la Figura 15. La conclusión es que cuanto mayor sea la conductividad de dicho material, el frente es más horizontal. La Figura 16 presenta el frente de solidificación obtenido con esta nueva geometría al cabo de 10 horas, siendo la conductividad eléctrica del material 1000 W/mK.

6 Líneas de investigación en curso: modelado de un horno de inducción para la purificación de silicio.

En el contexto de la sección anterior, actualmente la empresa Ferroatlántica I+D está investigando nuevos sistemas de purificación de silicio. En concreto, el diseño de un horno destinado al calentamiento por inducción electromagnética, del silicio contenido en un crisol es uno de los actuales temas de investigación. En esta sección, se describen brevemente los modelos matemáticos que se están desarrollando por parte del Departamento de Matemática Aplicada para contribuir al diseño de dicho horno.

Cabe señalar que el calentamiento por inducción electromagnética es ampliamente utilizado en la industria actual en procesos como fundición de



Rango de isotermas entre 1412°C y 1414°C, a las 10 horas

```
 4/12/02
mailg
coorg
vel

  1665   POINTS
  1665   NODES
  3120   ELEMENTS
  3120   TRIANGLES

  UNKNOWN :  1  MNEMONIC :VN
                               1521.
                 __     1414.
                 __     1414.
                 __     1414.
                 __     1414.
                 __     1414.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1413.
                 __     1412.
                 __     1412.
                 __     1412.
                 __     1412.
                 __     1412.

    20 ISOVALUES

END            : 0
CONTINUE       : 1
NEXT VIEW      : 2
PREVIOUS VIEW  : 3
REFRESH        : 4
ZOOM +         : 5 M
ZOOM -         : 6
VALUE          : 7 S F X A Z W
COPY:SOFT-HARD: 8 9
```

Figura 14: Frente de solidificación utilizando 3 materiales.

metales, precalentamiento para operaciones de soldadura y, en general, en aquellos procesos que requieren una velocidad alta de calentamiento en zonas localizadas de la pieza de un material conductor. Se trata de un proceso complejo que involucra fenómenos electromagnéticos, fenómenos térmicos con cambio de estado y fenómenos hidrodinámicos en las fases líquidas del metal. Por ello, la simulación numérica es una herramienta muy eficaz en este campo. En el caso concreto del horno de inducción, la simulación numérica permite conocer y controlar los parámetros que intervienen en su funcionamieto.

El objetivo que se persigue es estudiar los diferentes modelos matemáticos que permiten simular los procesos físico-químicos mencionados anteriormente y desarrollar métodos numéricos adecuados para su resolución numérica. Se pretende, por tanto, la realización de un programa de ordenador que resuelva numéricamente las ecuaciones de la termo-electromagneto-hidrodinámica de modo que se puedan obtener, en un tiempo de cálculo razonable, las corrientes inducidas, la distribución de la temperatura junto con la posición de la interfase sólido-líquido y el campo de velocidades en el material líquido.

Desde un punto de vista matemático, el modelado completo de un horno de inducción implica el acoplamiento de las ecuaciones de Maxwell con la ecuación del calor con términos de convección-difusión y cambio de estado, así como las ecuaciones de Navier-Stokes para un fluido incompresible. Las relaciones entre

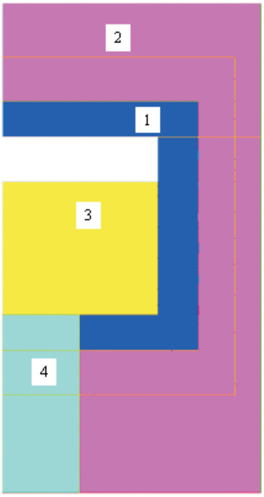
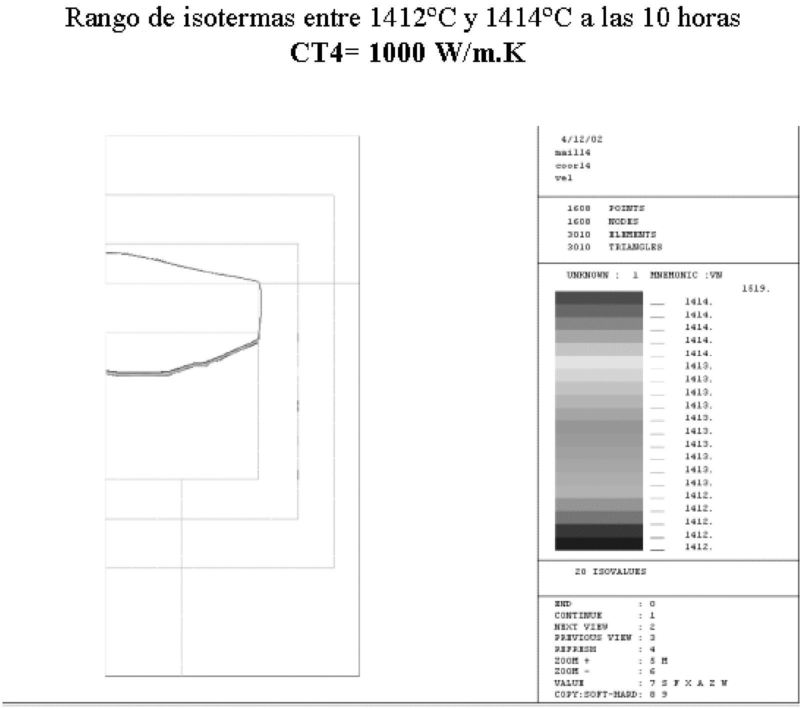


Figura 15: Sección radial de la cuchara: manta de alúmina (2), cemento (1), silicio (3), material muy conductor (4).

los modelos anteriores se esquematizan en el diagrama que se presenta en la Figura 17.

Este proceso presenta características muy complejas entre las que cabe destacar:

- **Acoplamiento de los fenómenos:**
  - Modelo Electromagnético - Modelo Térmico: el campo de temperaturas depende de la densidad de corriente a través del efecto Joule, mientras que las conductividades eléctricas de los diferentes materiales dependen de la temperatura.
  - Modelo Electromagnético - Modelo hidrodinámico: la fuerza de Lorentz combinada con la gravitatoria y las diferencias de densidad con la temperatura provocan el movimiento del metal fundido y dicho movimiento produce, a su vez, variaciones de los campos electromagnéticos a través de la Ley de Ohm.
  - Modelo térmico - Modelo hidrodinámico: La resolución del primero dará como resultado la posición del frente de solidificación del metal y, en consecuencia, la zona líquida donde se resolverán numéricamente las ecuaciones de la hidrodinámica. Téngase en cuenta que la viscosidad del metal líquido también depende de la temperatura. Asimismo, el campo de velocidades que proporcionará la resolución numérica del segundo modelo se incluirá en la parte convectiva de la ecuación del calor.



```
 4/12/02
mailll4
coorl4
vel

  1608   POINTS
  1608   NODES
  3010   ELEMENTS
  3010   TRIANGLES

  UNKNOWN :  1  MNEMONIC :VN
                                  1819.
            —    1414.
            —    1414.
            —    1414.
            —    1414.
            —    1414.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1413.
            —    1412.
            —    1412.
            —    1412.
            —    1412.
            —    1412.

    20 ISOVALUES

END            : 0
CONTINUE       : 1
NEXT VIEW      : 2
PREVIOUS VIEW  : 3
REFRESH        : 4
ZOOM +         : 5 M
ZOOM -         : 6
VALUE          : 7 S F X A Z W
COPY:SOFT-HARD: 8 9
```

Figura 16: Frente de solidificación utilizando 4 materiales.

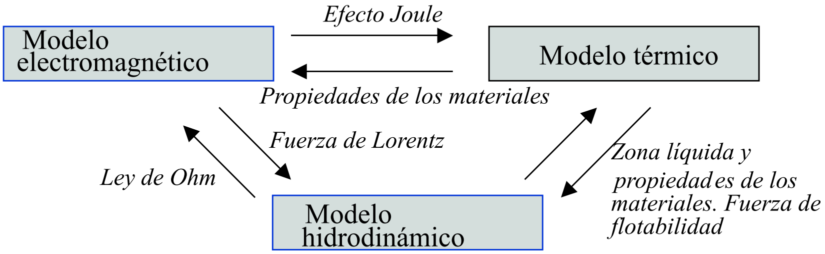

Modelo electromagnético
Efecto Joule
Modelo térmico
Propiedades de los materiales
Fuerza de Lorentz
Ley de Ohm
Zona líquida y propiedades de los materiales. Fuerza de flotabilidad
Modelo hidrodinámico

Figura 17: Acoplamiento entre los diferentes modelos.

- **Presencia de términos no lineales:**
  - La fuente de calor en el modelo térmico es la densidad de potencia del efecto Joule, que es proporcional al cuadrado de la densidad de corriente.
  - Las propiedades físicas de los materiales que forman el horno y del propio metal que se funde dependen, en general de forma no lineal, de la temperatura.

  - La fusión del metal dentro del horno requiere que se resuelva la ecuación del calor evolutiva con cambio de estado, lo que introduce una fuerte no linealidad en el problema.
  - Las condiciones de contorno del modelo térmico serán de convección-radiación con lo cual el flujo de calor en la frontera depende de la potencia cuarta de la temperatura.

- **Existencia de fronteras libres:** la interfase sólido-líquido del material que se introduce en el horno de inducción. Téngase en cuenta que el conocimiento de dicha interfase es crucial para los modelos electromagnético e hidrodinámico, puesto que las propiedades físicas varían considerablemente al cambiar de estado y, además, el modelo hidrodinámico solo se desarrolla en la fase líquida.

En la literatura pueden encontrarse algunos trabajos donde se acoplan el modelo térmico y el modelo electromagnético (ver por ejemplo [8], [9]) o bien el modelo electromagnético y el modelo hidrodinámico ([13]). Sin embargo, el acoplamiento de los tres modelos constituye una línea de investigación novedosa. Recientemente, en [11] se presenta la simulación numérica de un horno de inducción, también destinado a la purificación del silicio, que se lleva a cabo mediante códigos comerciales.

Finalmente, las técnicas numéricas que se emplearán para la resolución de las ecuaciones que modelan los tres procesos anteriormente descritos son:

- Modelo electromagnético: se utilizará un método híbrido de elementos finitos (FEM) y elementos de contorno (BEM). La combinación de ambas técnicas permite utilizar sus respectivas ventajas numéricas: el método de elementos finitos nos permite aproximar bien la geometría y tratar eficazmente los términos no lineales, y las matrices resultantes son "huecas". Por otro lado, los elementos de contorno proporcionan un método eficaz para tratar dominios homogéneos no acotados. Los BEM reducen el número de incógnitas pero tienen el inconveniente de que dan lugar a matrices llenas.

- Modelo térmico: La ecuación del calor se integrará en tiempo mediante un esquema implícito mientras que, en espacio, se utilizará el método de los elementos finitos. Además, se emplearán técnicas de operadores maximales monótonos para el tratamiento del cambio de estado.

- Modelo hidrodinámico: se integrarán en tiempo las ecuaciones de Navier-Stokes utilizando un esquema implícito donde el término inercial será aproximado mediante un método de características. En cada paso de tiempo, se utilizará un método de elementos finitos; en particular, se considerarán elementos continuos lineales a trozos con "burbujas"para las componentes de la velocidad y lineales a trozos para la presión.

Agradecimientos

Este trabajo ha sido financiado por FERROATLANTICA I+D y los proyectos FEDER-CICYT 1FD97.0280 y PGIDT00PXI20701PR.

Referencias

[1] A. Bermúdez, J. Bullón and F. Pena, "A finite element method for the thermoelectrical modelling of electrodes", *Commun. Numer. Meth. Engng.* **14**, (1998), 581-593.

[2] A. Bermúdez and R. Muñoz, "Existence of solution of a coupled problem arising in the thermoelectrical simulation of an electrode", *Quart. of Appl. Math.* **57**, (1999), no. 4, 621-636.

[3] A. Bermúdez, J. Bullón, F. Pena y P. Salgado, "Modelado y simulación numérica de electrodos para hornos metalúrgicos", CD-ROM Actas del XVII Congreso de Ecuaciones Diferenciales y Aplicaciones/VII Congreso de Matemática Aplicada (XVII CEDYA/VII CMA) Salamanca, España, 2001.

[4] A. Bermúdez, J. Bullón y F. Pena, "Simulación estacionaria de una colada de silicio" CD-ROM Actas del XVII Congreso de Ecuaciones Diferenciales y Aplicaciones/VII Congreso de Matemática Aplicada (XVII CEDYA/VII CMA) Salamanca, España, 2001.

[5] A. Bermúdez, R. Rodríguez and P. Salgado, "A finite element method with Lagrange multipliers for low-frequency harmonic Maxwell equations", *SIAM J. Numer. Anal.*, **40**, (2002), 1823-1849.

[6] A. Bermúdez, R. Rodríguez and P. Salgado, "Numerical treatment of realistic boundary conditions for the eddy current problem in an electrode via Lagrange multipliers", Preprint 2002-11, Universidad de Concepción, Chile, 2002.

[7] A. Bermúdez, J. Bullón, F. Pena and P. Salgado, "A numerical method for transient simulation of metallurgical compound electrodes", *Finite Elem. Anal. Des.*, **39**, (2003), 283-299.

[8] C. Chaboudez, S. Clain, R. Glardon, D. Mari, J. Rappaz, and M. Swierkosz, "Numerical Modeling in Induction Heating for axisymmetric geometries", *IEEE Transactions on Magnetics*, **33**(1), (1997), 739-745.

[9] S. Clain, J. Rappaz, M. Swierkosz and R. Touzani, "Numerical modelling of induction heating for two-dimensional geometries", *Mathematical Models and Methods in Applied Sciences*, **3**(6), (1993), 805-822.

[10] P. D'Ambrosio and I. Letizia, "Temperature and stress distribution on carbon electrodes for silicon metal production under transient temperature conditions", *16th Biennial Conference on Carbon*, Baden-Baden, (1983).

[11] Y. Delannoy, C. Alemany, K.I. Li, P. Proulx and C. Trassy, "Plasma-refining process to provide solar-grade silicon", *Solar Energy Materials & Solar Cells*, **72**, (2002), 69-75.

[12] R. Innvær and L. Olsen, "Practical use of mathematical models for Søderberg electrodes", *Elkem Carbon Technical Paper presented at the A.I.M.E. Conference* (1980).

[13] T.T. Natarajan and N. El-Kaddah, "A methodology for two-dimensional finite element analysis of electromagnetically driven flow in induction stirring systems", *IEEE Transactions on Magnetics*, **35**(3), (1999), 1773-1776.

[14] F. Pena, *Contribución al modelado matemático de algunos problemas de la metalurgia del silicio*, Tesis Doctoral, Universidade de Santiago de Compostela, 2003

[15] P. Salgado, *Mathematical and numerical analysis of some electromagnetic problems. Application to the simulation of metallurgical electrodes*, Tesis Doctoral, Universidade de Santiago de Compostela, 2002

[16] A. Schei, J.K. Tuset and H. Tveit, *High Silicon alloys,* Tapir Forlag, Trondheim, Noruega, 1998.