Computational Methods In Applied Mathematics

Volume 10 (2010) **Number 2**

CONTENTS

Computational Methods in Applied Mathematics (ISSN 1609–4840) is published four times a year.

Founder
Institute of Mathematics of the National Academy of Sciences of Belarus

Manuscripts and correspondence concerning manuscripts under review should be addressed to the Editor, Prof. P. Matus by e-mail: cmam@cmam.info .
Editorial Board of the Journal "Computational Methods in Applied Mathematics", 11 Surganov St., 220072 Minsk, Belarus. Tel./Fax: (37517) 284 1963. Information for contributors is available at http://www.cmam.info

$60 \times 84/8$

○

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 119–136

A PRECONDITIONED MINIMAL RESIDUAL SOLVER FOR A CLASS OF LINEAR OPERATOR EQUATIONS

O. AWONO[1] AND J. TAGOUDJEU[2]

Abstract — We consider the class of linear operator equations with operators admitting self-adjoint positive definite and m-accretive splitting (SAS). This splitting leads to an ADI-like iterative method which is equivalent to a fixed point problem where the operator is a 2 by 2 matrix of operators. An infinite dimensional adaptation of a minimal residual algorithm with Symmetric Gauss-Seidel and polynomial preconditioning is then applied to solve the resulting matrix operator equation. Theoretical analysis shows the convergence of the methods, and upper bounds for the decrease rate of the residual are derived. The convergence of the methods is numerically illustrated with the example of the neutron transport problem in 2-D geometry.

2000 Mathematics Subject Classification: 65J10, 65Jxx, 47Bxx, 47B44,82D75 65Bxx, 65-XX.

Keywords: minimal residual methods, preconditioning, neutron transport, self-adjoint operator, m-accretive, operator splitting.

1. Introduction

Iterative methods are widely used for solving linear operator equations (see [1, 3, 32, 17, 18, 30, 15, 20, 27] and the references therein). The GMRES algorithm for linear equations with bounded operators in a separable Hilbert space was studied in [15]. It was shown that the results of the finite dimensional case can be generalized in the continuous case if the operator is algebraic [15]. Recently, some new iterative methods for solving linear operator equations with bounded [20] and unbounded [30] operators have been introduced and analyzed. These methods make use of the adjoint operator in the transformation of the initial equation. For the particular case of the neutron transport equation, extensive use of iterative methods for continuous and discrete problems has been made (see [2, 4, 5, 14, 13, 22, 24, 28, 29, 33, 34, 37, 38] and the references therein). The standard method is the source iteration method based on the decoupling between the differential and integral parts of the transport operator. This method becomes extremely slow in the critical case. Several acceleration techniques of convergence of the source iteration method such as Diffusion Synthetic Acceleration (DSA) [2, 38] and multigrid algorithms have been introduced and studied [2, 13, 22]. Based on the natural splitting of the integral part of the transport operator, other methods such as Jacobi, Gauss-Seidel [34] and Successive overrelaxation (SOR) iteration have been successfully applied to the transport problem by solving

[1] *Ecole Nationale Supérieure Polytechnique, University of Yaoundé I, PO.Box 8390 Yaoundé, Cameroon.* E-mail: awonoonana@gmail.com

[2] *Faculty of Science and Ecole Nationale Supérieure Polytechnique, University of Yaoundé I, PO.Box 8390 Yaoundé, Cameroon.* E-mail: jtagoudjeu@gmail.com

a fixed point problem derived from the source iteration method. Using the same splitting, an adaptation to the continuous case of the minimal residual iteration method [4, 5] was proposed for the solution of the transport in slab geometry, in 2-D cartesian geometry and in 1-D spherical geometry. This method was proved to be efficient and it competes with the SOR method. Further, its preconditioned versions have been analyzed [35]. Recently, an ADI-like iterative method [25] based on positive definite and m-accretive splitting for linear operator equations with operators admitting such splitting has been proposed and analyzed [6]. This method converges unconditionally and its SOR acceleration [6] yields convergence results similar to those obtained in the presence of finite dimensional systems with matrices possessing the *Young property A* [11, 19, 39] that are matrices with non null diagonal permutationally similar to 2×2 block matrices with diagonal blocks being diagonal matrices. In a particular case where the positive definite part of the linear equation operator is self-adjoint, an upper bound for the contraction factor of the iterative method which depends solely on the spectrum of the self-adjoint part was derived [7]. As such, this method has been successfully applied to the neutron transport equation in slab and 2-D Cartesian geometry [7] and in 1-D spherical geometry [9].

Self-adjoint and m-accretive splitting leads to a fixed point problem where the operator is a 2 by 2 matrix of operators. A preconditioned minimal residual algorithm using symmetric Gauss-Seidel and polynomial preconditioning is then applied to solve the matrix operator equation. Theoretical analysis shows that the methods converge unconditionally and the upper bounds of the residual decrease rate which depend solely on the spectrum of the self-adjoint part of the operator are derived. Each step of the proposed iterative methods requires finding solutions of two linear equations, one with a bounded self-adjoint operator and the other with an m-accretive operator. These linear operator equations can be solved approximately using appropriate methods with respect to the properties of each operator. The convergence of these solvers is numerically illustrated with the example of the neutron transport problem in 2-D geometry. Various test cases, including pure scattering and optically thick domains [31] were considered.

The remainder of this paper is organized as follows: in Section 2 we give the description and the convergence properties of the SAS and minimal residual iteration method. The analysis of the preconditioned version of the minimal residual method using symmetric Gauss-Seidel preconditioning and polynomial preconditioning is considered in Section 3. Section 4 is devoted to the application of the method to the 2-D neutron transport equation and the numerical illustration. Concluding remarks are given in Section 5.

2. The SAS and Minimal Residual Iteration Method

Let us consider a Hilbert space H with inner product $(.,.)$ and the associated norm $\|.\|$. Let X be an unbounded linear operator on H with a domain $\mathcal{D}(X)$. Let I denote the identity operator on H.

Definition 2.1. *The operator X is said to be m-accretive if*

$$\forall u \in D(X), \;\; (Xu, u)_H \geqslant 0$$

and

$$\forall q \in H, \;\; \exists u \in D(X) \;\; \text{such that} \;\; Xu + u = q.$$

We have the following results [12, 14]:

Theorem 2.1. *Assume that X is an m-accretive operator on H. Then*

1. *$D(X)$ is dense in H;*
2. *the operator X is closed;*
3. *$\forall\alpha > 0$, $(I + \alpha X)$ is bijective from $D(X)$ to H, the operator $(I + \alpha X)^{-1}$ is bounded and $\|(I + \alpha X)^{-1}\| \leqslant 1$.*

It follows from Theorem 2.1 that if X is an m-accretive operator, for any positive constant α, $(\alpha I + X)$ is positive definite and $\|(\alpha I + X)^{-1}\| \leqslant \frac{1}{\alpha}$. Thus $(\alpha I + X)^{-1}$ is bounded on H.

Let T be a linear operator on H with a domain $\mathcal{D}(T)$ and a range $\mathcal{R}(T) = H$. We denote by I, the identity operator. Suppose that we need to solve in $\mathcal{D}(T)$ the following problem:

$$Tu = q, \tag{2.1}$$

where $q \in H$ is given and $u \in \mathcal{D}(T)$ is the unknown.

We assume that the operator T admits the following splitting [7, 10]:

$$T = S + A, \tag{2.2}$$

where S is a bounded self-adjoint positive definite operator and A is an m-accretive operator. Therefore, the operator T is positive definite and equation (2.1) admits a unique solution in H.

Let α be a positive constant. The following two-step splitting is obtained from (2.2)

$$\left\{ \begin{array}{l} T = (\alpha I + S) - (\alpha I - A) \\ T = (\alpha I + A) - (\alpha I - S) \end{array} \right. , \tag{2.3}$$

which leads to the following self-adjoint and m-accretive splitting (SAS) iteration method [7] Given an initial guess $u^{(0)} \in D(T)$, for $k = 0, 1, \ldots$ until $\{u^{(k)}\}$ converges, calculate

$$\left\{ \begin{array}{l} (\alpha I + S)u^{(k+\frac{1}{2})} = (\alpha I - A)u^{(k)} + q \\ (\alpha I + A)u^{(k+1)} = (\alpha I - S)u^{(k+\frac{1}{2})} + \end{array} \right. . \tag{2.4}$$

The exact solution u^* of the problem (2.1) verifies [7, 10]

$$\|u^{(k+1)} - u^*\|_{A(\alpha)} \leqslant \beta(\alpha)\|u^{(k)} - u^*\|_{A(\alpha)}, \tag{2.5}$$

where $\|.\|_{A(\alpha)}$ is a norm defined on $\mathcal{D}(T)$ by

$$\|u\|_{A(\alpha)} = \|(\alpha I + A)u\|, \tag{2.6}$$

and

$$\beta(\alpha) = \sup_{\lambda \in \sigma(S)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right|, \tag{2.7}$$

with $\sigma(S)$ denoting the spectrum of S. It follows from the positivity of α and λ that $\beta(\alpha) < 1$. Thus, the SAS iteration (2.4) converges unconditionally to the solution of (2.1) with respect to norm $\|.\|_{A(\alpha)}$. Since for $u \in D(T)$, we have $\alpha\|u\| \leqslant \|u\|_{A(\alpha)}$, the convergence of the SAS iteration with respect to the norm $\|.\|$ follows. The theoretical optimal parameter α_{opt} for the bound $\beta(\alpha)$ is $\alpha_{opt} = \sqrt{\lambda_{min}\lambda_{max}}$ with λ_{min} and λ_{max} denoting respectively the lower and upper bounds of $\sigma(S)$ [11, 25]. The convergence analysis of the incomplete version of SAS iteration where each subproblem of (2.4) is solved approximately is given in [7].

The following fixed point equation can be derived from the definition of the SAS iteration (2.4):

$$\begin{cases} (\alpha I + S)u_1 = (\alpha I - A)u_2 + q \\ (\alpha I + A)u_2 = (\alpha I - S)u_1 + q \end{cases}. \tag{2.8}$$

Let us define the matrix of operators $\mathbf{T}(\alpha)$ and the vector functions $\mathbf{u}$ and $\mathbf{q}$ as follows:

$$\mathbf{T}(\alpha) = \begin{pmatrix} (\alpha I + S) & -(\alpha I - A) \\ -(\alpha I - S) & (\alpha I + A) \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \text{ and } \mathbf{q} = \begin{pmatrix} q \\ q \end{pmatrix}. \tag{2.9}$$

Therefore, system (2.8) reads

$$\mathbf{T}(\alpha)\mathbf{u} = \mathbf{q}. \tag{2.10}$$

From the m-accretive property of A and the positive definiteness of S it follows that the solution $\mathbf{u}^*$ of the linear operator equation (2.10) exits and is unique in $\mathcal{D}(T) \times \mathcal{D}(T)$. This solution verifies [6]

$$\mathbf{u}^* = (u^*, u^*)^T$$

where u^* is the solution of (2.1). Then it follows that problems (2.1) and (2.10) are equivalent.

Let $\mathbf{P}(\alpha)$ be the matrix operator defined in $\mathcal{D}(T) \times \mathcal{D}(T)$ by

$$\mathbf{P}(\alpha) = \begin{pmatrix} (\alpha I + S) & 0 \\ 0 & (\alpha I + A) \end{pmatrix}. \tag{2.11}$$

Preconditioning of system (2.10) from the right by $[\mathbf{P}(\alpha)]^{-1}$ leads to the following system:

$$\mathbf{T}_1(\alpha)\mathbf{v} = \mathbf{q}_1 \tag{2.12}$$

where $\mathbf{q}_1 = \mathbf{q}$ and $\mathbf{T}_1(\alpha) = \begin{pmatrix} I & -A_1(\alpha) \\ -S_1(\alpha) & I \end{pmatrix}$, with $A_1(\alpha) = (\alpha I - A)(\alpha I + A)^{-1}$ and $S_1(\alpha) = (\alpha I - S)(\alpha I + S)^{-1}$. The operators $A_1(\alpha)$ and $S_1(\alpha)$ satisfy [7]

$$\|A_1(\alpha)\| \leqslant 1 \quad \text{and} \quad \|S_1(\alpha)\| \leqslant \beta(\alpha).$$

The solution $\mathbf{u}^*$ of problem (2.10) reads

$$\mathbf{u}^* = [\mathbf{P}(\alpha)]^{-1}\mathbf{v}^*, \tag{2.13}$$

where $\mathbf{v}^*$ is the solution of (2.12). Since all operators of the matrix $\mathbf{T}_1(\alpha)$ are bounded on H, $\mathbf{T}_1(\alpha)$ is bounded on $H \times H$.

We consider in $H \times H$ the inner product $\langle , \rangle$ defined for $\mathbf{u} = (u_1, u_2)^t$, $\mathbf{v} = (v_1, v_2)^t \in H \times H$ by

$$\langle \mathbf{u}, \mathbf{v} \rangle = (u_1, v_1) + (u_2, v_2), \tag{2.14}$$

and the associated norm

$$\||\mathbf{u}\||^2 = \|u_1\|^2 + \|u_2\|^2. \tag{2.15}$$

The minimal residual iteration method for the solution of problem (2.12) results from the minimization of the residual functional $\varepsilon(\mathbf{u}) = |||\mathbf{q}_1 - \mathbf{T}_1(\alpha)\mathbf{u}|||^2$ [4, 5, 19, 35]. The following estimate on the residual follows from the analysis of this minimal residual algorithm [19]: Given an initial guess $\mathbf{u}^{(0)}$, if the functions $\mathbf{u}^{(k)}$ $(k > 0)$ are computed by the minimal residual algorithm, then

$$\varepsilon(\mathbf{u}^{(k+1)}) \leqslant \lambda_1(\alpha, k)\varepsilon(\mathbf{u}^{(k)}), \tag{2.16}$$

where

$$\lambda_1(\alpha, k) = \left(1 - \frac{\langle \mathbf{r}^{(k)}, \mathbf{T}_1(\alpha)\mathbf{r}^{(k)}\rangle}{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)}\rangle} \frac{\langle \mathbf{r}^{(k)}, \mathbf{T}_1(\alpha)\mathbf{r}^{(k)}\rangle}{\langle \mathbf{T}_1(\alpha)\mathbf{r}^{(k)}, \mathbf{T}_1(\alpha)\mathbf{r}^{(k)}\rangle}\right). \tag{2.17}$$

We have the following convergence results of the minimal residual method for the solution of (2.12) [10]:

Theorem 2.2. *Let α be a positive constant. Given an initial guess $\mathbf{u}^{(0)} \in H \times H$, if the sequence $\left\{\mathbf{u}^{(k)}\right\}_{k\geqslant 0}$ is obtained by the minimal residual algorithm, then the following error estimations hold:*

$$\begin{aligned} \varepsilon(\mathbf{u}^{(k+1)}) &\leqslant \frac{3+\beta(\alpha)}{4}\varepsilon(\mathbf{u}^{(k)}), && (2.18)\\ |||\mathbf{u}^{(k+1)} - \mathbf{u}^*||| &\leqslant \frac{2}{1-\beta(\alpha)}\varepsilon(\mathbf{u}^{(k+1)})^{\frac{1}{2}}. && (2.19) \end{aligned}$$

where $\mathbf{u}^*$ is the exact solution of problem (2.12). Thus, $\left\{\mathbf{u}^{(k)}\right\}_{k\geqslant 0}$ converges to $\mathbf{u}^*$.

3. Preconditioned Minimal Residual Iteration Method

We present in this section two split type preconditioning strategies of problem (2.12). The first strategy is symmetric Gauss-Seidel preconditioning and the second one is coupled symmetric Gauss-Seidel and polynomial preconditioning.

3.1. Symmetric Gauss-Seidel and polynomial preconditioning

Let us consider in $H \times H$ the following operators:

$$\mathbf{M}_1(\alpha) = \begin{pmatrix} I & 0 \\ -S_1(\alpha) & I \end{pmatrix}, \quad \mathbf{M}_2(\alpha) = \begin{pmatrix} I & -A_1(\alpha) \\ 0 & I \end{pmatrix}. \tag{3.1}$$

The operators $\mathbf{M}_1(\alpha)$ and $\mathbf{M}_2(\alpha)$ are bounded and have bounded inverses defined by

$$\mathbf{M}_1^{-1}(\alpha) = \begin{pmatrix} I & 0 \\ S_1(\alpha) & I \end{pmatrix}, \quad \mathbf{M}_2^{-1}(\alpha) = \begin{pmatrix} I & A_1(\alpha) \\ 0 & I \end{pmatrix}. \tag{3.2}$$

The symmetric Gauss-Seidel preconditioner of problem (2.12) is defined by

$$\mathbf{M}_{SGS}(\alpha) = \mathbf{M}_1(\alpha)\mathbf{M}_2(\alpha). \tag{3.3}$$

The split preconditioning of (2.12) using $\mathbf{M}_{SGS}$ leads to the following equivalent problem:

$$\mathbf{T}_2(\alpha)\mathbf{v} = \mathbf{q}_2(\alpha), \tag{3.4}$$

$$\mathbf{u} = \mathbf{M}_2^{-1}(\alpha)\mathbf{v}, \tag{3.5}$$

where

$$\mathbf{T}_2(\alpha) = \mathbf{M}_1^{-1}(\alpha)\mathbf{T}_1(\alpha)\mathbf{M}_2^{-1}(\alpha) = \begin{pmatrix} I & 0 \\ 0 & I - M(\alpha) \end{pmatrix}, \ \mathbf{q}_2(\alpha) = \mathbf{M}_1^{-1}(\alpha)\mathbf{q}_1,$$

with $M(\alpha) = S_1(\alpha)A_1(\alpha)$. The operator of Eq. (3.4) can be written as

$$\mathbf{T}_2(\alpha) = \mathbf{I} - \mathbf{M}(\alpha), \tag{3.6}$$

where $\mathbf{I}$ denotes the identity operator in $H \times H$ and $\mathbf{M}(\alpha) = \begin{pmatrix} 0 & 0 \\ 0 & M(\alpha) \end{pmatrix}$. Since $\|M(\alpha)\| \leqslant \beta(\alpha)$, we have for $\mathbf{u} = (u_1, u_2)^t \in H \times H$

$$|||\mathbf{M}(\alpha)\mathbf{u}|||^2 = \|M(\alpha)u_2\|^2 < \beta^2(\alpha)\|u_2\|^2 < \beta^2(\alpha)|||\mathbf{u}|||^2.$$

Thus,

$$|||\mathbf{M}(\alpha)||| \leqslant \beta(\alpha) < 1, \tag{3.7}$$

and

$$\mathbf{T}_2^{-1}(\alpha) = (\mathbf{I} - \mathbf{M}(\alpha))^{-1} = \sum_{k=0}^{\infty} \mathbf{M}^k(\alpha). \tag{3.8}$$

Therefore, the operator $\mathbf{T}_2^{-1}(\alpha)$ can be approximated by the following truncated Neumann series:

$$\mathbf{P}_n(\alpha) = \sum_{k=0}^{n} \mathbf{M}^k(\alpha). \tag{3.9}$$

Setting

$$\mathbf{T}_2(\alpha, n) = \mathbf{P}_n(\alpha)\mathbf{T}_2(\alpha) = \mathbf{I} - \mathbf{M}^{n+1}(\alpha) \quad \text{and} \quad \mathbf{q}_2(\alpha, n) = \mathbf{P}_n(\alpha)\mathbf{q}_2(\alpha), \tag{3.10}$$

we obtain the following operator equation:

$$\mathbf{T}_2(\alpha, n)\mathbf{u} = \mathbf{q}_2(\alpha, n), \tag{3.11}$$

which is equivalent to (3.4). We have

$$\mathbf{T}_2(\alpha, n) = \mathbf{C}_1^{-1}(\alpha)\mathbf{T}_1(\alpha)\mathbf{M}_2^{-1}(\alpha) \text{ and } \mathbf{q}_2(\alpha, n) = \mathbf{C}_1^{-1}(\alpha)\mathbf{q}_1, \tag{3.12}$$

where $\mathbf{C}_1(\alpha) = \mathbf{P}_n^{-1}(\alpha)\mathbf{M}_1(\alpha)$. Thus, Eq. (3.11) follows from the split preconditioning of Eq. (2.12), using $\mathbf{C}_1(\alpha)\mathbf{M}_2(\alpha)$ as a preconditioner. This can be regarded as a coupled symmetric Gauss-Seidel and polynomial preconditioning of eq. (2.12). It can be noticed that Eq. (3.4) is a particular case of Eq.(3.11), when $n = 0$.

To use the minimal residual algorithm for solving Eq. (3.4), we have to make clear how $\mathbf{T}_2(\alpha, n)$ is computed, since $\mathbf{T}_2(\alpha, n)$ contains some inverse operators. Let $\mathbf{u} = (u_1, u_2)^t \in H \times H$. The components of $\mathbf{T}_2(\alpha)\mathbf{u} = (v_1, v_2)^t$ are expressed as $v_1 = u_1$ and $v_2 = u_2 - M^{n+1}(\alpha)u_2$. The main task is to compute the product $M^{n+1}(\alpha)u_2$ obtained after $n+1$ successive computations of products of the form $M(\alpha)u = \varphi$. Let $u \in H$. We demonstrate in the following how to compute $\varphi = M(\alpha)u$. We have

$$\begin{cases} \varphi_2 = (A - \alpha I)\varphi_1 \\ \varphi = (S - \alpha I)\varphi_3, \end{cases} \tag{3.13}$$

where $\varphi_1 \in \mathcal{D}(T)$ satisfies the differential equation

$$(A+\alpha I)\varphi_1 \;=\; u \tag{3.14}$$

and $\varphi_3 \in H$ satisfies the integral equation

$$(S+\alpha I)\varphi_3 \;=\; \varphi_2. \tag{3.15}$$

Once φ_1, φ_2 and φ_3 have been calculated, the product φ is easy to compute. The differential equation (3.14) and the integral equation (3.15) can be solved numerically.

3.2. Convergence analysis of the preconditioned minimal residual method

We present in this section the convergence results of the minimal residual algorithm of Section 2 applied to (3.11).

The following properties are characteristic of the operator $\mathbf{T}_2(\alpha, n)$:

Theorem 3.1. *Let α be a positive constant. For all $\mathbf{u} \in H\times H$, the following inequalities hold true:*

$$\begin{aligned}\langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{u}\rangle &\geqslant (1-\beta^{n+1}(\alpha))|||\mathbf{u}|||^2, &(3.16)\\ \langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{u}\rangle &\geqslant \frac{1}{2}\langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{T}_2(\alpha,n)\mathbf{u}\rangle. &(3.17)\end{aligned}$$

Proof. Let $\mathbf{u} = (u_1, u_2)^t$. We have

$$\begin{aligned}\langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{u}\rangle &= \|u_1\|^2+\|u_2\|^2-(M^{n+1}(\alpha)u_2,u_2)\\ &\geqslant \|u_1\|^2+\|u_2\|^2-\|M^{n+1}(\alpha)\|\|u_2\|^2\\ &\geqslant |||\mathbf{u}|||^2-\|M^{n+1}(\alpha)\||||\mathbf{u}|||^2\\ &\geqslant (1-\beta^{n+1}(\alpha))|||\mathbf{u}|||^2.\end{aligned}$$

We also have

$$\begin{aligned}\langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{u}\rangle-\tfrac{1}{2}\langle \mathbf{T}_2(\alpha,n)\mathbf{u},\mathbf{T}_2(\alpha,n)\mathbf{u}\rangle &= \tfrac{1}{2}|||\mathbf{u}|||^2-\tfrac{1}{2}\|M^{n+1}(\alpha)u_1\|^2\\ &\geqslant \tfrac{(1-\beta^{2n+2}(\alpha))}{2}|||\mathbf{u}|||^2 \geqslant 0.\end{aligned}$$

Thus, the inequality (3.17) is satisfied. □

Theorem 3.2. *Convergence results.*

Let α be a positive constant. Given an initial guess $\mathbf{u}^{(0)} \in H\times H$, if the sequence $\{\mathbf{u}^{(k)}\}_{k\geqslant 0}$ for the approximation of the solution u^* of (3.11) is obtained by the minimal residual algorithm, then the following error estimates hold:

$$\begin{aligned}\varepsilon(\mathbf{u}^{(k+1)}) &\leqslant \frac{1+\beta^{n+1}(\alpha)}{2}\varepsilon(\mathbf{u}^{(k)}), &(3.18)\\ |||\mathbf{u}^{(k+1)}-\mathbf{u}^*||| &\leqslant \frac{1}{1-\beta^{n+1}(\alpha)}\varepsilon(\mathbf{u}^{(k+1)})^{\frac{1}{2}}. &(3.19)\end{aligned}$$

where $\mathbf{u}^*$ is the exact solution of problem (3.11). Thus $\{\mathbf{u}^{(k)}\}_{k\geqslant 0}$ converges to $\mathbf{u}^*$.

Proof. Replacing in (2.17) $\mathbf{T}_1(\alpha)$ by $\mathbf{T}_2(\alpha, n)$, we deduce from inequalities (3.16) and (3.17) the following bound for the residual decrease rate:

$$\lambda_3(\alpha, k) \leqslant \frac{1 + \beta^{n+1}(\alpha)}{2}, \quad k \geqslant 0$$

and inequality (3.18) follows from (2.16).

Replacing in (3.16) $\mathbf{u}$ by $(\mathbf{u}^{(k+1)} - \mathbf{u}^*)$ yields

$$\begin{aligned} |||\mathbf{u}^{(k+1)} - \mathbf{u}^*|||^2 &\leqslant \tfrac{1}{1-\beta^{n+1}(\alpha)} \langle \mathbf{T}_2(\alpha, n)\mathbf{u}^{(k+1)} - \mathbf{q_2}(\alpha, n), \mathbf{u}^{(k+1)} - \mathbf{u}^* \rangle \\ &\leqslant \tfrac{1}{1-\beta^{n+1}(\alpha)} \left(\varepsilon(u^{(k+1)}\right)^{\frac{1}{2}} .|||\mathbf{u}^{(k+1)} - \mathbf{u}^*|||, \end{aligned}$$

and the estimate (3.19) then follows. □

Let $\nu_1(\alpha)$ and $\nu_2(\alpha, n)$ denote the upper bounds for the residual decrease rate of the minimal residual solver applied respectively to Eqs. (2.12) and (3.11). It follows from Theorem 2.2 and Theorem 3.2 that for $\alpha > 0$ and $n \geqslant 0$,

$$\nu_1(\alpha) - \nu_2(\alpha, n) = \frac{(1 - \beta^{n+1}(\alpha)) + \beta(\alpha)(1 - \beta^n(\alpha))}{4} > 0. \tag{3.20}$$

It follows that $\nu_1(\alpha) > \nu_2(\alpha, n)$ and the preconditioned minimal residual solver is theoretically faster than the minimal residual solver. Moreover, $\nu_2(\alpha, n) > \nu_2(\alpha, m)$ for $m > n$. Thus, the convergence is theoretically faster and faster with increasing n. Figure 3.1 plots the estimates of the upper bounds for the residual decrease rate as a function of $\beta(\alpha)$ of the minimal residual method and its preconditioned versions for several values of n.

Remark 3.1. Since we focus on the solution of Eq. (2.1), for the computational purpose we need only to solve the second sub-equation of problem (3.11) which reads

$$(I - M^{n+1}(\alpha))v = q_2, \tag{3.21}$$

where q_2 is the second component of the vector $\mathbf{q_2}(\alpha, n)$. The solution u^* of (2.1) is computed from the solution v^* of (3.21) as follows:

$$u^* = (\alpha I + A)^{-1} v^*. \tag{3.22}$$

Proceeding similarly as in the proof of Theorem 3.1 and Theorem 3.2, we have the following convergence results of the minimal residual method applied to (3.21):

$$\|q_2 - (I - M^{n+1}(\alpha))v^{(k+1)}\|^2 \leqslant \frac{1 + \beta^{n+1}(\alpha)}{2} \|q_2 - (I - M^{n+1}(\alpha))v^{(k)}\|^2, \tag{3.23}$$

$$\|v^{(k+1)} - v^*\| \leqslant \frac{1}{1 - \beta^{n+1}(\alpha)} \|q_2 - (I - M^{n+1}(\alpha))v^{(k+1)}\|. \tag{3.24}$$

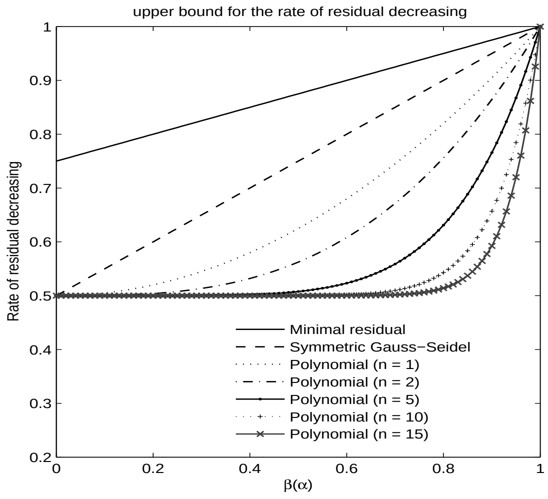


Fig. 3.1. Comparison of upper bounds for the residual decrease rate of the methods

4. Numerical Results for the 2-D Neutron Transport Equation

We apply the above minimal residual algorithm for solving the neutron transport equation in 2-D Cartesian geometry.

4.1. The 2-D Neutron Transport Equation

Consider the following single-group steady-state first-order neutron transport equation in 2-D Cartesian geometry [4, 7, 10]:

$$\left\{\begin{array}{l} Tu(r,\omega) := Au(r,\omega) + \Sigma u(r,\omega) - Ku(r,\omega) = q(r,\omega), \quad (r,\omega) \in R \times B \\ u(r,\omega) \in D(T), \end{array}\right. \tag{4.1}$$

where the operators A, Σ and K are defined by

$$Au = \omega \nabla_r u; \quad \Sigma u = \sigma(r) u; \quad Ku = \int_B \kappa(r,\omega,\omega') u(r,\omega') d\omega' \tag{4.2}$$

and

$$D(T) = \{u \in L^2(Q) : \omega \nabla_r u \in L^2(Q) \text{ and } u_{|\Gamma_-} = 0\}, \tag{4.3}$$

where $Q = R \times B$, $\Gamma_- = \{(r,\omega) \in \partial R \times B : \mu n_x + \eta n_y < 0\}$ ($n = (n_x, n_y)$ being the outer normal to ∂R), $r = (x,y)$, $\omega = (\mu,\eta)$, $R =]0,1[\times]0,1[$ and $B = \{\omega \in \mathbb{R}^2 : |\omega| < 1\}$. The function $\sigma(r)$ represents the total cross section and $\kappa(r,\omega,\omega')$ is a nonnegative kernel describing the scattering of particles. The function q is a nonnegative source term.

The operator A is m-accretive [14]. In the following it is assumed that:

$(a1)$ $\sigma \in L^\infty(Q)$ and $\exists \sigma_0 > 0$ such that $\sigma(r) \geqslant \sigma_0$ a.e on $R \times B$;

$(a2)$ $\kappa(r,\omega,\omega') = \kappa(x,\omega',\omega)$ and κ is non negative;

$(a3)$ $\exists c \in [0,1), \displaystyle\int_B \kappa(r,\omega,\omega') d\omega' \leqslant \sigma_0 c$ a.e on $R \times B$.

Therefore, the operator $S = \sigma I - K$ is self-adjoint and positive definite [14]. Thus, T is positive definite and the existence and uniqueness of the solution of problem (4.1) follows. Moreover, T admits a self-adjoint positive definite and m-accretive splitting (SAS) which yields the SAS iteration method. The SAS iteration method, the minimal residual method and its preceding preconditioned versions for the solution of Eq. (4.1) converge. Equation (4.1) is known to be near singular when $c \approx 1$ [14].

In the case of isotropic scattering where the integral operator is defined by

$$K\psi = \sigma_s(x)Pu, \tag{4.4}$$

with

$$Pu = \frac{1}{\pi}\int\limits_B u(x,\Omega')d\Omega',$$

the inverse of the operator $(\alpha I + S)$ is given by [7]

$$(\alpha I + S)^{-1} = \frac{1}{\sigma(x) - \sigma_s(x) + \alpha}P + \frac{1}{\sigma(x) + \alpha}(I - P). \tag{4.5}$$

Therefore, the linear integral equation (3.15) can be solved explicitly. Moreover, the function φ in (3.13) can be calculated as follows:

$$\varphi = P_1\varphi_2, \tag{4.6}$$

where

$$P_1 = \left(\frac{\alpha - \sigma - \sigma_s}{\alpha + \sigma - \sigma_s} - \frac{\alpha - \sigma}{\alpha + \sigma}\right)P + \frac{\alpha - \sigma}{\alpha + \sigma}I. \tag{4.7}$$

Here, σ_s and $\sigma_a = \sigma - \sigma_s$ denote the scattering and the absorption cross sections respectively. The scattering ratio and the optical coefficient are defined respectively as follows:

$$\gamma = \max_{x\in R}\left(\frac{\sigma_s(x)}{\sigma_s(x) + \sigma_a(x)}\right) \quad \text{and} \quad \nu = \min_{x\in R}\left(\sigma_s(x) + \sigma_a(x)\right)\operatorname{diam}(R), \tag{4.8}$$

where diam(R) denotes the diameter of the domain R. The values $\gamma = 1$ and $\nu >> 1$ ($\sigma_a >> 1$) correspond to the pure scattering and optically thick domains, respectively, and represent two extreme situations in the computational transport where conventional discretization methods such as piecewise linear finite elements using the Galerkin formulation [23], the classical finite difference scheme [16] and the upwind difference scheme [22] yield inaccurate solutions unless the spatial grid is very fine. As mentioned in [28, 21], as $\sigma_t = \sigma_a + \sigma_s$ tends to infinity and γ tends to 1, the problem becomes singularly perturbed. Therefore, the discrete approximation of the transport problem using these methods will have operators with condition numbers of the order of at least σ_t^2 regardless of the mesh size [21].

4.2. Discretization and Numerical Results

Discretization is carried out by the discrete ordinates and Diamond difference schemes [7, 14]. For the angular discretization a set of L discrete angular directions $\Omega_L = \{\omega_i = (\mu_i, \eta_i),\ 1 \leqslant i \leqslant L\} \subset B$ is used. The set Ω_L satisfies for all $(\mu, \eta) \in \Omega_L$: a) $\mu \neq 0$ and $\eta \neq 0$; b) $(-\mu, -\eta) \in \Omega_L$. A finite difference method based on volume control and cell averaging is considered for the space discretization. The numerical grid is defined by

$$R_h = \{(x_i, y_j), 0 \leqslant i \leqslant N, 0 \leqslant i \leqslant M\}, \tag{4.9}$$

where $x_0 = 0$, $x_i = x_{i-1} + (\Delta x)_i$, $x_N = 1$, $y_0 = 0$, $y_j = y_{j-1} + (\Delta y)_j$, $y_M = 1$ and $h = \max\limits_{ij} ((\Delta x)_i, (\Delta y)_j)$. The cell center grid points are defined as:

$$x_{i+\frac{1}{2}} = \frac{x_{i+1} - x_i}{2},\ y_{j+\frac{1}{2}} = \frac{y_{j+1} - y_j}{2},\ (\Delta x)_{i+\frac{1}{2}} = x_{i+1} - x_i \text{ and } (\Delta y)_{j+\frac{1}{2}} = y_{j+1} - y_j.$$

Therefore, Eqs. (3.14) and (3.15) can be solved using respectively the direct sweeping algorithm [7, 14] and the conjugate gradient method in the anisotropic case [7].

For the numerical results, we took particular data for which an exact solution of problem (4.1) is known in each case. For the iterative methods tested here, the iterations are stopped when the relative error $\|U - U_{exact}\|_2 / \|U_{exact}\|_2$ is less than the prescribed $\epsilon > 0$.

For $x = (x_1, x_2) \in R$ and $\Omega = (\mu, \eta) \in B$, we set $\sigma(x) = \sigma$, $\kappa(x, \Omega, \Omega') = \frac{\sigma c}{\pi}$ and

$$q(x,\mu) = \begin{cases} \mu x_2 + \eta x_1 + \sigma x_1 x_2 - \frac{\sigma c}{4}, & \mu > 0, \eta > 0; \\ -\mu x_2 + \eta(1 - x_1) + \sigma(1 - x_1)x_2 - \frac{\sigma c}{4}, & \mu < 0, \eta > 0; \\ -\mu(1 - x_2) - \eta(1 - x_1) + \sigma(1 - x_1)(1 - x_2) - \frac{\sigma c}{4}, & \mu < 0, \eta < 0; \\ \mu(1 - x_2) - \eta x_1 + \sigma x_1(1 - x_2) - \frac{\sigma c}{4}, & \mu > 0, \eta < 0. \end{cases}$$

The exact solution of this test problem is given by

$$\psi(x,\mu) = \begin{cases} x_1 x_2, & \mu > 0, \eta > 0; \\ (1 - x_1)x_2, & \mu < 0, \eta > 0; \\ (1 - x_1)(1 - x_2), & \mu < 0, \eta < 0; \\ x_1(1 - x_2), & \mu > 0, \eta < 0. \end{cases}$$

In this problem, c is the scattering ratio and σ is the optical coefficient. The quantities $\sigma_s = \sigma c$ and $\sigma_a = \sigma(1 - c)$ are respectively the scattering and absorption cross sections of neutrons in R.

For the numerical test, we take $\Delta x = \Delta y = \frac{1}{10}$ and L=100. We study the behavior of the preconditioned minimal residual methods with respect to the parameters σ, c and α. For the exemplary problem, the theoretical optimal parameter minimizing the bound $\beta(\alpha)$ is $\alpha_t = \sigma_a = \sigma(1 - c)$. It was observed in [7] that for fixed c and σ the optimal numerical value of α can be localized in the interval $[\sigma(1 - c), \sigma(1 - c/2)]$. The value of $\alpha*$ given in [10] yielded good convergence results for the SAS iteration applied to the exemplary problem as compared to the standard source iteration method, the spatial multigrid algorithm and some Krylov subspace methods such as GMRES and BiCGStab iterative algorithms [7]. It was observed in [10] that for some values of σ_a and c ($\sigma_a \geqslant 4$ and $c > 0.5$.) the convergence of the minimal residual seemed to be faster as compared to the SAS method using α_t. This result also holds for large values of σ in using the SAS method with $\alpha*$. The Gauss-Seidel preconditioned version of the minimal residual algorithm gave excellent results compared to SAS and its successive overrelaxation acceleration [10].

We present comparative numerical results (the number of iterations and the CPU time in s) of the previous minimal residual algorithm with: symmetric Gauss-Seidel preconditioning (SGS-Minres), polynomial preconditioning (PMinres[n]) with n denoting the order of the truncated Neumann series, and SAS iterations using $\alpha = \alpha_*$ and $\alpha = \alpha_t + c$. There are two sets of tests: one at fixed σ and the other at fixed c. As shown in Figs. 4.1 to 4.7, all the methods converge. At fixed $\sigma = 50$ and $\sigma = 100$, we compare the $c-$dependence of the iterative methods used here. Figs. 4.1 and 4.2 plot the number of iterations and the CPU time of the methods as a function of c, respectively, for $c \in [0.1, 0.99]$ and $c \in [0.98, 0.99999]$ with $\sigma = 50$, using $\alpha = \alpha_*$. We observe that the PMinres[n] iterations ($n = 1, 2, 10$] are faster than the SGS-Minres one which is faster than the SAS, particularly for values of c closed to 1 (Fig. 4.2). As can be seen from Figs. 4.3 and 4.4, these observations remain true in performing the same tests for $\sigma = 100$. Next, we compare the $\sigma-$dependence of the iterative methods. Figures 4.5 and 4.6 plot the number of iterations of the methods and the CPU time as a function of the total cross section ($\sigma \in [1, 100]$) at fixed $c = 0.5$ and $c = 0.99$ respectively. We can see that the PMinres[n] method is still more efficient than the SGS-Minres one which is faster than the SAS method. The same observations hold for large values of σ at $c = 0.99$ (see Fig. 4.7) and for the critical case where $c = 1$ (see Table 4.1). We set $\alpha = \sigma(1 - c) + c$. Table 4.2 and Table 4.3 present comparative numerical results of the methods for $1 \leqslant \sigma \leqslant 1000$ at fixed $c = 0.98$ and for $0.98 \leqslant c \leqslant 1$ at fixed $\sigma = 5$, respectively. The SAS method remains slower than the preconditioned minimal residual methods. We also remark that for $\alpha = \alpha^*$, PMinres[n] is more and more efficient with increasing n. This confirms the theoretical convergence results obtained.

We now consider another set of tests where the mesh size decreases : $\Delta x = \Delta y = h$, with $h \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\}$. At fixed $\sigma = 100$, we test the behavior of the methods with decreasing mesh size, for $c = 0.5$ and $c = 0.99$. The Table 4.4 presents the number of iterations of each method for $c = 0.5$. It can be observed that for each method the number of iterations remains roughly constant with decreasing mesh size. For $c = 0.99$, we set $\alpha = \sigma(1 - 23c/32)$. Table 4.5 presents the number of iterations for each method. The tested methods converge for a mesh size less than $\frac{1}{64}$. For $h = \frac{1}{128}$, a convergence of SGS-Minres and PMinres[1] is noted. For the other methods, we observe a divergence of the SAS method at the second iteration and a stagnation of the residual for Pminres[2] and Pminres[10]. This drawback is essentially due to the fact that the discretization method applied to the first subproblem of the SAS iteration generates a negative flux for fixed values of σ, c, and α. This drawback is overcome by setting $\alpha = \sigma$. Table 4.6 gives the number of iterations of the methods tested here. It can be observed that for each method the number of iterations is roughly constant for $h \geqslant \frac{1}{16}$ and the preconditioned Minres methods accelerate the SAS iterations. The convergence behavior (relative error as a function of the iteration) of the SAS, SGS-Minres, PMinres[n] ($n = 1, 2, 10$) is plotted in Fig. 4.8 for $h = \frac{1}{128}$, $\sigma = 100$ and $c = 1$ with $\alpha = \sigma$. The efficiency of the preconditioned minimal residual methods can be observed.

Additionally, we present comparative convergence behaviors of the preconditioned Minres methods and the spatial multigrid method using the bi-conjugate gradient stabilized method as a smoothing method MG(n_1,n_2), with n_1 and n_2 denoting the number of pre-smoothing and post-smoothing steps, respectively. Iterations are stopped when the relative residual error $\|B - GU\|/\|B\|$ is less than $1E - 05$, where G and B denote respectively, the matrix and the right hand side of the discrete system . The convergence history of the multigrid and the SGS-Minres methods for $\sigma = 100$ and $c = 0.5$ with $h = \frac{1}{16}$ is plotted in Fig. 4.9. It can be seen that the MG(1,1) method diverges and the MG(20,20) method converges but is less efficient than the SGS-Minres method. We set $c = 0.99$. It is seen from Fig. 4.10 that for the mesh size $h = \frac{1}{16}$, the MG(1,1) method diverges and the preconditioned Minres methods are efficient compared to the MG(20,20) method. Figure 4.11 plots the convergence behavior of the methods tested here for the mesh size $h = \frac{1}{32}$ at fixed $c = 0.99$ and $\sigma = 100$. Divergence of the MG(1,1) and MG(20,20) methods can be observed. We note the efficiency of the preconditioned Minres methods compared to the spatial multigrid methods considered.

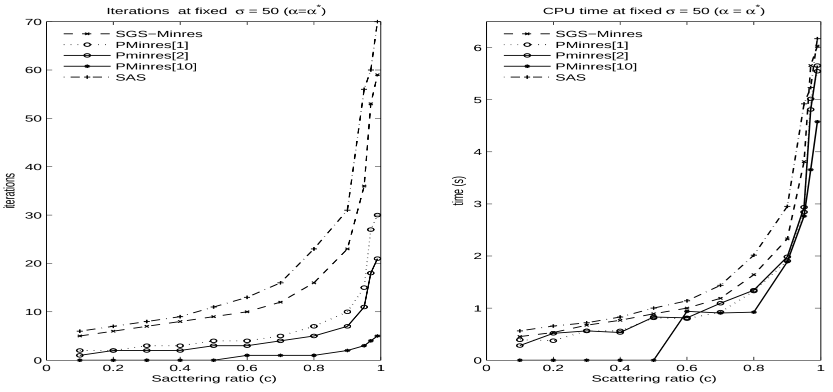


Fig. 4.1. Comparison of the methods at fixed $\sigma = 50$, for $c \in [0.1, 0.99]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

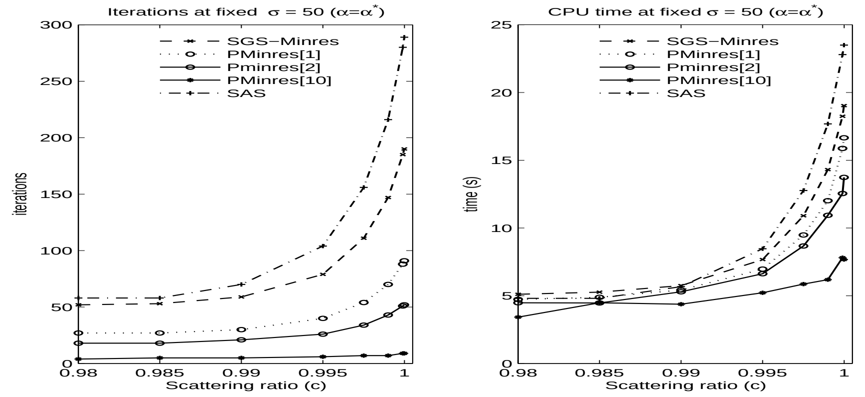


Fig. 4.2. Comparison of the methods at fixed $\sigma = 50$, for $c \in [0.98, 0.99999]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

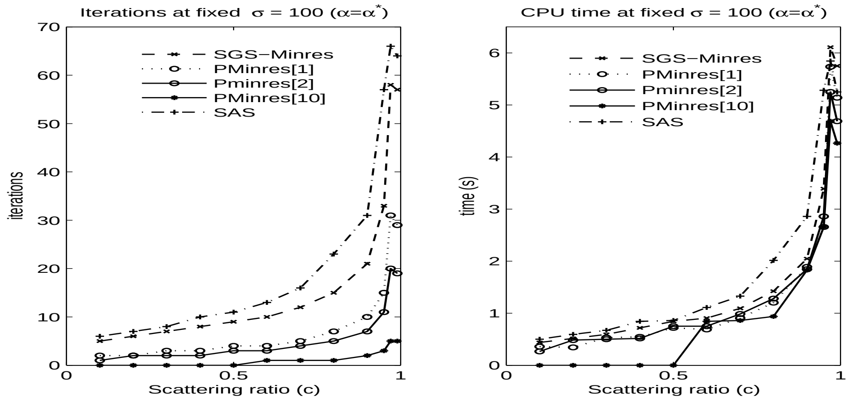


Fig. 4.3. Comparison of the methods at fixed $\sigma = 100$, for $c \in [0.1, 0.99]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

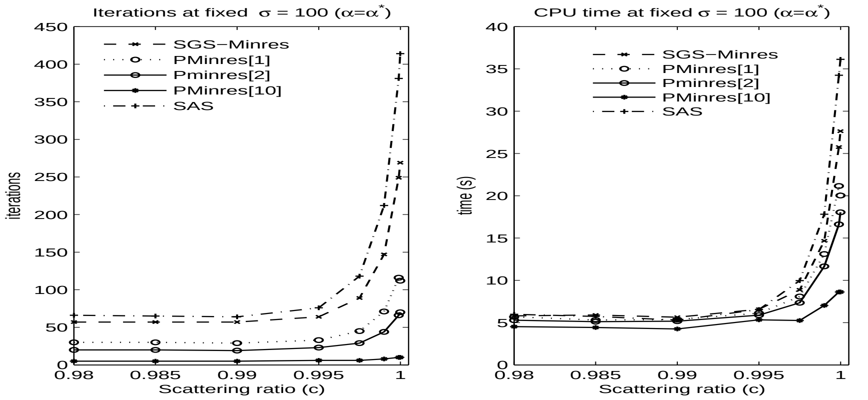


Fig. 4.4. Comparison of the methods at fixed $\sigma = 100$, for $c \in [0.98, 0.99999]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

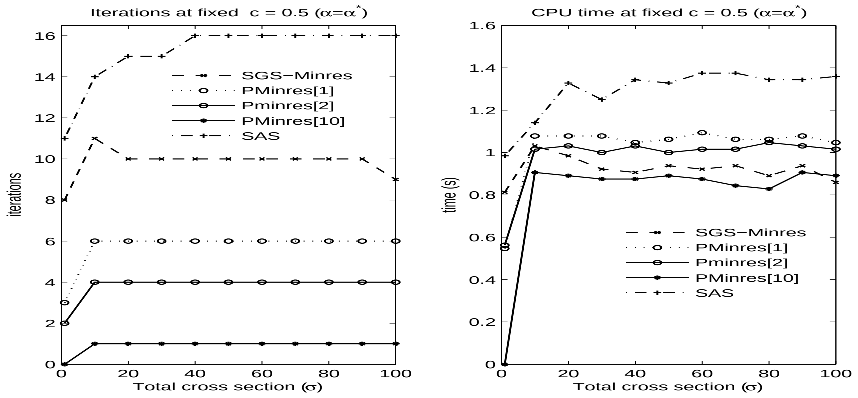


Fig. 4.5. Comparison of the methods at fixed $c = 0.5$, for $\sigma \in [1, 100]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

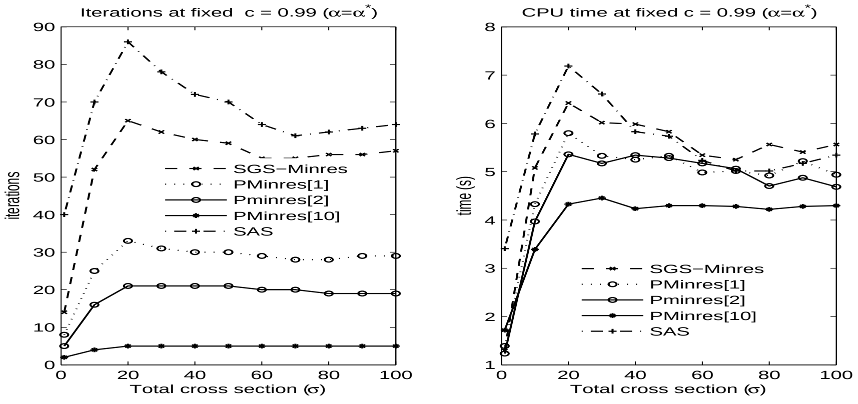


Fig. 4.6. Comparison of the methods at fixed $c = 0.99$, for $\sigma \in [1, 100]$ $(\epsilon = 1E - 08)$: (left) number of iterations;(right) CPU time in s

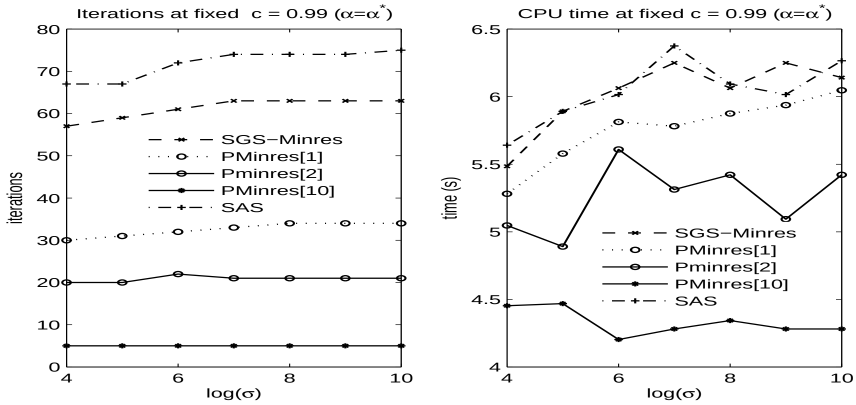


Fig. 4.7. Comparison of the methods at fixed $c = 0.99$, for large σ ($\epsilon = 1E - 08$): (left) number of iterations;(right) CPU time in s

Table 4.1. **Number of iterations and CPU time in s (in brackets) for** $c = 1$ ($\alpha = 1$, $\epsilon = 1E - 08$)

σ	1	10	50	100	500	1000
SAS	42(3.97)	86(7.16)	290(24.45)	418(34.95)	648(55.11)	696(58.03)
SGS-Minres	14(1.30)	61(6.27)	191(19.97)	271(28.22)	416(40.59)	446(44.41)
PMinres[1]	8(1.41)	28(4.95)	91(15.92)	112(20.06)	193(34.34)	204(35.42)
PMinres[2]	5(1.22)	19(4.80)	53(13.23)	71(17.69)	105(26.37)	112(28.86)
PMinres[10]	2(1.78)	4(3.37)	9(7.75)	10(8.45)	12(10.37)	12(10.16)

Table 4.2. **Number of iterations and CPU time in s (in brackets) for** $c = 0.98$ ($\alpha = \sigma(1 - c) + c$, $\epsilon = 1E - 08$)

σ	1	10	50	100	500	1000
SAS	38(3.53)	46(4.11)	43(3.70)	59(4.85)	185(15.97)	276(23.69)
SGS-Minres	14(1.44)	39(4.06)	42(4.17)	40(3.92)	24(2.28)	19(1.89)
PMinres[1]	8(1.56)	18(3.44)	21(3.61)	27(5.01)	70(12.06)	95(16.94)
PMinres[2]	5(1.36)	12(3.44)	14(3.48)	18(4.48)	23(5.87)	24(6.23)
PMinres[10]	2(1.98)	3(2.90)	3(2.59)	5(4.31)	15(12.80)	18(16.56)

Table 4.3. **Number of iterations and CPU time in s (in brackets) for** $\sigma = 5$ ($\alpha = \sigma(1 - c) + c$, $\epsilon = 1E - 08$)

c	0.98	0.99	0.995	0.9975	0.999	0.9999	1
SAS	39(3.72)	42(4.62)	49(5.61)	56(6.26)	60(6.56)	62(6.92)	63(6.80)
SGS-Minres	31(3.31)	33(4.20)	35(4.70)	36(4.70)	36(4.81)	36(4.73)	37(4.94)
PMinres[1]	14(2.81)	15(3.62)	17(4.20)	18(4.33)	19(4.39)	20(4.80)	20(4.87)
PMinres[2]	9(2.55)	10(3.31)	11(3.69)	11(3.72)	11(3.67)	11(3.64)	11(3.69)
PMinres[10]	3(2.83)	3(3.58)	3(3.39)	3(3.44)	3(3.47)	3(3.55)	3(3.43)

Table 4.4. **Number of iterations for** $\sigma = 100$, $c = 0.5$ ($\alpha = \alpha^*$, $\epsilon = 1E-06$)

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
SAS	8	8	8	8	8	8
SGS-Minres	6	6	7	7	7	7
PMinres[1]	3	3	3	3	3	3
PMinres[2]	2	2	2	2	2	2
PMinres[10]	0	0	0	0	0	0

Table 4.5. **Number of iterations for** $\sigma = 100$, $c = 0.99$ ($\alpha = \sigma(1 - 23c/32)$, $\epsilon = 1E-06$)

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
SAS	182	184	186	186	186	-
SGS-Minres	49	78	91	99	128	205
PMinres[1]	26	32	35	37	46	88
PMinres[2]	18	19	20	21	24	-
PMinres[10]	6	6	6	7	7	-

Table 4.6. **Number of iterations for** $\sigma = 100$, $c = 0.99$ ($\alpha = \sigma$, $\epsilon = 1E-06$)

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
SAS	633	637	640	641	642	643
SGS-Minres	20	34	46	51	49	53
PMinres[1]	7	21	32	32	33	33
PMinres[2]	7	18	26	26	26	26
PMinres[10]	7	13	14	14	13	13

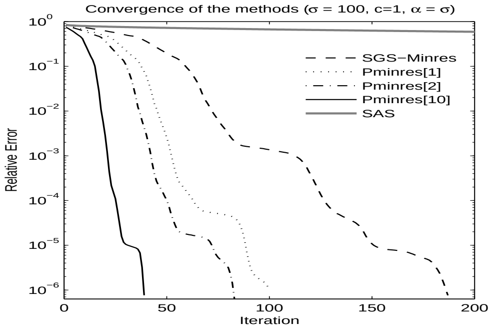


Fig. 4.8. Convergence behavior of the Preconditioned Minres and SAS Methods for $h = \frac{1}{128}$, $\sigma = 100$ and $c = 1$ with $\alpha = \sigma$

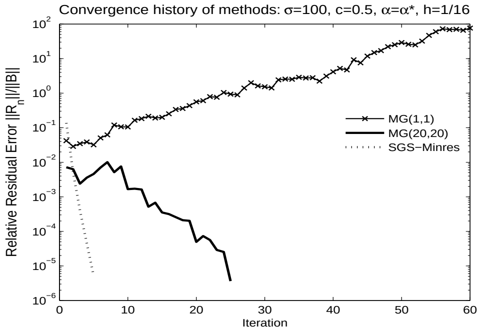


Fig. 4.9. Convergence behavior of the Multigrid and SGS-Minres Methods for $h = \frac{1}{16}$, $\sigma = 100$ and $c = 0.5$ with $\alpha = \alpha*$

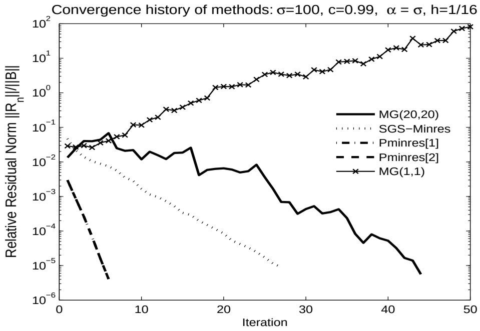


Fig. 4.10. Convergence behavior of the Multigrid and Preconditioned Minres Methods for $h = \frac{1}{16}$, $\sigma = 100$ and $c = 0.99$ with $\alpha = \sigma$

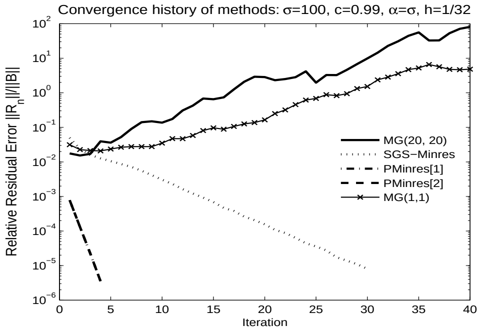


Fig. 4.11. Convergence behavior of the Multigrid and Preconditioned Minres Methods for $h = \frac{1}{32}$, $\sigma = 100$ and $c = 0.99$ with $\alpha = \sigma$

5. Conclusions

We have presented a symmetric Gauss-Seidel and polynomial preconditioning of a minimal residual method for solving a class of linear operator equations, with a positive definite operator admitting self-adjoint and m-accretive splitting in a Hilbert space H. Theoretical analysis shows that these methods converge unconditionally to the solution of the equation. Theoretical proof of the convergence of methods is independent of the discretization. Previous numerical results illustrate the feasibility and efficiency of these methods in solving a 2-D neutron transport problem. The methods converge for critical cases (c close to 1 and/or large σ). Moreover, the above preconditioned Minres methods give better results than the SAS iteration method does.

Acknowledgments. The authors thank the referees for their careful reading of the manuscript and valuable comments. The second author gratefully acknowledges the partial support of this work by the African Millennium Mathematical Science Initiative (AMMSI).

References

1. N. G. Abrashina-Zhadaeva and A. A. Egorov, *Multicomponent iterative methods solving stationary problems of mathematical physics* Mathematical Modelling and Analysis, **13** (2008), no. 3, pp. 313–326.

2. M. L. Adams and E. W. Larsen, *Fast Iterative Methods for Discrete-Ordinates Particle Transport Calculation*, Prog. Nucl. Energy, **40** (2002), pp. 3–159.

3. A. B. Antonevich, J. Appell, V. A. Prokhorov, and P. P. Zabrejko, *Quasi-iteration methods of Chebyshev type for the approximate solution of operator equations*, REND. SEM. MAT. UNIV. PADOVA, **93** (1995), pp. 127–141.

4. S. Akesbi and E. Maître, *Theoretical and numerical analysis of a minimal residual solver for 2d Boltzmann equation*, Journal of Computation and Applied Mathematics, **150** (2003), no. 2, pp. 357-374.

5. S. Akesbi and E. Maître, *Minimal residual method applied to the transport equation*, Journal of Numerical Algorithms, **26** (2001), pp. 235-249.

6. O. Awono and J. Tagoudjeu, *Iterative methods for a class of linear operator equations* Int. J. Contemp. Math. Sci., **4** (2009), no. 12, pp. 549–564.

7. Awono Onana and J. Tagoudjeu, *A splitting iterative method for solving the neutron transport equation*, *Mathematical Modelling and Analysis*, **14** (2009), no. 3, pp. 271–289.

8. O. Awono and J. Tagoudjeu, *A Self-adjoint and m-accretive splitting iterative method for solving the neutron transport equation in 1-D sphérical geometry*, in: Proceeding of the 9^{th} African Conference on Research in Computer Science and Applied Mathematics, Morocco, (2008), pp. 331-338.

9. O. Awono and J. Tagoudjeu, *A SOR acceleration of self-Adjoint and m-Accretive splitting iterative solver for 2-D neutron transport equation*, in: Proceeding of the 9^{th} International Conference JANO'9, Mohammedia-Morocco, (2008), pp. 318-321.

10. O. Awono and J. Tagoudjeu, *A Minimal residual solver for the neutron transport equation*, Int. J. Contemp. Math. Sci., **4** (2009), no. 34, pp. 1671–1684.

11. Z.-Z. Bai, G.H. Golub, and M.K. NG, *On successive overrelaxation of hermitian and skew-hermitian splitting methods for non-hermitian positive definite linear systems*, Technical Report, SCCM02-06, Stanford University, 2002.

12. H. Brezis, *Analyse Fonctionnelle. Theorie et Application*, 2ème tirage, Masson, Paris, 1987.

13. B. Chang and B. Lee, *A multigrid algorithm for solving the multi-group anisotropic scattering Boltzmann equation using first-order system least-squares methodology*, Electronic Transactions on Numerical Analysis, **15** (2003), pp. 132-151.

14. R. Dautray and J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Tome 3, Masson, Paris, 1985.

15. M. G. Gasparo, A. Papini, and A. Pasquali, *Some properties of GMRES in Hilbert spaces*, Numer. Funct. Anal. Optim., **29** (2008), no. 11-12, pp. 1276–1285.

16. E. W. Larsen and J. E. Morel, Advances in Discrete-Ordinates Methodology: *in Nuclear Computational Science: A century of Review*, Eds. Y. Amzi and E. Sartori, Springer Berlin, 2006.

17. L. Knizhnerman, *On GMRES-Equivalent Bounded Operator*, SIAM J. Matrix Anal. Appl., **23** (2000), no. 1, pp. 195–212.

18. M.A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitskii, and V. Ya. Stetsenko, *Approximate solution of operator problem*, Wolters-Noordhoff Publishing, Groningen, 1972.

19. P. Lascaux and R. Théodor, *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur*, Volume 2, Masson, Paris, 1987.

20. P. Malits, *Certain approximate methods for solving linear operator equations*, Appl. Math. Lett., **20** (2007), no. 3, pp. 306–311.

21. T. A. Manteuffel, S. McCormick, J. E. Morel, S. Oliveira, and G. Yang, *A Parallel Version of A multigrid algorithm for isotropic transport Equations.* SIAM Journal on Scientific Computing, **15** (1994), no. 2, pp. 474–493.

22. T. A. Manteuffel, S. McCormick, J. E. Morel, and G. Yang, *A fast multigrid algorithm for isotropic transport problems ii. with absorption.* SIAM Journal on Scientific Computing, **17** (1996), pp. 1449–1474.

23. T. A. Manteuffel and K. Ressel, *A Systematic Solution Approach for Neutron Transport Problems in Diffusive Regimes,* in Seventh Copper Mountain Conference on Multigrid Methods , N. D. Melson, T. A. Manteuffel, S. McCormick, and C. C. Douglas, eds., NASA Hampton, VA, (1996), pp. 519–534.

24. T. A. Manteuffel and K. Ressel, *Least-squares finite element solution for the neutronic transport equation in diffusive regimes.* SIAM J. Numer. Anal., 1998.

25. G. Marchuk, *Decomposition methods*, Nauka, Moscou, 1988, in Russian.

26. G. Marchuk and V. Agochkov, *Introduction aux Méthodes des Eléments Finis*, Mir, Moscou, 1985.

27. O. Nevanlinna, *Convergence of Krylov methods for sums of two operators*, BIT, **36** (1996), no. 4, pp. 775–785.

28. S. Oliveira and Y. Deng, *Preconditioned Krylov subspace methods for transport equations.* Prog. Nucl. Energy, **33** (1998), no. 1/2, pp. 155–174.

29. B. W. Patton and J. P. Holloway, *Application of Preconditioned GMRES to the numerical solution of the neutron transport equation.* Ann. Nucl. Energy, **29** (2002), no. 2, pp. 109–136

30. A. G. Ramm, *Iterative solution of linear equations with unbounded operator*, J. Math. Anal. Appl., **330** (2007), pp. 1338-1346.

31. M. Seaïd, *A note on numerical methods for two-dimensional neutron transport equation*, Technical Report Nr. 2332, TU Darmstadt, 2002.

32. F. Schäpfer, A. K. Louis, and T. Schuster, *Nonlinear iterative methods for linear ill-posed problems in banach spaces*, Inverse Problems, **22** (2006), pp. 311–329.

33. F. D. Swesty, D. C. Smolarski and P. E. Saylor, *A comparison of algorithms for the efficient solution of the linear systems arising from multigroup flux-limited diffusion problems.* The Astrophysical Journal Supplement Series, **182** (2004), pp. 369–387.

34. A. Tizaoui, *Splitting operator for solving the neutron transport equation in 1-D spherical geometry,* International Journal of Mathematics and Statistics, **1** (2007), no. A07, pp. 31-45.

35. A. Tizaoui, *Polynomial Preconditioning and the Generalized Minimal Residual Algorithm Solver for the 2-D Boltzmann Transport Equation*, C. R. Acad. Sci. Paris, Ser. I, **345** (2007), pp. 178-181.

36. V.Trenoguine, *Analyse Fontionnelle*, Mir, Moscou, 1985.

37. J. S. Warsa, M. Benzi, T. A. Wareing, and J. E. Morel, Preconditioning a mixed discontinuous finite element method for radiation diffusion. *Linear Algebra Appl.*, **11** (2004), pp. 795–811.

38. J. S. Warsa, M. Benzi, T. A. Wareing and J. E. Morel, Fully consistent diffusion synthetic acceleration of linear discontinuous transport discretizations on three-dimensional unstructured meshes. *Nucl. Sci. Engr.*, **141** (2002), pp. 236–251.

39. D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York and London, 1971.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 137–163

ANALYSIS OF A CLASS OF PENALTY METHODS FOR COMPUTING SINGULAR MINIMIZERS

C. CARSTENSEN[1] AND C. ORTNER[2]

Abstract — Amongst the more exciting phenomena in the field of nonlinear partial differential equations is the Lavrentiev phenomenon which occurs in the calculus of variations. We prove that a conforming finite element method fails if and only if the Lavrentiev phenomenon is present. Consequently, nonstandard finite element methods have to be designed for the detection of the Lavrentiev phenomenon in the computational calculus of variations.

We formulate and analyze a general strategy for solving variational problems in the presence of the Lavrentiev phenomenon based on a splitting and penalization strategy. We establish convergence results under mild conditions on the stored energy function. Moreover, we present practical strategies for the solution of the discretized problems and for the choice of the penalty parameter.

2000 Mathematics Subject Classification: 65N12; 65N25; 65N30; 65N50.

Keywords: Lavrentiev phenomenon, finite element method, computational calculus of variations.

1. Introduction

The calculus of variations is concerned with the minimisation problem

$$\inf E(\mathcal{A}_1) := \inf_{v \in \mathcal{A}_1} E(v), \tag{1.1}$$

where $E : \mathcal{A}_1 \to \mathbb{R} \cup \{+\infty\}$ and where $\mathcal{A}_1$ (or more generally $\mathcal{A}_p$) is the first-order Sobolev space

$$\mathcal{A}_p := W_0^{1,p}(\Omega; \mathbb{R}^m) = \{v \in W^{1,p}(\Omega)^m : v|_{\partial\Omega} = 0\},$$

based on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ with piecewise hyperplanar boundary $\partial\Omega$.

We shall assume throughout that E is *proper* on $\mathcal{A}_\infty$, i.e., there exists $v \in \mathcal{A}_\infty$ so that $E(v) < +\infty$. In particular, $\mathcal{A}_\infty \subset \mathcal{A}_1$ always implies

$$-\infty \leqslant \inf E(\mathcal{A}_1) \leqslant \inf E(\mathcal{A}_\infty) < +\infty.$$

The *Lavrentiev phenomenon*, named after its first occurence in the literature [18], is the surprising property that, in some some variational problems,

$$\inf E(\mathcal{A}_1) < \inf E(\mathcal{A}_\infty). \tag{L}$$

[1] *Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany; Department of Computational Science and Engineering, Yonsei University, 120-749 Seoul, Korea.* E-mail: cc@mathematik.hu-berlin.de

[2] *Mathematical Institute, University of Oxford*, 24-29 St Giles', Oxford OX1 3LB, UK. E-mail: ortner@maths.ox.ac.uk

Other well-known examples are the one-dimensional examples of Mania [23] and of Ball and Mizel [7, 6], or the convex example of Foss, Hrusa and Mizel [16]. In nonlinear elasticity, the Lavrentiev phenomenon is closely related to the occurence of cavitation [4].

For the conforming finite element discretization of (1.1) assume we are given a family of finite element spaces

$$V_0, V_1, V_2, \dots \subset \overline{\cup_{\ell=0}^{\infty} V_\ell} \subseteq \mathcal{A}_\infty,$$

to solve the discrete minimization problem

$$\inf E(V_\ell) := \inf_{v_\ell \in V_\ell} E(v_\ell). \tag{1.2}$$

The respective infimal energies are possibly convergent towards some limit

$$\inf E(\mathcal{A}_\infty) \leqslant \liminf_{\ell \to \infty} \inf E(V_\ell).$$

We say that the finite element method (FEM) is convergent if E and the sequence of discrete subspaces $V_0, V_1, V_2, \dots$ allow for

$$\inf E(\mathcal{A}_1) = \lim_{\ell \to \infty} \inf E(V_\ell). \tag{C}$$

Therein, the convergence of the entire sequence of energy minima (not merely of *some* subsequence but for all subsequences) is part of the statement as well as the equality of that limit to $\inf E(\mathcal{A}_1)$.

However, since conforming finite element functions are always Lipschitz continuous any finite element space V_ℓ is contained in $\mathcal{A}_\infty$ and hence standard finite element methods cannot compute singular minimisers, that is, if (L) holds then

$$\inf E(\mathcal{A}_1) < \inf E(\mathcal{A}_\infty) \leqslant \inf E(V_\ell).$$

In particular, it follows that (C) implies that (L) is false. Section 2 below provides a general framework that allows for the converse and establishes

$$\text{(C)} \quad \Longleftrightarrow \quad \text{NOT (L)},$$

under natural assumptions on the energy density.

A consequence of this equivalence is that conforming finite element methods are inappropriate tools for detecting the singular minimisers associated to the Lavrentiev phenomenon (L).

Several classes of numerical schemes have been introduced in the literature to allow for a numerical detection of (L), including the penalty method of Ball and Knowles [5, 17] and its extension to polyconvex integrands by Negron–Marrero [24], the element-removal method of Li [19, 20], and the truncation method of Li, and Bai and Li [1, 2, 21].

Section 3 introduces a general concept for the construction of a new class of splitting and penalty methods. We establish general convergence results in Sections 4 and 5. In Section 6 we discuss some connections of our results with the theory of Γ-convergence.

Similarly as in the methods of Ball & Knowles [5] and of Negron–Marrero [24] we decouple a problematic variable, for example the gradient ∇u, by introducing a new variable η in its place and then penalizing the difference $\nabla u - \eta$. The main difference between the methods

[5, 24] and our approach is how this penalization is achieved. While [5, 24] use a constraint of the form

$$\|\nabla u - \eta\|_{L^p} \leqslant \varepsilon,$$

we add a penalization term

$$\varepsilon^{-1}\Psi(\nabla u, \eta),$$

to the total energy functional. Moreover, we design this penalization term with practical implementation issues in mind. For example, by choosing a non-differentiable penalty functional (similar to an L^1-norm), we obtain the desirable property that the difference $\nabla u - \eta$ is non-zero only in a small subregion of the computational domain.

As a result of our careful design of the penalty functional our method is potentially easier to use and more efficient in practise. In particular, we also include a detailed description of a practical implementation and various computational examples in the final section of the paper.

In [25, 26] non-conforming finite element methods were analyzed as an alternative to the penalty methods discussed in the present paper. The main advantage of non-conforming methods is that they require no penalty parameter. However, even though this is a promising new direction, it is at present entirely unclear how to generalize the results in [25, 26] to the vectorial non-convex case. By contrast, our convergence results in the present paper hold under far less restrictive conditions on the stored energy functions.

2. Finite Element Failure is Equivalent to the Lavrentiev Phenomenon

This section is devoted to the proof of the equivalence of (C) and NOT (L), in a general setting which is entirely free of growth conditions and notions of convexity. However, we assume uniform convergence of the mesh-size to zero in the finite element methods as well as global continuity of the energy density.

Suppose that $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \ldots$ is a sequence of regular triangulations into simplices of a Lipschitz domain $\Omega \subset \mathbb{R}^n$ with piecewise flat boundary $\partial\Omega$ that is perfectly matched by the triangulations. Suppose that the triangulation is shape regular in the sense that the largest n dimensional ball inside each simplex T and the smallest ball outside have uniformly bounded ratios: There exists a universal positive constant $C_{\text{shaperegular}}$, which does not depend on T or ℓ, such that one finds midpoints m_T and M_T, and radii r_T and R_T, satisfying

$$B(m_T, r_T) \subset T \subset B(M_T, R_T) \quad \text{and} \quad R_T/r_T \leqslant C_{\text{shaperegular}}.$$

We assume throughout that the mesh-size tends to zero, written $h_\ell \to 0$, by which we mean that

$$\lim_{\ell\to\infty} \max_{T\in\mathcal{T}_\ell} R_T = 0.$$

The finite-dimensional space V_ℓ of piecewise affine finite element functions (piecewise with respect to the triangulation $\mathcal{T}_\ell$),

$$V_\ell := \{v_\ell \in C_0(\Omega; \mathbb{R}^m) : \forall T \in \mathcal{T}_\ell,\ v_\ell|_T \text{ affine }\},$$

belongs to $\mathcal{A}_\infty$. For future reference we also define

$$\begin{aligned} \mathrm{P}^0(\mathcal{T}_\ell) &:= \{v_\ell \in L^1(\Omega) : \forall T \in \mathcal{T}_\ell,\ v_\ell|_T \text{ constant }\},\\ \mathrm{P}^1(\mathcal{T}_\ell) &:= \{v_\ell \in C(\Omega;\mathbb{R}^m) : \forall T \in \mathcal{T}_\ell,\ v_\ell|_T \text{ affine }\}, \quad \text{and}\\ \mathrm{P}^1_0(\mathcal{T}_\ell) &:= \{v_\ell \in C_0(\Omega;\mathbb{R}^m) : \forall T \in \mathcal{T}_\ell,\ v_\ell|_T \text{ affine }\}. \end{aligned}$$

Note that with this notation, $V_\ell = \mathrm{P}^1(\mathcal{T}_\ell) \cap \mathcal{A}_\infty = \mathrm{P}^1_0(\mathcal{T}_\ell)$. In the following sections we will redefine V_ℓ in order take into account nonhomogeneous boundary conditions.

Let the energy density $W : \overline{\Omega} \times \mathbb{R}^m \times \mathbb{R}^{m\times n} \to \mathbb{R}$ be continuous and define the energy

$$E(v) := \int_\Omega W(x, v(x), Dv(x))dx,$$

for all $v \in \mathcal{A}_\infty$. In fact, if v is Lipschitz continuous, then the set of triples $\{(x, v(x), Dv(x)) : x \in \bar{\Omega}\}$ as well as the set $\{W(x, v(x), Dv(x)) : x \in \bar{\Omega}\}$ are contained in compact sets. Consequently, $E(v) \in \mathbb{R}$ and $E : \mathcal{A}_\infty \to \mathbb{R}$ is well defined. For an arbitrary function $v \in \mathcal{A}_1$ this is no longer clear. Throughout this section we simply assume that

$$E : \mathcal{A}_1 \to \mathbb{R} \cup \{+\infty\},$$

is some extension of $E|_{\mathcal{A}_\infty}$. In applications, this may be guaranteed by growth control from below and we refer to the literature (e.g., [12]) for this well-understood argument in the direct method of the calculus of variations. The question of attainment of a global or discrete minimum is irrelevant here and bypassed by a consequent discussion of infima instead of minima, e.g., for any $\ell = 0, 1, 2, \ldots,$

$$E_\ell := \inf E(V_\ell) := \inf_{v_\ell \in V_\ell} E(v_\ell) \in \mathbb{R} \cup \{\pm\infty\}.$$

We emphasize that there is no nestedness assumption on the finite element spaces and so the convergence of the infimal energies E_ℓ does *not* follow automatically. In fact, it is stated in the following theorem as a conclusion. We remark that an extension of the following result to non-homogeneous Dirichlet conditions is not straightforward since, by approximating the boundary condition, the discrete admissible set would not be contained in $\mathcal{A}_\infty$ any more.

Theorem 2.1 Finite Element Failure $\Leftrightarrow$ Lavrentiev Phenomenon. *If* $W : \bar{\Omega} \times \mathbb{R}^m \times \mathbb{R}^{m\times n} \to \mathbb{R}$ *is continuous then* $\lim_{\ell\to\infty} E_\ell = \inf E(\mathcal{A}_\infty)$ *and, in particular,*

$$\lim_{\ell\to\infty} E_\ell = \inf E(\mathcal{A}_1) \quad \Longleftrightarrow \quad \inf E(\mathcal{A}_1) = \inf E(\mathcal{A}_\infty).$$

The direction $\Longrightarrow$ in the theorem's assertion is obvious from the introduction and $V_\ell \subset \mathcal{A}_\infty$:

$$\inf E(\mathcal{A}_\infty) \leqslant \liminf_{\ell\to\infty} E_\ell = \inf E(\mathcal{A}_1) \leqslant \inf E(\mathcal{A}_\infty).$$

The converse $\Longleftarrow$ requires a density argument stated in terms of the nodal interpolation operator. Given a continuous function $v : \overline{\Omega} \to \mathbb{R}^m$ and a triangulation $\mathcal{T}_\ell$ the nodal interpolation $v_\ell := I_\ell v$ of v is defined on each simplex $T \in \mathcal{T}_\ell$ with vertices $z_1, \ldots, z_{n+1}$ through linear interpolation of the values $v(z_j)$ at the $n+1$ vertices z_j.

Lemma 2.1. *There exists a constant C, which depends only on $C_{\text{shaperegular}}$, such that, for any $v \in W^{1,\infty}(\Omega; \mathbb{R}^m)$, the piecewise affine function $v_\ell = I_\ell v$ satisfies*

$$\|v_\ell\|_{W^{1,\infty}(\Omega)} \leqslant C\|v\|_{W^{1,\infty}(\Omega)} \quad \textit{for all } \ell = 0, 1, 2, \ldots.$$

Moreover, $v_\ell \to v$ in $L^\infty(\Omega; \mathbb{R}^m)$, and $Dv_\ell \to Dv$ pointwise a.e. in Ω, as $\ell \to \infty$.

Proof. The stability of the nodal interpolation operator as well as the convergence in the L^∞-norm are standard results and can, for example, be found in [10].

The theorem of Rademacher implies that, for almost all x in some simplex T, $Dv(x)$ exists in the sense of a Fréchet derivative, i.e.,

$$Dv(x)(y - x) = v(y) - v(x) + o(|x - y|),$$

for some function $y \mapsto o(|x - y|)$ with

$$\lim_{y \to x} o(|x - y|)/|x - y| = 0.$$

Fix some $x \in \Omega$ so that, for any $\ell \in \mathbb{N}_0$, x lies in the interior of an element $T \in \mathcal{T}_\ell$ then

$$\begin{aligned} Dv(x)(z_j - z_k) &= v_\ell(z_j) - v_\ell(z_k) + o(|x - z_j|) + o(|x - z_k|) \\ &= Dv_\ell(x)(z_j - z_k) + o(|x - z_j|) + o(|x - z_k|) \quad \text{for all } j, k = 1, \ldots, n+1, \end{aligned}$$

where $|\cdot|$ denotes the ℓ^2-norm of a vector, or as below, the Frobenius norm of a matrix. Since the tangential vectors are linearly independent and the interior angles do not deteriorate we have

$$\sup_{j,k=1,\ldots,n+1} (Dv(x) - Dv_\ell(x))(z_j - z_k) \geqslant c|Dv(x) - Dv_\ell(x)| r_T,$$

where c depends only on $C_{\text{shaperegular}}$. It now follows easily that

$$\lim_{\ell \to \infty} |Dv(x) - Dv_\ell(x)| = 0.$$

□

Proof of Theorem 2.1. Given $v \in \mathcal{A}_\infty$ and its nodal interpolant $v_\ell := I_\ell v$ for all $\ell \in \mathbb{N}_0$, the previous lemma shows that

$$\lim_{\ell \to \infty} (v_\ell(x), Dv_\ell(x)) = (v(x), Dv(x)) \in \mathbb{R}^m \times \mathbb{R}^{m \times n} \quad \text{for a.e. } x \in \Omega.$$

Since W is continuous this yields pointwise convergence of the energy density

$$\lim_{\ell \to \infty} W(x, v_\ell(x), Dv_\ell(x)) = W(x, v(x), Dv(x)) \quad \text{for a.e. } x \in \Omega.$$

Furthermore, the boundedness of v_ℓ in $W^{1,\infty}(\Omega)$ and the assumption that W is continuous implies that $W(x, v_\ell(x), Dv_\ell(x))$ is bounded uniformly in x and ℓ. Consequently, Lebesgue's dominated convergence theorem shows

$$\lim_{\ell \to \infty} \int_\Omega W(x, v_\ell(x), Dv_\ell(x)) dx = \int_\Omega W(x, v(x), Dv(x)) dx = E(v).$$

Therefore,

$$\inf E(\mathcal{A}_\infty) \leqslant \liminf_{\ell\to\infty} E_\ell \leqslant \limsup_{\ell\to\infty} E_\ell \leqslant \lim_{\ell\to\infty} E(v_\ell) = E(v).$$

Since v was an arbitrary element in $\mathcal{A}_\infty$, we deduce

$$\liminf_{\ell\to\infty} E_\ell = \limsup_{\ell\to\infty} E_\ell = \inf E(\mathcal{A}_\infty).$$

In particular, we can conclude that $\lim_\ell E_\ell = \inf E(\mathcal{A}_\infty)$ exists. From this, the assertion of Theorem 2.1 follows immediately. □

3. Penalisation and Discrete Scheme

In many examples there exists a *coupling function*

$$\gamma : \Omega \times \mathbb{R}^m \times \mathbb{R}^{m\times n} \to \mathbb{M},$$

where $\mathbb{M} \equiv \mathbb{R}^\mu$ is a space of matrices, and an *extended energy density*

$$\phi : \Omega \times \mathbb{R}^m \times \mathbb{R}^{m\times n} \times \mathbb{M} \to \mathbb{R},$$

such that the energy density W is given by

$$W(x, v, F) := \phi(x, v, F, \gamma(x, v, F)),$$

for all $x \in \Omega$, $v \in \mathbb{R}^m$, $F \in \mathbb{R}^{m\times n}$. In this case, we also define

$$\Phi(v, \eta) := \int_\Omega \phi(x, v(x), Dv(x), \eta(x))dx \quad \text{for } (v, \eta) \in \mathcal{A}_1 \times L^1(\Omega; \mathbb{M}),$$

and, with the abbreviation $\gamma(\cdot, v, Dv)(x) := \gamma(x, v(x), Dv(x))$ for $x \in \Omega$, we observe that

$$E(v) = \Phi(v, \gamma(\cdot, v, Dv)). \tag{3.1}$$

Example 3.1 Polyconvex Materials. By definition, at almost all material points $x \in \Omega$ and all $v \in \mathbb{R}^m$, a polyconvex energy density $W(x, v, \cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ can be written in the form

$$W(x, v, F) = \phi(x, v, \gamma(F)),$$

where ϕ is convex in its third component (with x, v fixed), and $\gamma : \mathbb{R}^{m\times n} \to \mathbb{M}$ maps a deformation gradient F to the vector of minors (sub-determinants) of F and $\mathbb{M}$ is the space of all those minors (e.g. $\mathbb{M} = \mathbb{R}^{19}$ for $m = n = 3$ and $\mathbb{M} = \mathbb{R}^5$ for $m = n = 2$).

Example 3.2 Decoupling the Gradient. For stored energy functions $W : \Omega \times \mathbb{R}^m \times \mathbb{R}^{m\times n} \to \mathbb{R}$ where no obvious coupling mechanism is present, it is sometimes useful to let $\mathbb{M} = \mathbb{R}^{m\times n}$ and consider

$$\phi(x, v, F, \eta) := W(x, v, \eta) \quad \text{and} \quad \gamma(x, v, F) := F.$$

This decoupling of the gradient variable will help us to overcome the Lavrentiev gap phenomenon.

On the continuous level this looks as a trivial complication of the formulation but the point is that the discretisation relaxes the condition

$$\eta = \gamma(x, v, F) \quad \text{in} \quad W(x, v, F) = \phi(x, v, F, \eta).$$

Since the immediate substitution cannot detect singular minimisers with a Lavrentiev phenomenon the 'coupling' $\eta = \gamma(x, v(x), Dv(x))$ will be weakened by introducing a penalty functional,

$$\Psi_\ell : L^1(\Omega; \mathbb{M}) \times L^1(\Omega; \mathbb{M}) \to \mathbb{R} \cup \{+\infty\},$$

which is written, via some density $\psi_\ell : \Omega \times \mathbb{M} \times \mathbb{M} \to [0, \infty]$, as

$$\Psi_\ell(\eta, \zeta) := \int_\Omega \psi_\ell(x, \eta(x), \zeta(x))dx \qquad \text{for } \eta, \zeta \in L^1(\Omega; \mathbb{M}).$$

The proposed discrete minimisation problem reads: Minimise the discrete energy

$$E_\ell(v, \eta) := \Phi(v, \eta) + \Psi_\ell(\eta, \gamma(\cdot, v, Dv)),$$

over $(v, \eta) \in V_\ell \times Y_\ell$ where V_ℓ and Y_ℓ are suitable finite element spaces.

Example 3.3 Penalisation. A typical class of distance functionals is given for $1 \leqslant p < \infty$ and positive parameters ε_ℓ which possibly depend on the position x in the spatial domain (e.g., piecewise constant with respect to the triangulation $\mathcal{T}_\ell$) and

$$\psi_\ell(x, \eta, \zeta) := \varepsilon_\ell^{-1} |\eta - \zeta|^p,$$

for all $x \in \Omega$ and $\eta, \zeta \in \mathbb{M}$.

4. Polyconvex Energy Densities

An important class of energy functionals, especially in the field of nonlinear elasticity, are those where the stored energy density is polyconvex. As a prototypical model problem, we consider the stored energy density

$$W(x, u, F) = \phi(x, F, \det F) - f(x) \cdot u, \tag{4.1}$$

where $f \in L^q(\Omega)^n$ for some $q > 1$, and $\phi : \Omega \times \mathbb{R}^{n \times n} \times \mathbb{R} \to [0, +\infty]$, $n \geqslant 2$. We assume throughout this section that ϕ satisfies

$$\begin{aligned} &|F|^n + \Gamma(\eta) \lesssim \phi(x, F, \eta) \lesssim 1 + |F|^n + \Gamma(\eta), \quad \text{and} \\ &\phi(x, \cdot, \cdot) \;\; \text{is convex and l.s.c. in } \mathbb{R}^{n \times n} \times \mathbb{R} \quad \text{for a.a. } x \in \Omega, \end{aligned} \tag{4.2}$$

where $\Gamma : \mathbb{R} \to [0, +\infty]$ is convex and has superlinear growth, i.e., $\liminf_{|s| \to \infty} \Gamma(s)/s = +\infty$ [3, 11]. We remark that the growth condition $|F|^n + \Gamma(\eta)$ may be replaced by $|F|^p$ for some $p > n$. In fact, the latter implies the former.

The space of admissible functions is defined as

$$V = u_D + W_0^{1,n}(\Omega)^n,$$

where $u_D \in W^{1,n}(\Omega)^n$ and $E(u_D) < +\infty$. Under these conditions the minimization problem

$$u \in \operatorname{argmin} E(V), \tag{4.3}$$

has at least one solution [12, Theorem 2.10].

To discretize the problem we fix a sequence $u_{D,\ell} \in \mathrm{P}^1(\mathcal{T}_\ell)^n$ such that $u_{D,\ell} \to u_D$ strongly in $W^{1,n}(\Omega)^n$, and we discretize V and $L^1(\Omega)$, respectively, by

$$V_\ell = u_{D,\ell} + \mathrm{P}_0^1(\mathcal{T}_\ell)^n, \qquad \text{and} \quad Y_\ell = \mathrm{P}^0(\mathcal{T}_\ell).$$

We remark that, throughout, V denotes the admissible set, V_ℓ the discrete admissible set, and Y_ℓ the discrete admissible set for the penalty variable.

Further, we assume that we have a *penalty functional* $\Psi : L^1(\Omega)^2 \to [0, +\infty]$ such that, for all sequences (η_ℓ) and $(\zeta_\ell) \subset L^1(\Omega)$,

$$\Psi(\eta_\ell, \zeta_\ell) \to 0 \quad \Leftrightarrow \quad \|\eta_\ell - \zeta_\ell\|_{L^1} \to 0. \tag{4.4}$$

Given a sequence $\varepsilon_\ell \searrow 0$, we discretize (4.3) by

$$(u_\ell, \xi_\ell) \in \operatorname{argmin} E_\ell(V_\ell, Y_\ell),$$

where

$$\begin{aligned} E_\ell(v_\ell, \eta_\ell) = {} & \Phi(v_\ell, \eta_\ell) + \varepsilon_\ell^{-1}\Psi(\det Dv_\ell, \eta_\ell) \\ = {} & \int_\Omega \big(\phi(x, Dv_\ell, \eta_\ell) - f \cdot v_\ell\big)dx + \varepsilon_\ell^{-1}\Psi(\det Dv_\ell, \eta_\ell). \end{aligned}$$

Theorem 4.1. *Assume that* (4.1), (4.2), *and* (4.4) *hold. Then there exists a sequence* $\varepsilon_\ell \searrow 0$ *such that, for any sequence* $(u_\ell, \xi_\ell) \in V_\ell \times X_\ell$ *of approximate minimizers, that is,*

$$|E_\ell(u_\ell, \xi_\ell) - \inf E_\ell(V_\ell, Y_\ell)| \to 0 \qquad \text{as } \ell \to \infty,$$

we have

$$\Phi(u_\ell, \xi_\ell) \to \inf E(V) \quad \text{and} \quad \varepsilon_\ell^{-1}\Psi(\det Du_\ell, \xi_\ell) \to 0.$$

Moreover, the family $\{u_\ell; \ell \in \mathbb{N}\}$ *is precompact in the weak topology of* $W^{1,n}(\Omega)^n$ *and each accumulation point* u *is a minimizer of* E *in* V. *In particular, there exists a subsequence* $\ell_k \nearrow \infty$ *such that*

$$\begin{aligned} u_{\ell_k} &\rightharpoonup u && \text{weakly in } W^{1,n}(\Omega)^n, \\ \xi_{\ell_k} &\rightharpoonup \det Du && \text{weakly in } L^1(\Omega), \end{aligned}$$

where u *solves* (4.3).

The proof of Theorem 4.1 is contained in the following three lemmas.

Lemma 4.1. *Assume that* (4.1), (4.2) *and* (4.4) *hold. For every* $v \in V$ *there exists a sequence* $(v_\ell, \eta_\ell) \in V_\ell \times Y_\ell$ *such that*

$$v_\ell \to v \qquad \text{strongly in } W^{1,n}(\Omega)^n, \tag{4.5}$$

$$\lim_{\ell \to \infty} \Psi(\det Dv_\ell, \eta_\ell) = 0, \quad \text{and} \tag{4.6}$$

$$\lim_{\ell \to \infty} \Phi(v_\ell, \eta_\ell) = \Phi(v, \det Dv) = E(v). \tag{4.7}$$

Proof. Let $v \in V$. If $E(v) = +\infty$, then we take an arbitrary sequence $v_\ell \in V_\ell$ converging strongly in $W^{1,n}(\Omega)^n$ to v, and $\eta_\ell = \det Dv_\ell$. From the lower semicontinuity of E we obtain that $E(v_\ell) = \Phi(v_\ell, \eta_\ell) \to \infty$ as $\ell \to +\infty$, since, otherwise, $E(v)$ would be finite. Moreover, we have $\Psi(\det Dv_\ell, \eta_\ell) = 0$.

We may now assume that $E(v) < \infty$. We take an arbitrary sequence $v_\ell \in V_\ell$ such that $v_\ell \to v$ strongly in $W^{1,n}(\Omega)^n$ which also implies $\det Dv_\ell \to \det Dv$ strongly in $L^1(\Omega)$. The variable $\eta_\ell \in Y_\ell$ is defined as

$$\eta_\ell(x) = |T|^{-1} \int\limits_T \det Dv\, dx \qquad x \in T \in \mathcal{T}_\ell.$$

It follows that $\eta_\ell \to \det Dv$ strongly in $L^1(\Omega)$ and in particular that $\Psi(\det Dv_\ell, \eta_\ell) \to 0$. Thus, we have shown (4.5) and (4.6).

To prove (4.7) we first use Jensen's inequality to estimate, for $x \in T \in \mathcal{T}_\ell$,

$$\Gamma(\eta_\ell(x)) = \Gamma\Big(|T|^{-1} \int\limits_T \det Dv\, dx\Big) \leqslant |T|^{-1} \int\limits_T \Gamma(\det Dv) dx =: \Gamma_\ell(x),$$

i.e., Γ_ℓ is a majorant for $\Gamma(\eta_\ell)$. From its definition, and since $\Gamma(\det Dv) \in L^1(\Omega)$ (which follows from the fact that $E(v)$ is finite), it follows immediately that $\Gamma_\ell \to \Gamma(\det Dv)$ strongly in $L^1(\Omega)$.

Hence, we obtain that

$$\phi(x, Dv_\ell, \eta_\ell) \lesssim 1 + |Dv_\ell|^n + \Gamma_\ell =: a_\ell,$$

where a_ℓ is strongly convergent in $L^1(\Omega)$. For any subsequence we can extract a further subsequence such that $(Dv_\ell, \eta_\ell) \to (Dv, \eta)$ pointwise, and hence we can use a variant of Lebesgue's dominated convergence theorem [15, Sec. 1.3, Th. 4] to deduce (4.7). □

Lemma 4.2. *Assume that* (4.1), (4.2), *and* (4.4) *hold. There exists a sequence* $\varepsilon_\ell \searrow 0$ *such that*

$$\limsup_{\ell\to\infty} \min E_\ell(V_\ell, Y_\ell) \leqslant \min E(V). \tag{4.8}$$

Proof. Let $u \in \operatorname{argmin} E(V)$ and let (u_ℓ, ξ_ℓ) be the sequence constructed in Lemma 4.1 (for $v = u$). Then

$$\Psi(\det Du_\ell, \xi_\ell) \to 0,$$

and chosing $\varepsilon_\ell = \Psi(\det Du_\ell, \xi_\ell)^{1/2}$ we obtain

$$\limsup_{\ell\to\infty} \inf E_\ell(V_\ell, Y_\ell) \leqslant \limsup_{\ell\to\infty} E_\ell(u_\ell, \xi_\ell) = E(u).$$

□

In the previous lemma, we showed that it is possible to choose a sequence ε_ℓ such that the upper bound (4.8) holds. It remains to show that the limit is in fact equal.

Lemma 4.3. *Assume that* (4.1), (4.2), *and* (4.4) *hold. Suppose that a sequence* $\varepsilon_\ell \searrow 0$ *is fixed. Suppose furthermore that* $u_\ell \in V_\ell, \xi_\ell \in Y_\ell$ *such that*

$$\limsup_{\ell\to\infty} E_\ell(u_\ell, \xi_\ell) \leqslant \inf E(V), \tag{4.9}$$

then there exists a subsequence $\ell_k \uparrow \infty$ and $u \in \operatorname{argmin} E(V)$ such that

$$\begin{aligned} u_{\ell_k} &\rightharpoonup u && \textit{weakly in } W^{1,n}(\Omega)^n \\ \xi_{\ell_k} &\rightharpoonup \det Du && \textit{weakly in } L^1(\Omega), \end{aligned}$$

and moreover, we have separate convergence of the entire sequences of energy contributions:

$$\Phi(u_\ell, \xi_\ell) \to E(u), \qquad \textit{and} \qquad \varepsilon_\ell^{-1}\Psi(\det Du_\ell, \xi_\ell) \to 0.$$

Proof. It follows from (4.9) that $E_\ell(u_\ell, \xi_\ell)$ is bounded by some constant M. Using (4.2) and the assumption that u_D has finite energy, we obtain

$$M \geqslant E_\ell(u_\ell, \xi_\ell) \gtrsim \|\nabla u_\ell\|_{L^n}^n - C\|u_\ell\|_{L^{q'}} + \int\limits_\Omega \Gamma(\xi_\ell)dx + \varepsilon_\ell^{-1}\Psi(\det Du_\ell, \xi_\ell),$$

and since $W^{1,n}(\Omega)^n$ is continuously embedded in $L^{q'}(\Omega)^n$, there exists $M' \in \mathbb{R}$ such that

$$\|u_\ell\|_{W^{1,n}}^n + \int\limits_\Omega \Gamma(\xi_\ell)dx + \varepsilon_\ell^{-1}\Psi(\det Du_\ell, \xi_\ell) \leqslant M'.$$

We can therefore deduce the existence of a subsequence $\ell_k \nearrow \infty$, and of functions $u \in W^{1,n}(\Omega)^n$ and $\xi \in L^1(\Omega)$ such that

$$u_\ell \rightharpoonup u \ \text{ weakly in } W^{1,n}(\Omega)^n \quad \text{and } \xi_\ell \rightharpoonup \xi \ \text{ weakly in } L^1(\Omega).$$

(We note that the superlinear bound implies equi-integrability of the sequence (ξ_ℓ) which implies its precompactness in the weak tolopogy of $L^1(\Omega)$ [14, Cor. IV.8.11].)

Since $\det Du_\ell \rightharpoonup' \det Du$ in the sense of distributions [12, Sec. 4.2, Th. 2.6, (5)], and using (4.4), it follows that $\xi = \det Du$. Using sequential weak lower semi-continuity of energies with convex integrands [12, Sec. 3.3, Th. 3.4] we can estimate

$$\begin{aligned} E(u) \leqslant\ & \liminf_{k\to\infty} \int\limits_\Omega \Big(\phi(x, u_{\ell_k}, \xi_{\ell_k}) - f \cdot u_{\ell_k}\Big)dx \\ \leqslant\ & \liminf_{k\to\infty} \int\limits_\Omega \Big(\phi(x, u_{\ell_k}, \xi_{\ell_k}) - f \cdot u_{\ell_k}\Big)dx \\ & \qquad + \limsup_{k\to\infty} \varepsilon_{\ell_k}^{-1}\Psi(\det Du_{\ell_k}, \xi_{\ell_k}) \\ \leqslant\ & \limsup_{\ell\to\infty} E_\ell(u_\ell, \xi_\ell) \ \leqslant\ \inf E(V). \end{aligned}$$

It follows therefore that $E(u) = \inf E(V)$. Moreover, this implies that all inequalities in the above chain of estimates must be equalities, and hence,

$$\limsup_{k\to\infty} \varepsilon_{\ell_k}^{-1}\Psi(\det Du_{\ell_k}, \xi_{\ell_k}) = 0.$$

Since the proof applies also if we begin with an arbitrary subsequence, it follows that the energy of the entire sequence converges in this sense. □

Proof of Theorem 4.1. Lemma 4.2 guarantees the existence of a sequence $\varepsilon_\ell \searrow 0$ such that the conditions of Lemma 4.3 are satisfied. Hence, Lemma 4.3 guarantees the existence of a weakly convergent subsequence of approximate minimizers E_ℓ and establishes the various convergence statements in the theorem. □

Remark 4.1. 1. In practise, the condition that Ψ is continuous in the strong topology of $L^1(\Omega;\mathbb{R}^n)$ requires that Ψ takes the form

$$\Psi(\eta,\zeta) = \int_\Omega \psi(|\eta-\zeta|)dx,$$

where ψ has 1-growth at infinity. Typical penalty densities ψ are $\psi(t)=|t|$, or, if one prefers a smooth functional, $\psi(t)=(t^2+1)^{1/2}-1$. The condition (4.4) can be obtained, for example, by requiring that $\psi \geqslant 0$ and $\psi(t)=0$ if and only if $t=0$.

2. If ϕ satisfies a stronger growth condition, for example $\phi(x,F,g) \gtrsim |F|^p$ for some $p>n$ then this additional integrability allows us to use a penalty functional which is only continuous in $L^{p/n}(\Omega;\mathbb{R}^n)$.

3. We have only shown the existence of some sequence ε_ℓ for which we obtain convergence of the penalty method. We will show in Section 7 below how this sequence can be constructed in practise.

4. More general polyconvex material models where ϕ depends on all minors of the gradient can be easily incorporated in our analysis. One would then have to decouple all minors which appear in the definition of the functional. Similar convergence can then be obtained whenever the growth conditions from above and below are the same and are sufficiently strong so that the direct method can be applied.

5. Examples with Lavrentiev Phenomenon

In many problems decoupling the gradient is sufficient, and it is the goal of this section to make this precise. This is possible whenever W is convex in the third component, but it is also a useful approach if it is unclear which variable should be relaxed. We begin again with a more general discussion which we then make precise at two classes of problems, general one-dimensional functionals with continuous integrands, and higher-dimensional examples with mild v-dependence of the integrand.

We assume throughout that $W=\phi:\Omega\times\mathbb{R}^m\times\mathbb{R}^{m\times n}\to(-\infty,+\infty]$ is lower semi-continuous in all three variables, continuous at every point (x,v,η) where $\phi(x,v,\eta)<\infty$, and that it satisfies the lower bound

$$\phi(x,v,\eta) \gtrsim -1-|v|^q, \tag{5.1}$$

where $1\leqslant q<n/(n-1)$ if $n\geqslant 2$ and $1\leqslant q<\infty$ if $n=1$. This implies in particular that, for $v\in W^{1,1}(\Omega)^m$ and $\eta\in L^1(\Omega)^{m\times n}$, the functionals

$$\Phi(v,\eta)=\int_\Omega W(x,v,\eta)dx, \quad \text{and } E(v)=\Phi(v,Dv),$$

are well-defined in $(-\infty,+\infty]$. Let $u_D\in W^{1,1}(\Omega)^m$ such that $E(u_D)<\infty$ and define $V=u_D+W_0^{1,1}(\Omega)^m$.

We will assume that the penalty has 1-growth, namely, that there exists a continuous penalty density $\psi : \mathbb{R}^{m\times n} \to [0,\infty)$ satisfying

$$\begin{aligned} |\eta| - 1 \lesssim \psi(\eta) \lesssim |\eta| + 1 \qquad & \text{for all } \eta \in \mathbb{R}^{m\times n}, \quad \text{and} \\ \psi(\eta) = 0 \qquad & \text{if and only if } \eta = 0, \end{aligned} \tag{5.2}$$

such that the functional Ψ is of the form

$$\Psi(\eta,\zeta) = \int_\Omega \psi(\eta - \zeta)dx \qquad \text{for all } \eta, \zeta \in L^1(\Omega)^{m\times n}. \tag{5.3}$$

To discretize the problem of minimizing E over V we take $u_{D,\ell} \in \mathrm{P}^1(\mathcal{T}_\ell)$ such that $u_{D,\ell} \to u_D$ strongly in $W^{1,1}(\Omega)^m$, and define

$$V_\ell = u_{D,\ell} + \mathrm{P}^1_0(\mathcal{T}_\ell)^m \qquad \text{and} \qquad Y_\ell = \mathrm{P}^0(\mathcal{T}_\ell)^{m\times n},$$

to discretize, respectively, the variables u and η. We approximate Φ using the midpoint rule: For $v_\ell \in \mathrm{P}^1(\mathcal{T}_\ell)^m$, we set $\bar{v}_\ell(x) = (v_\ell)_T := |T|^{-1} \int_T v_\ell \, dx$ for $x \in T \in \mathcal{T}_\ell$, and for $v_\ell \in V_\ell$ and $\eta_\ell \in Y_\ell$, we define

$$\Phi_\ell(v_\ell, \eta_\ell) = \int_\Omega \phi(\bar{x}_\ell, \bar{v}_\ell, \eta_\ell)dx = \sum_{T\in\mathcal{T}_\ell} |T|\phi\big((x)_T, (v_\ell)_T, \eta_\ell|_T\big).$$

The functional Φ_ℓ is extended in an obvious way to $V_\ell \times L^1(\Omega)^{m\times n}$.

Remark 5.1. We could have included a quadrature approximation in our analysis in Section 4 as well. For the sake of simplicity, we decided not to do so. In the present case, we are in fact unable to prove convergence of the penalty method *without* the quadrature approximation. The reason for this is essentially that we have chosen $\eta_\ell \in \mathrm{P}^0(\mathcal{T}_\ell)^{m\times n}$ and hence we can only adjust its value to a single point within each element. Since we assume no control on ϕ from above we cannot control an integral over an element from information at a single quadrature point.

Our first aim is an approximation result akin to Lemma 4.1. In Lemma 5.1 below we reduce this task to the following general condition which can be quite easily checked for different problems: *for all $v \in V$ there exists a function $\zeta \in L^1(\Omega)^{m\times n}$ and a sequence $v_\ell \in V_\ell$ such that the following conditions are satisfied:*

$$\begin{aligned} &\textit{(i)} \qquad \phi(x, v, \zeta) \in L^1(\Omega), \\ &\textit{(ii)} \qquad v_\ell \to v \textit{ strongly in } W^{1,1}(\Omega)^m, \textit{ and} \\ &\textit{(iii)} \qquad \limsup_{\ell\to\infty} \Phi_\ell(v_\ell, \zeta) \leqslant \Phi(v, \zeta). \end{aligned} \tag{5.4}$$

Example 5.1 1D Examples. Suppose that $n = 1$, that $\phi : \Omega \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is globally continuous, and assume that $u_{D,\ell} = u_D$ for all ℓ. This class includes in particular problems of Maniá type [7, 23].

We now prove that (5.4) holds under this assumption. Let $v \in V$ and let v_ℓ be its piecewise affine nodal interpolant. Then $v_\ell \to v$ strongly in $W^{1,1}(\Omega)^m$, $(\bar{x}_\ell, \bar{v}_\ell(x)) \to (x, v(x))$ uniformly in Ω and, since ϕ is globally continuous,

$$\Phi_\ell(v_\ell, \zeta) \to \Phi(v, \zeta),$$

for any fixed $\zeta \in \mathbb{R}^m$. □

Example 5.2 Weak Coupling of u and Du. Suppose that, in addition to (5.1),

$$\phi(x, v, \eta) \lesssim |v|^q + \Gamma(\eta), \tag{5.5}$$

where $\Gamma : \mathbb{R}^{m\times n} \to [0, +\infty]$ is proper. We note that this class includes in particular the example of Foss, Hrusa, and Mizel [16] and Ball's example of cavitation [4].

We now prove that (5.4) holds under this assumption. Let $v \in V$ and take $v_\ell \in V_\ell$ converging strongly in $W^{1,1}(\Omega)^m \cap L^q(\Omega)^m$ to v. In particular, we also have $\bar{v}_\ell \to v$ strongly in $L^q(\Omega)^m$ by Lebesgue's differentiation theorem. Further, let $\zeta \in \mathbb{R}^{m\times n}$ such that $\Gamma(\zeta) < +\infty$. In view of the growth condition imposed in (5.5) we obtain $\phi(x, v(x), \zeta) \in L^1(\Omega)$. Let $\ell_j \nearrow \infty$ be a subsequence such that

$$\limsup_{\ell\to\infty} \Phi_\ell(v_\ell, \zeta) = \lim_{j\to\infty} \Phi_{\ell_j}(v_{\ell_j}, \zeta).$$

Upon extracting a further subsequence we may assume that $(\bar{x}_{\ell_j}, \bar{v}_{\ell_j}) \to (x, v)$ pointwise a.e. in Ω. Since $\phi(x, v(x), \zeta) \in L^1(\Omega)$ it is finite for a.a. $x \in \Omega$ and hence continuous at those points. We therefore obtain

$$\lim_{j\to\infty} \phi(\bar{x}_j, \bar{v}_j, \zeta) = \phi(x, v, \zeta) \quad \text{pointwise a.e. in } \Omega.$$

The majorant

$$\phi(\bar{x}_{\ell_j}, \bar{v}_{\ell_j}, \zeta) \leqslant |\bar{v}_{\ell_j}|^q + \Gamma(\zeta),$$

is strongly convergent in $L^1(\Omega)$ and hence we can use Fatou's Lemma to obtain (5.4) (iii). □

Having shown that (5.4) indeed holds for several interesting problem classes we establish the basic approximation result, which it implies.

Lemma 5.1. *Fix $\varepsilon > 0$ and suppose that* (5.2), (5.3) *and* (5.4) *hold; then, for every $v \in V$ there exists a sequence $(v_\ell, \eta_\ell) \in V_\ell \times Y_\ell$ such that*

$$\limsup_{\ell\to\infty} \left[\Phi(v_\ell, \eta_\ell) + \varepsilon^{-1}\Psi(Dv_\ell, \eta_\ell)\right] \leqslant E(v).$$

Moreover, the sequence v_ℓ can be chosen independent of the value of ε.

Proof. We take the sequence v_ℓ specified in (5.4). For every $T \in \mathcal{T}_\ell$ and $x \in T$ we define

$$\bar{\phi}_\ell(x) = \inf_{\xi\in\mathbb{R}^{m\times n}} \left[\phi(\bar{x}_\ell(x), \bar{v}_\ell(x), \xi) + \varepsilon^{-1}\psi(\xi - Dv_\ell(x))\right].$$

Since $\bar{v}_\ell$ and $\bar{x}_\ell$ are piecewise constant $\bar{\phi}_\ell$ may also be chosen as a piecewise constant function and it follows from the growth condition on ϕ from below that it is finite. In particular, it is measurable and its integral is well-defined with a value in $(-\infty, +\infty]$.

There exists a subsequence $\ell_j \nearrow \infty$ such that

$$\limsup_{\ell\to\infty} \int_\Omega \bar{\phi}_\ell \, dx = \lim_{j\to\infty} \int_\Omega \bar{\phi}_{\ell_j}, \quad \text{and}$$

$$(\bar{v}_{\ell_j}, Dv_{\ell_j}) \to (v, Dv) \;\; \text{pointwise a.e. in } \Omega.$$

From the definition of $\bar{\phi}_\ell$, we have

$$\bar{\phi}_\ell \leqslant \phi(\bar{x}_\ell, \bar{v}_\ell, Dv) + \varepsilon^{-1}\psi(Dv - Dv_\ell) \quad \text{for a.a. } x \in \Omega,$$

and since we assumed that ϕ is continuous at every point where it is finite, and that ψ is globally continuous, we obtain

$$\limsup_{j\to\infty} \bar{\phi}_{\ell_j}(x) \leqslant \phi(x, v(x), Dv(x)) \quad \text{for a.a. } x \in \Omega. \tag{5.6}$$

Again using the definition of $\bar{\phi}_\ell$ we obtain the majorant

$$\bar{\phi}_\ell \leqslant \phi(\bar{x}_\ell, \bar{v}_\ell, \zeta) + \varepsilon^{-1}\psi(\zeta - Dv_\ell) =: m_\ell,$$

where $\zeta \in L^1(\Omega)^{m\times n}$ is taken from (5.4). Since ϕ is continuous at $(x, v(x), \zeta(x))$, for a.a. $x \in \Omega$, it follows from (5.4) (ii) that

$$m_{\ell_j}(x) \to m(x) := \phi(x, v(x), \zeta(x)) + \varepsilon^{-1}\psi(\zeta(x) - Dv(x)) \quad \text{for a.a. } x \in \Omega.$$

Condition (5.4) (iii) translates as

$$\liminf_{j\to\infty} \int_\Omega m_{\ell_j}\,dx \leqslant \int_\Omega m\,dx.$$

Applying Fatou's lemma to the sequence $m_\ell - \bar{\phi}_\ell$ gives

$$\int_\Omega \liminf_{j\to\infty}(m_{\ell_j} - \bar{\phi}_{\ell_j})dx \leqslant \liminf_{j\to\infty} \int_\Omega (m_{\ell_j} - \bar{\phi}_{\ell_j})dx,$$

which can, equivalently, be written as

$$\int_\Omega \big(m - \limsup_{j\to\infty} \bar{\phi}_{\ell_j}\big)dx \leqslant \int_\Omega m dx - \limsup_{j\to\infty} \int_\Omega \bar{\phi}_{\ell_j} dx,$$

and hence we obtain, using (5.6) in the last inequality,

$$\limsup_{\ell\to\infty} \int_\Omega \bar{\phi}_\ell\,dx = \limsup_{j\to\infty} \int_\Omega \bar{\phi}_{\ell_j}\,dx \leqslant \int_\Omega \limsup_{j\to\infty} \bar{\phi}_{\ell_j}\,dx \leqslant E(v).$$

It remains to show that there exists a sequence $\eta_\ell \in Y_\ell$ such that

$$\limsup_{\ell\to\infty} \int_\Omega \bar{\phi}_\ell\,dx = \limsup_{\ell\to\infty} \Phi_\ell(v_\ell, \eta_\ell).$$

To this end we choose $\eta_\ell \in Y_\ell$, such that

$$\phi(\bar{x}_\ell, \bar{v}_\ell(x), \eta_\ell(x)) \leqslant \bar{\phi}_\ell(x) + 1/\ell \qquad \text{for a.e. } x \in \Omega.$$

The existence of such functions follows from the definition of $\bar{\phi}_\ell$. □

Next, we will deduce from Lemma 5.1 the existence of a sequence $\varepsilon_\ell \searrow 0$ for which the same upper bound still holds.

Lemma 5.2. *Suppose that* (5.2), (5.3) *and* (5.4) *hold; then there exists a sequence* $\varepsilon_\ell \searrow 0$ *such that*

$$\limsup_{\ell\to\infty} \inf E_\ell(V_\ell, Y_\ell) \leqslant \inf E(V).$$

Proof. Let $v_k \in V$ such that $E(v_k) \leqslant \inf E(V) + 1/k$. According to Lemma 5.1, for every $k \in \mathbb{N}$, there exists $\ell_k \in \mathbb{N}$ such that, for all $\ell \geqslant \ell_k$,

$$\inf_{(u_\ell,\xi_\ell)\in V_\ell\times Y_\ell} \big[\Phi_\ell(u_\ell,\xi_\ell) + k\Psi(\xi_\ell, Du_\ell)\big] \leqslant E(v_k) + 1/k \leqslant \inf E(V) + 2/k.$$

We may assume that $\ell_k \leqslant \ell_{k+1}$ for all k. If we define

$$\varepsilon_\ell = 1/k \quad \text{for } \ell_k \leqslant \ell < \ell_{k+1}, \quad k = 1, 2, \ldots,$$

and $\varepsilon_\ell = 1$ for $1 \leqslant \ell < \ell_1$, then $\varepsilon_\ell \searrow 0$ and

$$\inf E_\ell(V_\ell, Y_\ell) \leqslant \inf E(V) + 2\varepsilon_\ell \qquad \text{for all } \ell \geqslant \ell_1.$$

□

We only need to prove a lower bound now. Here, we distinguish two cases: whether ϕ is convex in the third component or only quasiconvex.

We adopt assumption (ii) in the following theorem as an abstract compactness assumption that we found difficult to verify for examples where we observe it in practise, such as the Foss/Hrusa/Mizel example in Section 7.5. Failure of this assumption will normally be displayed as an instability in the numerical calculation.

Theorem 5.1 Convex Energies. *Suppose that* (5.2), (5.3) *and* (5.4) *hold, and assume in addition that* ϕ *is* convex *in its third component. Let* $\varepsilon_\ell \searrow 0$ *be the sequence established in Lemma 5.2, and let* $(u_\ell, \xi_\ell) \in V_\ell \times Y_\ell$ *be a sequence satisfying the following conditions:*

(i) (u_ℓ, ξ_ℓ) *are* approximate minimizers, *i.e.,*

$$|E_\ell(u_\ell,\xi_\ell) - \inf E_\ell(V_\ell, Y_\ell)| \to 0 \quad \textit{as } \ell \to \infty. \tag{5.7}$$

(ii) There exists $u \in V$ *such that*

$$u_\ell \rightharpoonup u \qquad \textit{weakly in } W^{1,1}(\Omega)^m. \tag{5.8}$$

Then $u \in \operatorname{argmin} E(V)$,

$$\begin{aligned} \lim_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell) &= E(u), \\ \lim_{\ell\to\infty} \varepsilon_\ell^{-1}\Psi(\xi_\ell, Du_\ell) &= 0, \ \textit{and} \\ \xi_\ell \rightharpoonup Du & \qquad \textit{weakly in } L^1(\Omega)^{m\times n}. \end{aligned} \tag{5.9}$$

Proof. By the construction of ε_ℓ and assumption (5.7) we have

$$\limsup_{\ell\to\infty} E_\ell(u_\ell,\xi_\ell) \leqslant \inf E(V).$$

In particular, $\Psi(\xi_\ell, Du_\ell) \lesssim \varepsilon_\ell \to 0$ which implies $\xi_\ell \rightharpoonup Du$ weakly in $L^1(\Omega)^{m\times n}$. We can therefore deduce that

$$E(u) \leqslant \liminf_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell).$$

Using the same arguments as in the proof of Lemma 4.3 we can conclude the proof of the theorem. □

In addition to assumption (ii) in Theorem 5.1 we require another stability assumption in the quasiconvex case. Assumption (iii) in the following theorem will be satisfied whenever singularities occur only in localized regions. This is again observed in typical numerical experiments but would be very difficult to prove rigorously.

Theorem 5.2 Quasiconvex Energies. *Suppose that* (5.2), (5.3) *and* (5.4) *hold, and assume in addition that* ϕ *is* quasiconvex *in its third component. Let* $\varepsilon_\ell \searrow 0$ *be the sequence established in Lemma 5.2 and let* $(u_\ell,\xi_\ell) \in V_\ell \times Y_\ell$ *be a sequence satisfying (i) and (ii) in Theorem 5.1, as well as:*

(iii) There exists a monotone family of subsets $\Omega_k \nearrow \Omega$ *such that*

$$\lim_{\ell\to\infty} \big\|\phi(\bar x_\ell,\bar u_\ell,Du_\ell) - \phi(\bar x_\ell,\bar u_\ell,\xi_\ell)\big\|_{L^1(\Omega_k)} = 0 \quad \textit{and} \tag{5.10}$$

$$\forall k\in\mathbb{N} \quad \sup_{\ell\geqslant k} \|u_\ell\|_{W^{1,\infty}(\Omega_k)} < \infty. \tag{5.11}$$

Then $u \in \operatorname{argmin} E(V)$ *and the conclusion* (5.9) *remains true as well.*

Proof. In view of the bound (5.11), for fixed $k\in\mathbb{N}$, we have

$$u_\ell \overset{*}{\rightharpoonup} u \qquad \text{weakly-}{*} \text{ in } W^{1,\infty}(\Omega_k)^m.$$

Since ϕ is quasiconvex in its third component it follows from (5.10) that

$$\begin{aligned}\int_{\Omega_k} \phi(x,u,Du)dx &\leqslant \liminf_{\ell\to\infty} \int_{\Omega_k} \phi(\bar x_\ell,\bar u_\ell,Du_\ell)dx \\ &= \liminf_{\ell\to\infty} \int_{\Omega_k} \phi(\bar x_\ell,\bar u_\ell,\xi_\ell)dx.\end{aligned}$$

Using the lower bound (5.1), the compactness of the embedding $W^{1,1}(\Omega)^m \subset L^q(\Omega)^m$, and setting $\Omega_k' = \Omega\setminus\Omega_k$, we obtain

$$\begin{aligned}\int_{\Omega_k} \phi(x,u,Du)dx &\leqslant \liminf_{\ell\to\infty}\Big(\Phi_\ell(u_\ell,\xi_\ell) - \int_{\Omega_k'} \phi(\bar x_\ell,\bar u_\ell,\xi_\ell)dx\Big) \\ &\leqslant \liminf_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell) + \limsup_{\ell\to\infty} C(|\Omega_k'| + \|u_\ell\|^q_{L^q(\Omega_k')}) \\ &= \liminf_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell) + C(|\Omega_k'| + \|u\|^q_{L^q(\Omega_k')}).\end{aligned}$$

Setting $\delta_k = C(|\Omega_k'| + \|u\|^q_{L^q(\Omega_k')})$ we can further estimate

$$\begin{aligned}
\int\limits_{\Omega_k} \phi(x,u,Du)dx &\leqslant \liminf_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell) + \delta_k \\
&\leqslant \liminf_{\ell\to\infty} \Phi_\ell(u_\ell,\xi_\ell) + \limsup_{\ell\to\infty} \varepsilon_\ell^{-1}\Psi(Du_\ell,\xi_\ell) + \delta_k \\
&\leqslant \limsup_{\ell\to\infty} E_\ell(u_\ell,\xi_\ell) + \delta_k \\
&\leqslant \inf E(V) + \delta_k \qquad \text{for all } k \in \mathbb{N}.
\end{aligned} \tag{5.12}$$

Adding the term $C(1+|u|^q)$ to the integral on the left-hand side the integrand becomes non-negative and the bound becomes

$$\int\limits_{\Omega_k} \big[\phi(x,u,Du) + C(1+|u|^q)\big]dx \leqslant \inf E(V) + \int\limits_{\Omega} C(1+|u|^q)dx.$$

Taking the supremum over k on the left-hand side (employing, for example, the Beppo-Levi theorem), it follows that $\phi(x,u,Du)$ is integrable and that $u \in \operatorname{argmin} E(V)$. Furthermore, we can let $k \to \infty$ and thus $\delta_k \to 0$ in (5.12) from which we can deduce the separate convergence of the energy contributions (compare also with the proof of Lemma 4.3). □

6. Connection with Γ-Convergence

Our main results, Theorems 4.1, 5.1 and 5.2, can be understood as Γ-convergence (also known as epi-convergence) results. The purpose of the present section is to briefly explain this connection. We refer to the monographs of Braides [9] and Dal Maso [13] for an introduction to Γ-convergence.

We will demonstrate this point of view at the example of the polyconvex case. To this end, suppose that (4.1), (4.2), (4.4) and (5.2) hold, and define, for $v \in W^{1,n}(\Omega)^n$, $\eta \in L^1(\Omega)$ and $\varepsilon \in [0,\infty)$,

$$F(v,\eta,\varepsilon) = \begin{cases} E(v) & \text{if } v \in V, \eta = \det Dv, \varepsilon = 0, \\ \Phi(v,\eta) + \varepsilon^{-1}\Psi(\det Dv,\eta) & \text{if } v \in V, \varepsilon \in (0,\infty), \\ +\infty & \text{otherwise;} \end{cases}$$

$$F_\ell(v,\eta,\varepsilon) = \begin{cases} \Phi(v,\eta) + \varepsilon^{-1}\Psi(\det Dv,\eta) & \text{if } v \in V_\ell, \eta \in Y_\ell, \varepsilon \in (0,\infty), \\ +\infty & \text{otherwise.} \end{cases}$$

A minor modification of Lemma 4.2 shows that, for each $u \in V$, $\xi = \det Du$, there exists a sequence $u_\ell \to u$ strongly in $W^{1,n}(\Omega)^n$, $\xi_\ell \to \xi$ strongly in $L^1(\Omega)$, and $\varepsilon_\ell \to 0$ such that

$$\limsup_{\ell\to\infty} F_\ell(u_\ell,\xi_\ell,\varepsilon_\ell) \leqslant F(u,\xi,0). \tag{6.1}$$

If $\xi \neq \det Du$ then $F(u,\xi,0) = +\infty$ and hence (6.1) is trivially satisfied.

On the other hand, in Lemma 4.3, we have proven that, whenever $u_\ell \rightharpoonup u$ weakly in $W^{1,n}(\Omega)^n$, $\xi_\ell \rightharpoonup \xi$ weakly in $L^1(\Omega)$, and $\varepsilon_\ell \to 0$, then

$$F(u,\xi,0) \leqslant \liminf_{\ell\to\infty} F_\ell(u_\ell,\xi_\ell,\varepsilon_\ell). \tag{6.2}$$

Strictly speaking we have shown this for the case $\xi = \det Du$, but we have also shown that all accumulation points of families with bounded energy satisfy this. Hence, (6.2) is indeed correct.

In the language of Γ-convergence (6.1) and (6.2) are, respectively, called the *limsup* and *liminf conditions* (here only for $\varepsilon = 0$), and together they can be written as

$$\Gamma\text{–}\lim_{\ell\to\infty} F_\ell(v,\eta,0) = F(v,\eta,0) \qquad \text{for all } v \in V, \eta \in L^1(\Omega), \tag{6.3}$$

where Γ-convergence is understood with respect to the weak $W^{1,n}(\Omega)^n \times L^1(\Omega) \times [0,\infty)$-topology. In fact, it is straightforward to verify that

$$\Gamma\text{–}\lim_{\ell\to\infty} F_\ell = F,$$

holds in the entire space $W^{1,n}(\Omega)^n \times L^1(\Omega) \times [0,\infty)$, however, this is less relevant for our purposes.

Thus, Theorem 4.1 can be interpreted as a Γ-convergence result in the sense of (6.3). In an obvious way, Theorems 5.1 and 5.2 can also be written in this way. We note however, that our original statements are slightly stronger in that we obtain separate convergence of the different contributions to the energy.

To conclude, we note that the statement

$$\Gamma\text{–}\lim_{\ell\to\infty} F_\ell(\cdot,\cdot,\varepsilon_\ell) = F(\cdot,\cdot,0),$$

for a fixed sequence $\varepsilon_\ell \to 0$, is in general *false*. To see this, observe that to obtain (6.1), the choice of the sequence (ε_ℓ) may strongly depend on the limit point u which we are aiming to approximate.

7. Algorithms and Numerical Examples

In the preceding sections we have formulated a general class of numerical methods for the solution of problems of the calculus of variations. The purpose of the present section is to demonstrate how they can be efficiently implemented and to demonstrate their practicality at several examples. We aim to give as much detail as possible so that our numerical results may be easily reproduced.

7.1. Optimization of non-differentiable energies

We begin by describing the implementation of the non-differentiable functionals which arise in our penalization procedure. Recall that we are aiming to minimize an energy which can be written in the form

$$\begin{aligned} E(v) &= \int_\Omega W(x,v,Dv)dx \\ &= \int_\Omega \phi\big(x,v,Dv,\gamma(Dv)\big)dx, \end{aligned}$$

over a convex and closed subset $V \subset W^{1,1}(\Omega)^m$, where $\phi(x,v,F,\eta)$ and $\gamma(F)$ are assumed to be smooth (at least twice differentiable) in v, F, and η. For the sake of simplicity we do not consider $\gamma = \gamma(x,v,Dv)$, but this is not a true restriction.

We shall consider general penalty functionals of the type

$$E_\varepsilon(v,\eta) = \int_\Omega \phi(x,v,Dv,\eta)dx + \varepsilon^{-1}\int_\Omega \big|\gamma(Dv)-\eta\big|_1 dx, \tag{7.1}$$

defined for $v \in V_\ell = u_{D,\ell}+\mathrm{P}^1(\mathcal{T}_\ell)^m$, $\eta \in Y_\ell = \mathrm{P}^0(\mathcal{T}_\ell)^\mu$, and where $|\cdot|_1$ denotes the ℓ^1-norm. We will see in numerical experiments that the L^1-type penalty functional guarantees a compact support of the difference $\gamma(Dv)-\eta$. This gives us information about the location of the singularities and also significantly reduces the complexity of the optimization (the optimization software TRON [22] automatically removes the unnecessary degrees of freedom).

By a simple variable transformation, we can replace η by $\eta+\gamma(F)$ to obtain a new functional

$$\int_\Omega \phi\big(x,v,Dv,\gamma(Dv)+\eta\big)dx + \int_\Omega |\eta|_1 dx.$$

Next, we split the variable η into $\eta=\eta^+-\eta^-$ where $\eta_j^+ = \max(\eta_j,0)$ and $\eta_j^- = -\min(\eta_j,0)$, $j=1,\dots,\mu$, and hide $\gamma(F)$ within a newly defined energy density

$$\widetilde{\phi}(x,u,F,\eta) = \phi\big(x,u,F,\gamma(F)+\eta\big),$$

to rewrite the functional as

$$\widetilde{E}_\varepsilon(v,\eta^+,\eta^-) = \int_\Omega \widetilde{\phi}(x,v,Dv,\eta^+-\eta^-)dx + \varepsilon^{-1}\int_\Omega |\eta^+|_1 + |\eta^-|_1 dx. \tag{7.2}$$

Upon making η^+ and η^- independent variables but imposing the bound constraints $\eta^+ \geqslant 0$ and $\eta^- \geqslant 0$ we have thus turned the original non-differentiable problem to minimize (7.1) into a smooth but constrained optimization problem. In particular, we define (7.2) for all $v \in V_\ell$ and for all $\eta^+,\eta^- \in Y_\ell^+$, where

$$Y_\ell^+ = \{\eta \in Y_\ell : \eta_j \geqslant 0 \ \text{ in } \Omega, j=1,\dots,\mu\}.$$

The uunctionals in (7.2) can be easily implemented with its gradient and hessian provided exactly. Our own implementation uses the trust region software TRON [22] to solve the *local* minimization problem

$$\min_{\substack{u\in V_\ell\\ \xi^\pm \in Y_\ell^+}} \widetilde{E}_\varepsilon(u,\xi^+,\xi^-). \tag{7.3}$$

7.2. Adaptive mesh refinement for the penalty method

At several points in the continuation algorithm for the penalty method, described in the following section, we have to refine the mesh based on one of two principles: (i) either to reduce the overall energy or (ii) to reduce the contribution from the penalty term.

(i) To reduce the overall energy we use a DWR-type idea [8]. Let $(u_\ell,\xi_\ell^+,\xi_\ell^-)$ be a local minimum of $\widetilde{E}_\varepsilon$, computed using the method described above. We then define the error indicators

$$\eta_{\mathrm{e}} = \sum_{T\in\mathcal{T}_\ell} \eta_T, \quad \text{where}$$

$$\eta_T = \left| \int_T \partial_F \widetilde{\phi}(x,u_\ell,Du_\ell,\xi_\ell^+-\xi_\ell^-) : (Du_\ell - G_\ell)dx \right|,$$

where $G_\ell \in \mathrm{P}^1(\mathcal{T}_\ell)^{m\times n}$ is a gradient recovery defined at each node z of the mesh $\mathcal{T}_\ell$ by

$$G_\ell(z) = -\!\!\!\!\!\!\int\limits_{\cup\{T\in\mathcal{T}_\ell : z\in T\}} Du_\ell \, dx.$$

The value η_{e} gives an indication how much the "elastic" energy may be lowered by local mesh refinement. On the other hand, the value of the penalty integral

$$\eta_{\mathrm{p}} = \varepsilon^{-1} \int\limits_\Omega |\xi_\varepsilon^+|_1 + |\xi_\varepsilon^-|_1 dx,$$

indicates how much the "penalty" energy can be lowered. If $\eta_{\mathrm{e}} > C_{\mathrm{e,p}}\eta_{\mathrm{p}}$ then the mesh is refined by marking a fraction of all elements which have the largest indicators η_T for refinement. Otherwise all those elements are marked where $\xi^+ + \xi^-$ is non-zero (up to a threshhold which takes round-off errors and premature termination of the optimization into account).

(ii) To reduce the penalty energy we use the very same procedure. All those elements are marked for refinement where $\xi^+ + \xi^-$ is non-zero.

7.3. Continuation algorithm

A major difficulty one encounters when solving problems involving the Lavrentiev phenomenon is the so-called repulsion property. For example, if $u_j \to x^{1/3}$ strongly in $L^1(0,1)$, but $u_j \in W^{1,\infty}(0,1)$ for all j, then

$$\int\limits_0^1 |u_{j,x}|^6 (u_j^3 - x)^2 dx \to +\infty.$$

We can imagine this effect as a huge energy barrier that needs to be overcome (or a complicated path to be found) when moving from a Lipschitz function to the global minimum. In our computations, we see this effect in that even for sufficiently small meshes it is often difficult to find the correct minimizers and that the penalty method converges to the Galerkin solution instead. (By "*Galerkin solution*", we mean any P^1-minimizer of the original non-penalized functional.) In particular, we observed that a local minimum when ε is chosen too small in relation to the current mesh since in that case the penalty method becomes in effect a Galerkin method again.

Thus the problem may be overcome by, either increasing ε, or decreasing the mesh size. The former is clearly not desirable while the latter may be prohibitively expensive. Our solution therefore was to consider a continuation with respect to the parameter ε. By initially choosing ε very large the Galerkin solution is automatically discarded even for coarse meshes. We then gradually decrease ε and adapt the mesh whenever there is a danger that we may "fall out" of the basin of attraction of the exact minimizer because ε has become too small for the current mesh. This may be controlled by requiring that at all times the total energy $\widetilde{E}_\varepsilon$ must be below a critical value which should be less than the energy of the Galerkin solution.

(1) Choose $\varepsilon_{\mathrm{dec}} \in (0,1)$, $E_{\mathrm{goal}} \in \mathbb{R}$, ε_0, an initial mesh $\mathcal{T}_0$, and two bounds $N_{\mathrm{opt}}^1, N_{\mathrm{opt}}^2$ (see remarks below how to coose them) for number of iterations of the optimization. Set $\ell = 0$ and $N_{\mathrm{opt}} = N_{\mathrm{opt}}^2$.

(2) Minimize $\widetilde{E}_{\varepsilon_\ell}$, allowing at most N_{opt} iterations.

(3) Determine next action:

(3.1) If the optimization converged and $\widetilde{E}_{\varepsilon_\ell} \leqslant E_{\text{goal}}$ accept the step, set $\ell \leftarrow \ell + 1$, $\varepsilon_\ell = \varepsilon_{\ell-1} \cdot \varepsilon_{\text{dec}}$, $\mathcal{T}_\ell = \mathcal{T}_{\ell-1}$, $N_{\text{opt}} = N^1_{\text{opt}}$ and continue at (2).

(3.2) If the optimization converged but $\widetilde{E}_{\varepsilon_\ell} > E_{\text{goal}}$ use refinement strategy (i) of the previous section to obtain a new mesh $\mathcal{T}_\ell$, set $N_{\text{opt}} = N^2_{\text{opt}}$, and redo step (2).

(3.3) If the optimization did not converge use refinement strategy (ii) of the previous section to obtain a new mesh $\mathcal{T}_\ell$, set $\varepsilon_\ell = \varepsilon_{\ell-1}$, $N_{\text{opt}} = N^2_{\text{opt}}$, and redo step (2).

Some further comments to refine the continuation algorithm are required.

- The initial parameters for step (1) have to be chosen in such a way that the first step is always succesful.

- The algorithm terminates unsuccesfully when a maximum number of elements is reached, and succesfully when a prescribed goal $\varepsilon_{\text{goal}}$ for ε_ℓ is achieved.

- If the algorithm has terminated succesfully we usually "postprocess" the solution by performing a few additional mesh refinements (but fixing ε) using strategy (i) to confirm that the penalty energy and support of $\xi^+_\ell + \xi^-_\ell$ tend to zero.

- After ε_ℓ is decreased in step (3.1) we only expect a small change in the solution. Therefore the optimization should essentially behave like Newton's method and terminate in few steps. We therefore set the maximum number of iterations to a relatively small number (say $N^1_{\text{opt}} = 20$). This setting prevents us from spending many iterations on finding an entirely new equilibrium when ε_ℓ becomes too small for the current mesh and the penalty solution ceases to be a local minimizer.

- On the other hand, after the mesh is refined in either step (3.2) or (3.3) we expect a large change in the solution because the support of $\xi^+ + \xi^-$ may shrink and we therefore allow a larger number of iterations (say $N^2_{\text{opt}} = 10^6$, but we usually observe termination in far fewer iterations).

We have not addressed the question under which the algorithm is considered to have failed. When no Lavrentiev phenomenon occurs, we observe, in general, that for large ε a state satisfying the requirement $\widetilde{E}_{\varepsilon_\ell} \leqslant E_{\text{goal}}$ is found but that eventually, the algorithm will keep refining the mesh without being able to uphold this bound. We have therefore implemented a safety check which terminates the algorithm when a prescribed number of elements is reached.

As a warning, we also note that for sufficiently large ε it is sometimes possible to find *reasonably looking* solutions which indicate a Lavrentiev gap, but which may disappear as ε becomes small. It is therefore crucial to be able to drive ε as close to zero as possible.

7.4. Maniá-type examples

In this section we present numerical results for one-dimensional problems of the type

$$\begin{aligned} E(v) = & \int_0^1 \Big(|v_x|^n (v^m - x^k)^2 + \nu |v_x|^2 \Big) dx \\ V = & \big\{ v \in W^{1,1}(0,1) : v(0) = 0, v(1) = 1 \big\} = \mathrm{id} + W_0^{1,1}(0,1), \end{aligned} \tag{7.4}$$

where $k, m, n \in \mathbb{N}$ and $\nu \geqslant 0$. This class includes in particular Maniá's original example [23] ($n = 6, m = 3, k = 1, \nu = 0$), and the *regular* example of Ball and Mizel [6, 7] ($n = 14, m = 3, k = 2, 0 < \nu < 2.4 \times 10^{-3}$). The idea behind these examples is that, for $\nu = 0$ the infimum of the energy is always zero with exact solution $u^*(x) = x^{k/m}$, but that the power n can be chosen large to make approximation *difficult*. Moreover, if m and k are chosen such that $u^* \in H^1(0,1)$ then a perturbation of the functional with sufficiently small positive ν does not change whether E exhibits a Lavrentiev phenomenon or not [6, 7].

The $x^{1/3}$ singularity for the original Maniá example is expensive (though not impossible) to resolve and so we have chosen to compute the solution for $n = 8, m = 2, k = 1, \nu = 0$ instead. We have plotted an accurate Galerkin solution, the solution of the penalty method for $\varepsilon = 10^{-1}$ and $\sharp\mathcal{T} = X$ in Fig. 7.1, and the iterations of the contributions to the energy of the penalty method as well as the support of $\xi_\ell^+ + \xi_\ell^-$ in Fig. 7.2.

In addition, we also computed the solution for the regular example of Ball and Mizel with $n = 14, m = 3, k = 2, \nu = 10^{-3}$, and we have plotted the solution in Fig. 7.3. The evolution of the energy and of the support of the penalty variable is similar as in the previous example.

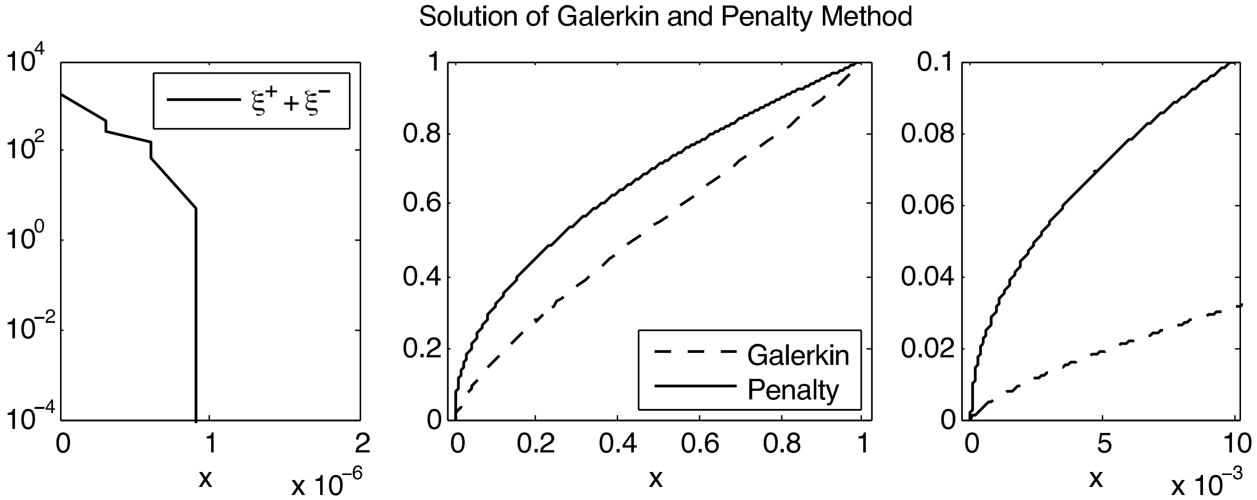


Fig. 7.1. Final solutions of the Galerkin and the Penalty methods for the Maniá problem (7.4) with parameters $n = 8, m = 2, k = 1, \nu = 0$ *before* the reduction step. The error of the penalty solution in the L^∞-norm is $\|u_\ell - u^*\|_{L^\infty} \approx 7.83 \times 10^{-5}$

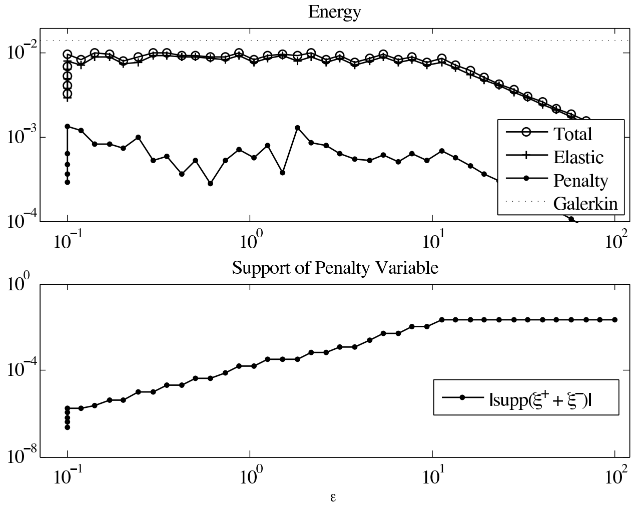


Fig. 7.2. Evolution of the contributions to the penalty energy $\widetilde{E}_{\varepsilon_\ell}$ and of the support of the penalty variables at each step of the continuation algorithm outlined in Section 7.3. The clear convergence of $|\mathrm{supp}(\xi^+ + \xi^-)|$ to zero is a strong indicator for the convergence of the method

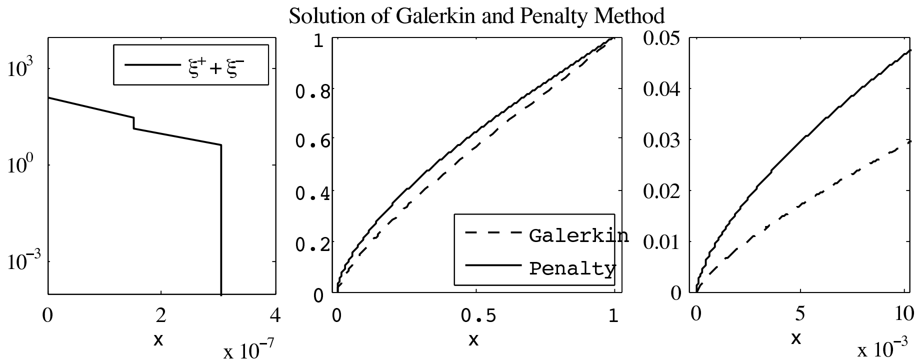


Fig. 7.3. Final solution of the Galerkin and Penalty methods for Ball and Foss' [7] version of the Maniá problem (7.4) with parameters $n = 17, m = 3, k = 2, \nu = 10^{-3}$ *before* the reduction step. The different orders of the singularity at the origin are a clear indication for a Lavrentiev gap

7.5. A convex example in 2D

In this section, we present numerical results for a modification of the example provided by Foss, Hrusa and Mizel [16]. In their original example a semi-circle Ω is transformed into a quarter-circle $y(\Omega)$ with stored energy

$$E(y) = \int\limits_{\Omega} \Big[\big(|Dy|^2 - 2 \det Dy \big)^4 + \nu \big(\frac{\kappa}{\det Dy} + 3^{2-\kappa 2} (1 + |Dy|^2)^{\kappa/2} \big) \Big] dx,$$

where κ and ν are parameters, creating a singularity at the origin. The idea of the example is similar as in the regular examples of Ball and Mizel. For $\nu = 0$ the map $y^*(x) =$

$r^{1/2}(\cos(\theta/2), \sin(\theta/2))$ gives zero energy but the large power makes approximation difficult and it can be shown that the problem exhibits the Lavrentiev phenomenon. Further, the deformation y^* has finite energy for $\nu > 0$ and hence, for ν sufficiently small the Lavrentiev effect remains [16].

We note that the map $F \mapsto (|F|^2 - 2\det F)$ is a non-negative quadratic form and hence the stored energy density

$$W_0(F) = \big(|F|^2 - 2\det F\big)^4,$$

is convex. The polyconvex terms are fairly unimportant for the Lavrentiev effect and hence we decided to ignore them completely (though we should mention that we also performed succesful computations with the full Foss/Hrusa/Mizel example). Instead, upon noting that $y^* \in H^1(\Omega)$ we regularize W_0 by a quadratic and define

$$\begin{aligned} E(v) = &\int_\Omega \big[W_0(Dy) + \nu|Dy|^2\big]\,dx, \qquad (7.5)\\ V = &\big\{v \in W^{1,1}(\Omega) : v(x) = r(\cos(\theta/2), \sin(\theta/2)) \text{ if } |x| = 1,\\ &v_1(\{x_2 = 0, x_1 < 0\}) = \{0\} \text{ and } v_2(\{x_2 = 0, x_1 > 0\} = \{0\}\big\}. \end{aligned}$$

The solution and the evolution of the energy during optimization for the case $\nu = 0$ are plotted in Fig. 7.4, 7.5 and 7.6. For the case $\nu = 10^{-3}$, we have only plotted the radial component of the solution in Fig. 7.7. The evolutions of energy and support of the penalty variables during the optimization is similar as in Fig. 7.6.

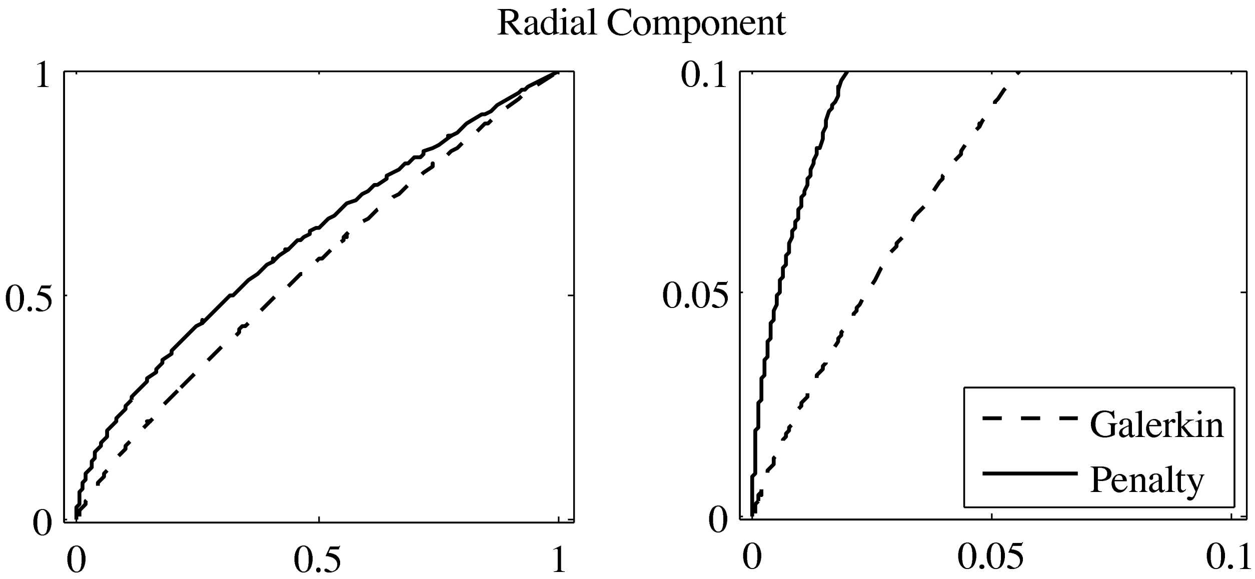


Fig. 7.4. Radial components of the solution of the Galerkin and the Penalty methods for the modified Foss/Hrusa/Mizel problem (7.5) with $\nu = 0$ *before* the reduction step. The different orders in the singularities at the origin are a clear indicator for a Lavrentiev gap. The error of the penalty solution *before* reduction is $\|u_\ell - u^*\|_{L^\infty} \approx 1.07 \times 10^{-1}$

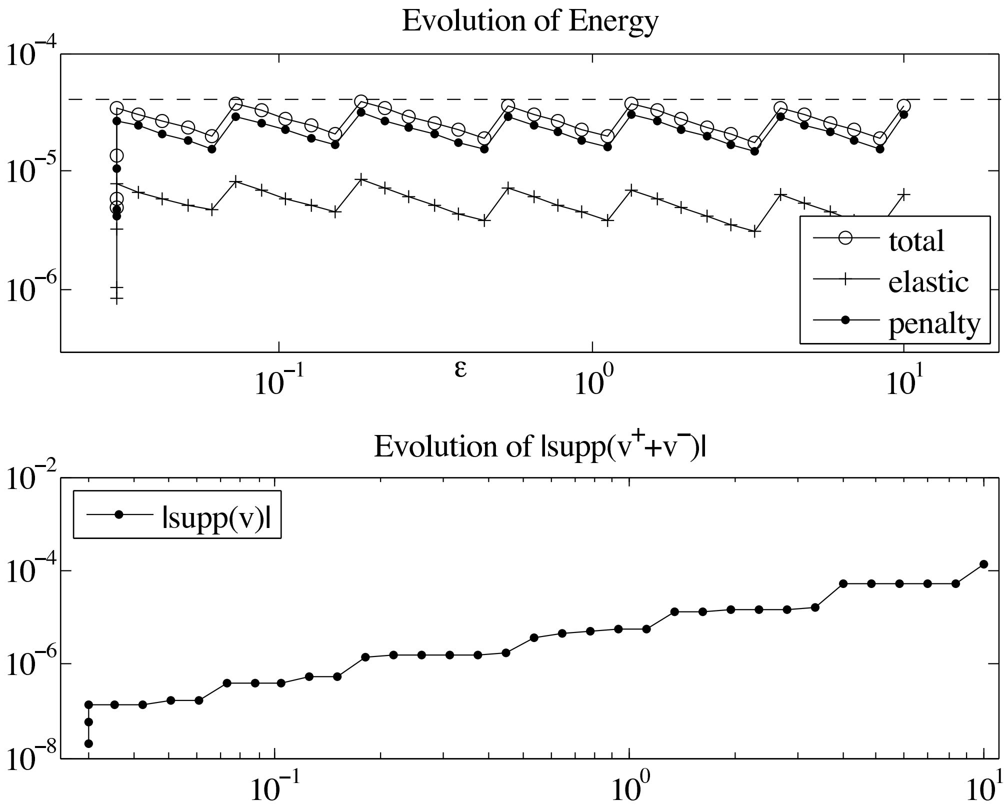


Fig. 7.6. Evolution of the contributions to the penalty energy $\widetilde{E}_{\varepsilon_\ell}$ and of the support of the penalty variables at each step of the continuation algorithm outlined in Section 7.3. The apparent convergence of $|\mathrm{supp}(\xi^+ + \xi^-)|$ to zero is a strong indicator for the convergence of the method

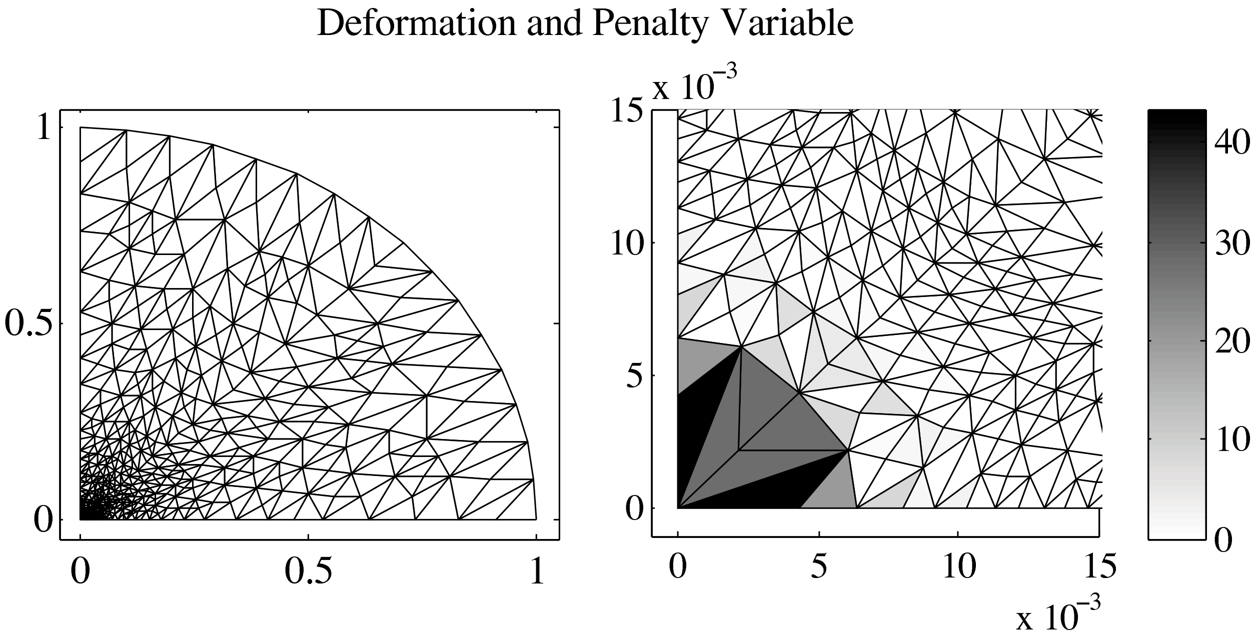


Fig. 7.5. Plot of the deformation given by the solution of the Penalty method for the Foss/Hrusa/Mizel problem (7.5) with $\nu = 0$. The shade of the elements represents the size of the penalty variables $\xi^+ + \xi^-$

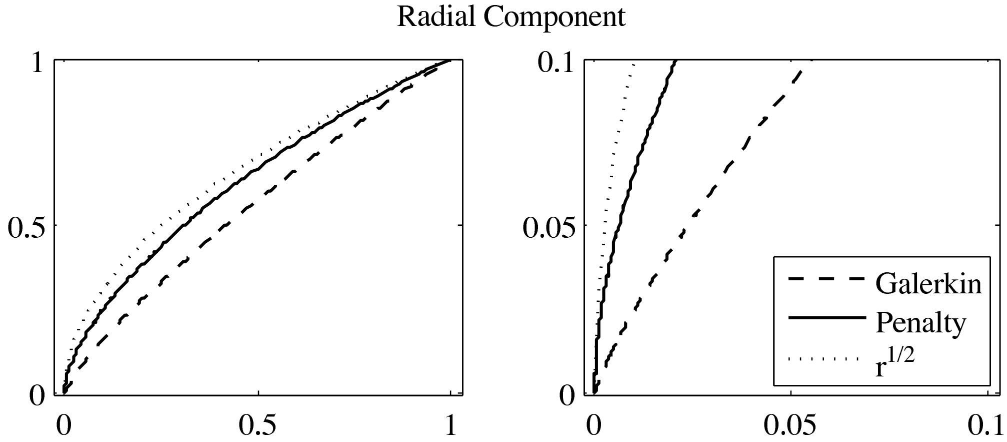


Fig. 7.7. Radial components of the solution of the Galerkin and the Penalty methods for the modified Foss/Hrusa/Mizel problem (7.5) with $\nu = 0$ *before* the reduction step. The different orders in the singularities at the origin are a clear indivator for a Lavrentiev gap. For comparison, the exact solution for the case $\nu = 0$ is plotted as well

To conclude, we briefly outline the result of an experiment that does not exhibit a Lavrentiev gap. We modify (7.5) as follows:

$$E(v) = \int_\Omega \big[W_0(Dy) + \nu|Dy|^p\big]\,dx,$$

keeping the same admissible set V. We choose $\nu = 1/60$ and $p = 6$ a case for which numerical experiments in [25, 26] indicate the absence of a Lavrentiev gap.

An adaptive Galerkin solution suggests that the infimum of the energy in the space of Lipschitz functions is approximately $\inf E(V \cap W^{1,\infty}(\Omega;\mathbb{R}^2)) \approx 0.0093 + O(3 \times 10^4)$. Hence, we try to minimize the penalty functional with target energy $E_{\mathrm{goal}} = 0.0085$. We observe that up to $\varepsilon \approx 2$ the algorithm behaves similar as in the case $p = 2$ above. However, at this point it stagnates and is unable to lower the penalty parameter further without increasing the energy above E_{goal}. This is strong indication that no Lavrentiev gap exists or, more precisely, that no gap larger than 10^{-3} exists, which is consistent with [25, 26].

Next, we considered the case $\nu = 1/40$ and $p = 4$. This is a borderline case that is particularly difficult to resolve. In this case the adaptive Galerkin solution suggests that $\inf E(V \cap W^{1,\infty}(\Omega;\mathbb{R}^2)) \approx 0.0212 + O(3 \times 10^4)$. We tried to minimize the penalty functional with $E_{\mathrm{goal}} = 0.02$. Our algorithm once again managed to decrease the penalty parameter to approximately $\varepsilon \approx 1.8$ but not further, thus indicating the absence of a Lavrentiev gap. However, this is in contradiction with the numerical experiments shown in [26]. Due to the relative simplicity of the method used in [26] it is conceivable that their results are correct, and thus shows that in particularly difficult borderline cases our method may still require some improvements.

Acknowledgments. This work was supported by the DFG Research Center MATHEON at Berlin, Germany, the World Class University (WCU) program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology R31-2008-000-10049-0, the German Research Group Microplast (Forschergruppe 797 der Deutschen Forschungsgemeinschaft), the Hausdorff Institute of Mathematics at Bonn, Germany, and the EPSRC Critical Mass programme 'New Frontiers in the Mathematics of Solids' (OxMOS), Oxford, UK.

References

1. Y. Bai and Z. Li, *Numerical solution of nonlinear elasticity problems with Lavrentiev phenomenon*, Math. Models Methods Appl. Sci., **17** (2007), no. 10, pp. 1619–1640.

2. Y. Bai and Z.-P. Li, *A truncation method for detecting singular minimizers involving the Lavrentiev phenomenon*, Math. Models Methods Appl. Sci., **16** (2006), no. 6, pp. 847–867.

3. J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., **63** (1976/77), no. 4, pp. 337–403.

4. J. M. Ball, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London Ser. A, **306** (1982), no. 1496, pp. 557–611.

5. J. M. Ball and G. Knowles, *A numerical method for detecting singular minimizers*, Numer. Math., **51** (1987), no. 2, pp. 181–197.

6. J. M. Ball and V. J. Mizel, *Singular minimizers for regular one-dimensional problems in the calculus of variations*, Bull. Amer. Math. Soc. (N.S.), **11** (1984), no. 1, pp. 143–146.

7. J. M. Ball and V. J. Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rational Mech. Anal., **90** (1985), no. 4, pp. 325–388.

8. W. Bangerth and R. Rannacher, *Adaptive finite element methods for differential equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2003.

9. A. Braides, *Γ-convergence for beginners*, vol. 22 of *Oxford Lecture Series in Mathematics and its Applications*, Oxford University Press, Oxford, 2002.

10. S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of *Texts in Applied Mathematics*, 3rd edn., Springer, New York, 2008.

11. P. G. Ciarlet, *Mathematical elasticity. Vol. I*, vol. 20 of *Studies in Mathematics and its Applications*, North-Holland Publishing Co., Amsterdam, 1988, three-dimensional elasticity.

12. B. Dacorogna, *Direct methods in the calculus of variations*, vol. 78 of *Applied Mathematical Sciences*, Springer-Verlag, Berlin, 1989.

13. G. Dal Maso, *An introduction to Γ-convergence*, Progress in Nonlinear Differential Equations and their Applications, 8, Birkhäuser Boston Inc., Boston, MA, 1993.

14. N. Dunford and J. T. Schwartz, *Linear operators. Part I*, Wiley Classics Library, John Wiley & Sons Inc., New York, 1988, general theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

15. L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

16. M. Foss, W. J. Hrusa, and V. J. Mizel, *The Lavrentiev gap phenomenon in nonlinear elasticity*, Arch. Ration. Mech. Anal., **167** (2003), no. 4, pp. 337–365.

17. G. Knowles, *Finite element approximation to singular minimizers, and applications to cavitation in nonlinear elasticity*, in: *Differential equations and mathematical physics (Birmingham, Ala., 1986)*, vol. 1285 of *Lecture Notes in Math.*, Springer, Berlin, 1987, pp. 236–247.

18. A. Lavrentiev, *Sur quelques problémes du calcul des variations*, Ann. Mat. Pura Appl., **41** (1926), pp. 107–124.

19. Z.-P. Li, *Element removal method for singular minimizers in variational problems involving Lavrentiev phenomenon*, Proc. Roy. Soc. London Ser. A, **439** (1992), no. 1905, pp. 131–137.

20. Z.-P. Li, *Element removal method for singular minimizers in problems of hyperelasticity*, Math. Models Methods Appl. Sci., **5** (1995), no. 3, pp. 387–399.

21. Z.-P. Li, *A numerical method for computing singular minimizers*, Numer. Math., **71** (1995), no. 3, pp. 317–330.

22. C.-J. Lin and J. J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., **9** (1999), no. 4, pp. 1100–1127 (electronic), dedicated to John E. Dennis, Jr., on his 60th birthday.

23. B. Maniá, *Sopra un esempio di Lavrentieff*, Boll. Un. Mat. Ital., **13** (1934), pp. 147–153.

24. P. V. Negrón-Marrero, *A numerical method for detecting singular minimizers of multidimensional problems in nonlinear elasticity*, Numer. Math., **58** (1990), no. 2, pp. 135–144.

25. C. Ortner, *Non-conforming finite element discretisation of convex variational problems*, to appear in IMA J. Numer. Anal.

26. C. Ortner and D. Praetorius, *On the convergence of adaptive non-conforming finite element methods*, Tech. rep., Mathematical Institute, University of Oxford, 2008, oxMOS Preprint 14.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 164–176

ADDITIVE AVERAGE SCHWARZ METHODS FOR DISCRETIZATION OF ELLIPTIC PROBLEMS WITH HIGHLY DISCONTINUOUS COEFFICIENTS

M. DRYJA[1] AND M. SARKIS[2]

Abstract — A second order elliptic problem with highly discontinuous coefficients has been considered. The problem is discretized by two methods: 1) continuous finite element method (FEM) and 2) composite discretization given by a continuous FEM inside the substructures and a discontinuous Galerkin method (DG) across the boundaries of these substructures. The main goal of this paper is to design and analyze parallel algorithms for the resulting discretizations. These algorithms are additive Schwarz methods (ASMs) with special coarse spaces spanned by functions that are almost piecewise constant with respect to the substructures for the first discretization and by piecewise constant functions for the second discretization. It has been established that the condition number of the preconditioned systems does not depend on the jumps of the coefficients across the substructure boundaries and outside of a thin layer along the substructure boundaries. The algorithms are very well suited for parallel computations.

2000 Mathematics Subject Classification: 65F10; 65N20; 65N30.

Keywords: domain decomposition methods, additive Schwarz method, finite element method, discontinuous Galerkin method, elliptic problems with highly discontinuous coefficients, heterogeneous coefficients.

1. Introduction

In this paper, a second order elliptic problem with a highly discontinuous coefficient $\varrho(x)$ in a 2-D polygonal region Ω is considered. For simplicity of the presentation we assume Dirichlet homogeneous boundary conditions. The region Ω is partitioned into disjoint polygonal substructures $\Omega_i, \overline{\Omega} = \cup_i \overline{\Omega}_i,\ i = 1, \cdots, N$, and we denote by $\varrho_i(x)$ the restriction of $\varrho(x)$ to Ω_i. For this partition, let us denote by Ω_i^h the layer around $\partial\Omega_i$ with width h_i and define $\overline{\alpha}_i = \sup_{x\in\Omega_i^h} \varrho_i(x)$ and $\underline{\alpha}_i = \inf_{x\in\Omega_i^h} \varrho_i(x)$. We say that the coefficient $\varrho_i(x)$ has moderate variations on Ω_i^h if $\overline{\alpha}_i/\underline{\alpha}_i = O(1)$. The coefficient ϱ can be highly discontinuous in $\Omega_i \backslash \Omega_i^h$ and across $\partial\Omega_i$.

We consider two discretization methods: the standard continuous finite element method (FEM), see [3], and a composite discretization FEM with discontinuous Galerkin (DG) (see [1, 10]. The latter means that in each Ω_i the problem is discretized by a continuous FEM inside Ω_i and a DG method across $\partial\Omega_i$, see [4, 5]. This discretization is determined by the regularity of $\varrho(x)$ and the regularity of the solution.

[1] *Department of Mathematics, Warsaw University, Banacha 2, 00-097 Warsaw, Poland.* E-mail: dryja@mimuw.edu.pl

[2] *Instituto Nacional de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro 22460-320, Brazil; Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA01609, USA.* E-mail: msarkis@impa.br and msarkis@wpi.edu

The main goal of this paper is to design and analyze parallel algorithms for these two considered discretizations. They are additive Schwarz methods (ASMs) with coarse space functions which are piecewise constant on each $\Omega_i \backslash \Omega_i^h$ for the first discretization and piecewise constant on each Ω_i for the second one. The unknowns associated with these coarse spaces are related to the average values on $\partial\Omega_i$. These algorithms are called additive average Schwarz methods (AASMs) and they are generalizations of the algorithms considered in [2] for the case of continuous FEMs and for regular coefficients.

In this paper, we have proved that the condition number of the preconditioned systems obtained by AASMs for the first discretization is bounded by $C \max_i (\frac{H_i}{h_i})^2 \frac{\overline{\alpha}_i}{\underline{\alpha}_i}$, where C is independent of the jumps of ϱ, the size of the substructures $H_i := diam(\Omega_i)$ and the triangulation parameters h_i in Ω_i, $i = 1, \cdots, N$. For the second discretization (the composite discretization), we have proved that the condition number is bounded by $C \max_i \max_{j \in \mathcal{I}_i} (\frac{H_i^2}{h_i h_{ij}}) \frac{\overline{\alpha}_i}{\underline{\alpha}_i}$ where $\mathcal{I}_i$ is a set of indices j such that $|\partial\Omega_i \cap \partial\Omega_j| \neq 0$ and $h_{ij} := 2h_i h_j/(h_i + h_j)$, as the harmonic average of h_i and h_j. These estimates can be improved when $\underline{\alpha}_i$ and $\overline{\alpha}_i$ are of the same order and $\overline{\alpha}_i \leqslant \varrho_i(x)$ on $\Omega_i \backslash \Omega_i^h$. In this case, we get estimates with $C \max_i (H_i/h_i)$ for the first discretization and $C \max_i \max_{j \in \mathcal{I}_i} (H_i/h_{ij})$ for the second one.

The discussed algorithms can be straightforwardly extended to the 3-D case. In this paper, the 2-D case is considered only for the simplicity of the presentation.

Parallel algorithms for the considered discretizations in the case of piecewise constant coefficients with respect to Ω_i have been discussed in many papers (see [11] and the references therein). The case of coefficients with highly discontinuous coefficients inside Ω_i and across $\partial\Omega_i$ has been discussed only in a few papers. For the first discretization, the standard Schwarz method with overlap and the FETI method were considered in [8] and [9], respectively. In [6], the FETI-DP was discussed, where the estimate of the condition number of the preconditioned system is better than in [9]. In the present paper, we consider simpler coarse spaces and smaller local problems than in the papers mentioned above and with better condition number estimates. For the second discretization, parallel algorithms have not been discussed in the literature to our knowledge, i.e., in the case where the coefficients are highly discontinuous inside of Ω_i and across $\partial\Omega_i$. In the literature, only the case where $\varrho(x)$ is piecewise constant with respect Ω_i has been discussed (see for example [7], [5] and the references therein).

This paper is organized as follows. In Section 2, the differential problem and assumptions about the triangulations and coefficients are introduced. In Section 3, the continuous finite element discretization on matching triangulation is formulated, and in Section 4, an additive average Schwarz method (AASM) for the resulting discrete problem is designed and analyzed. The main result is Theorem 4.1, where we establish the estimate of the condition number of the preconditioned system. In Section 5, the original problem is discretized on nonmatching triangulation across $\partial\Omega_i$ by a continuous FEM in each Ω_i and DG with an interior penalty term across $\partial\Omega_i$, and in Section 6, we design and analyze an AASM for the resulting discrete problem. The main result is Theorem 6.1, where we estimate the condition number of the preconditioned system. In Section 7, we discuss the implementation of these preconditioned systems.

2. Differential problems and assumptions

In this section, we formulate a differential problem with discontinuous coefficient and describe some of the assumptions about the coefficients and triangulations.

2.1. Differential problem

Find $u^* \in H_0^1(\Omega)$ such that

$$a(u^*, v) = f(v), \qquad v \in H_0^1(\Omega), \tag{2.1}$$

where

$$a(u, v) := (\varrho(\cdot)\nabla u, \nabla v)_{L^2(\Omega)}, \qquad f(v) := \int\limits_{\Omega} f v dx. \tag{2.2}$$

We assume that $\varrho \in L^\infty(\Omega)$ and $\varrho(x) \geqslant \varrho_0 > 0, f \in L^2(\Omega)$, and Ω is a 2-D polygonal region. Under these assumptions the problem has a unique solution (see, e.g., [3]).

2.2. Assumptions

We suppose that Ω is decomposed into disjoint polygonals Ω_i, $\overline{\Omega} = \cup_i \overline{\Omega}_i, i = 1, \cdots, N$. Inside each Ω_i we introduce a shape regular and quasi-uniform triangulation $\mathcal{T}^h(\Omega_i)$ with mesh parameter h_i and $H_i := diam(\Omega_i)$. For the first discretization we assume that the global mesh is regular (no hanging nodes) while for the second discretization we allow nonmatching meshes across substructure boundaries. Denote Ω_i^h as the layer around $\partial\Omega_i$ which is a union of $e_k^{(i)}$ triangles of $\mathcal{T}^h(\Omega_i)$ which touch $\partial\Omega_i$, and we introduce

$$\overline{\alpha}_i := \sup_{x \in \overline{\Omega}_i^h} \varrho(x), \qquad \underline{\alpha}_i := \inf_{x \in \overline{\Omega}_i^h} \varrho(x). \tag{2.3}$$

3. Discrete continuous problem

To define the first discretization, the continuous finite element method for problem (2.1), we introduce the space of piecewise linear continuous functions as

$$V_h(\Omega) := \{v \in C_0(\Omega); v_{|e_k} \in P_1(x)\},$$

where e_k are the triangles of $\mathcal{T}^h(\Omega)$ and $P_1(x)$ is a set of linear polynomials.

The discrete problem is defined as: Find $u_h^* \in V_h(\Omega)$ such that

$$a(u_h^*, v) = f(v), \quad v \in V_h(\Omega). \tag{3.1}$$

4. Additive average Schwarz method for (3.1)

In this section, we design and analyze an additive average Schwarz method for the discrete problem (3.1). For that we use the general theory of additive Schwarz methods (ASMs) described in [11].

4.1. Decomposition of $V_h(\Omega)$

Let us decompose

$$V_h(\Omega) = V_0(\Omega) + V_1(\Omega) + \cdots + V_N(\Omega), \tag{4.1}$$

where for $i = 1, \cdots, N$, we define $V_i(\Omega) = V_h(\Omega) \cap H_0^1(\Omega_i)$ on Ω_i and extended by zero outside of Ω_i. The coarse space $V_0(\Omega)$ is defined as the range of the following interpolation operator I_A. For $u \in V_h(\Omega)$, let $I_A u \in V_h(\Omega)$ be defined so that on $\overline{\Omega}_i$

$$I_A u := \begin{cases} u(x), & x \in \partial\Omega_{ih} \\ \bar{u}_i, & x \in \Omega_{ih} \end{cases}, \tag{4.2}$$

where

$$\bar{u}_i := \frac{1}{n_i} \sum_{x \in \partial\Omega_{ih}} u(x). \tag{4.3}$$

Here Ω_{ih} and $\partial\Omega_{ih}$ are the sets of nodal points of Ω_i (interior) and $\partial\Omega_i$, respectively, and n_i is the number of nodal points of $\partial\Omega_{ih}$.

4.2. Inexact solvers

For $i = 1, \cdots, N$, let us introduce

$$b_i(u, v) := a_i(u, v), \quad u, v \in V_i(\Omega), \tag{4.4}$$

and $a_i(\cdot, \cdot)$ is the restriction of $a(\cdot, \cdot)$ to Ω_i.

For $i = 0$, let us introduce

$$b_0(u, v) := \sum_{i=1}^{N} \sum_{x \in \partial\Omega_{ih}} \overline{\alpha}_i (u(x) - \bar{u}_i)(v(x) - \bar{v}_i), \quad u, v \in V_0(\Omega). \tag{4.5}$$

Note that (4.5) reduces to

$$b_0(u, v) = \sum_{i=1}^{N} \overline{\alpha}_i \sum_{x \in \partial\Omega_{ih}} (u(x) - \bar{u}_i) v(x). \tag{4.6}$$

4.3. Operator equation

For $i = 0, \cdots, N$, we define the operators $T_i^{(A)} : V_h(\Omega) \to V_i(\Omega)$ by

$$b_i(T_i^{(A)} u, v) = a(u, v), \quad v \in V_i(\Omega). \tag{4.7}$$

Of course, each of these problems has a unique solution. Let us introduce

$$T_A := T_0^{(A)} + T_1^{(A)} + \cdots + T_N^{(A)}. \tag{4.8}$$

We replace (3.1) by the operator equation

$$T_A u_h^* = g_h, \tag{4.9}$$

where

$$g_h = \sum_{i=0}^{N} g_i, \qquad g_i = T_i^{(A)} u_h^*, \tag{4.10}$$

and u_h^* is the solution of (3.1). Note that to compute g_i we do not need to know u_h^*, see (4.7). We note also that the solutions of (3.1) and (4.9) are the same. This follows from the first main result of this paper:

Theorem 4.1. *For any $u \in V_h(\Omega)$ the following holds:*

$$C_1\beta_1^{-1}a(u,u) \leqslant a(T_Au,u) \leqslant C_2 a(u,u), \tag{4.11}$$

where $\beta_1 = \max_i(\overline{\alpha}_i/\underline{\alpha}_i)(H_i/h_i)^2$ and the positive constants C_1 and C_2 do not depend on ϱ_i, $\overline{\alpha}_i/\underline{\alpha}_i$, H_i, and h_i, $i = 1,\cdots,N$.

Remark 4.1. The estimate (4.11) can be improved when $\overline{\alpha}_i$ and $\underline{\alpha}_i$ are of the same order and $\underline{\alpha}_i \leqslant \varrho_i(x)$ on $\Omega_i\backslash\Omega_i^h$. In this case, $\beta_1 = \max_i(H_i/h_i)$.

Remark 4.2. The layer Ω_i^h can be replaced by Ω_i^δ, the layer around $\partial\Omega_i$ with width δ_i. In this case, $\beta_1 = \max_i(\frac{\overline{\alpha}_i}{\underline{\alpha}_i}\frac{H_i^2}{h_i\delta_i})$ where $\overline{\alpha}_i$ and $\underline{\alpha}_i$ here are defined on Ω_i^δ (see [6]).

Proof. of Theorem 4.1. To this end, we need to check the three key assumptions of the general theory of ASMs (see Theorem 2.7 of [11]). □

Assumption(i) We need to show that $\eta(\varepsilon)$, the spectral radius of $\varepsilon = \{\varepsilon_{ij}\}_{i,j=1,\cdots,N}$, defined by

$$a(u_i,u_j) \leqslant \varepsilon_{ij}a^{1/2}(u_i,u_i)a^{1/2}(u_j,u_j) \quad \forall u_i \in V_i \quad \text{and} \quad \forall u_j \in V_j,$$

is bounded by a constant that does not depend on the jumps of $\varrho_i(x), H_i$ and h_i. In our case, V_i and V_j are orthogonal for $i,j = 1,\cdots,N$ and $i \neq j$, therefore, $\eta(\varepsilon) = 1$.

Assumption (ii) We need to show that for $i = 0,\cdots,N$,

$$a(u,u) \leqslant \omega_i b_i(u,u), \qquad u \in V_i,$$

with $\omega_i \leqslant C$ where C is independent of the jumps of $\varrho_i(x), H_i$ and h_i.

For $i = 1,\cdots,N$, it is obvious that $\omega_i = 1$. For $i = 0$ and $u \in V_h(\Omega)$ we have

$$a(I_Au,I_Au) = \sum_{i=1}^N a_i(I_Au,I_Au),$$

and (see (4.2))

$$\begin{aligned} a_i(I_Au,I_Au) &\equiv (\varrho_i(\cdot)\nabla I_Au, \nabla I_Au)_{L^2(\Omega_i)} = \\ &= (\varrho_i(\cdot)\nabla(I_Au-\bar{u}_i), \nabla(I_Au-\bar{u}_i))_{L^2(\Omega_i)} = \\ &= (\varrho_i(\cdot)\nabla(I_Au-\bar{u}_i), \nabla(I_Au-\bar{u}_i))_{L^2(\Omega_i^h)} \leqslant \\ &\leqslant C\sum_{x\in\partial\Omega_{ih}} \overline{\alpha}_i(u_i(x)-\bar{u}_i)^2, \end{aligned} \tag{4.12}$$

where $\overline{\alpha}_i$ is defined in (2.3). We have used the inverse inequality. Hence

$$a(I_Au,I_Au) \leqslant Cb_0(u,u),$$

with $\omega_0 \leqslant C$. Thus $\max_{i=0}^N \omega_i \leqslant C$.

Assumption(iii) We prove that for $u \in V_h(\Omega)$ there exist $u_i \in V_i, i = 0.\cdots, N$, such that $u = \sum_{i=0}^N u_i$ and

$$\sum_{i=0}^N b_i(u_i, u_i) \leqslant C\beta_1 a(u,u). \tag{4.13}$$

Let $u_0 := I_A u$ for $u \in V_h(\Omega)$ and $u_i := u - u_0$ on $\overline{\Omega}_i$ and $u_i = 0$ outside of Ω_i. Of course, $u_i \in V_i(\Omega)$ for $i = 0, \cdots, N$, and $u = \sum_{i=0}^N u_i$. We have

$$\sum_{i=1}^N b_i(u_i, u_i) = \sum_{i=1}^N a_i(u - u_0, u - u_0) \leqslant 2\sum_{i=1}^N \{a_i(u,u) + a_i(u_0,u_0)\} = 2\{a(u,u) + a(u_0,u_0)\}. \tag{4.14}$$

To obtain β_1 in (4.13), we only need to estimate $a(u_0, u_0)$. We have

$$\begin{aligned} a_i(u_0,u_0) &\leqslant C \sum_{x\in\partial\Omega_{ih}} \overline{\alpha}_i (u(x) - \bar{u}_i)^2 \\ &\qquad\qquad \leqslant C\frac{\overline{\alpha}_i}{h_i} \parallel u - \bar{u}_i \parallel^2_{L^2(\partial\Omega_i)} \leqslant C\frac{H_i^2}{h_i}\overline{\alpha}_i |u|^2_{H^1(\partial\Omega_i)}, \end{aligned} \tag{4.15}$$

where we have used (4.12) and Friedrich's inequality. Note that

$$\overline{\alpha}_i |u|^2_{H^1(\partial\Omega_i)} \leqslant \frac{\overline{\alpha}_i}{\underline{\alpha}_i h_i} (\varrho_i(\cdot)\nabla u, \nabla u)_{L^2(\Omega_i^h)}. \tag{4.16}$$

Using this in (4.15) we obtain

$$\sum_{i=1}^N a_i(u_0,u_0) \leqslant \sum_{i=1}^N C\frac{\overline{\alpha}_i}{\underline{\alpha}_i}\frac{H_i^2}{h_i^2} a_i(u,u) \leqslant C\beta_1 a(u,u). \tag{4.17}$$

Using this in (4.14), we obtain (4.13). This completes the proof of Theorem 4.1.

5. Discrete discontinuous Galerkin problem

In this section, the original problem (2.1) is discretized by a composite discretization. We decompose Ω into disjoint polygonals $\Omega_i, i = 1, \cdots, N$, so $\overline{\Omega} = \cup_i \overline{\Omega}_i$ as in Section 4 and define $H_i = diam(\Omega_i)$. The problem (2.1) is discretized by a continuous FEM in each Ω_i and by a DG across $\partial\Omega_i$.

Let us introduce a triangulation $\mathcal{T}^h(\Omega_i)$ in each Ω_i with triangular elements $e_k^{(i)}$ and a mesh parameter h_i. We assume that this triangulation is shape-regular on $\overline{\Omega}_i$. The resulting triangulation is nonmatching across $\partial\Omega_i$. Let $X_i(\Omega_i)$ be the finite element space of piecewise linear continuous functions on Ω_i. We do not assume that the functions of $X_i(\Omega_i)$ vanish on $\partial\Omega_i \cap \partial\Omega$. Let us introduce

$$X_h(\Omega) := X_1(\Omega_1) \times \cdots \times X_N(\Omega). \tag{5.1}$$

The functions v of $X_h(\Omega)$ are represented as $v = \{v_i\}_{i=1}^N$ with $v_i \in X_i(\Omega_i)$. Note that $X_h(\Omega) \nsubseteq H^1(\Omega)$ but $X_h(\Omega) \subset L_2(\Omega)$.

The coefficients $\varrho(x)$ on the introduced triangulation can be discontinuous. We assume that $\varrho(x)$ on each element $e_k^{(i)} \subset \overline{\Omega}_i$ is a constant $\varrho_k^{(i)}$, which can be defined, for example, by $|e_k^{(i)}|^{-1} \int_{e_k^{(i)}} \varrho(x) ds$. It means that this is done in the formulation of the original problem.

Let Ω_i^h, as in Section 2, denote a layer with width h_i around $\partial\Omega_i$ which is the union of $e_k^{(i)}$ triangles that touch $\partial\Omega_i$. We will use also $\overline{\alpha}_i$ and $\underline{\alpha}_i$ defined in (2.3). Note that this time $\varrho(x)$ is piecewise constant on the triangles of Ω_i^h.

A discrete problem for (2.1) is obtained by a composite discretization, i.e., a regular continuous FEM in each Ω_i and a DG across of $\partial\Omega_i$ (see [1, 10, 4, 5]). The discretization is defined as follows: Find $u_h^* \in X_h(\Omega)$ such that

$$\hat{a}_h(u_h^*, v_h) = f(v_h), \qquad v_h \in X_h(\Omega), \tag{5.2}$$

where

$$\hat{a}_h(u,v) := \sum_{i=1}^{N} \hat{a}_i(u,v), \quad f(v) := \sum_{i=1}^{N} \int\limits_{\Omega_i} f v_i dx. \tag{5.3}$$

Each bilinear form $\hat{a}_i$ is given as a sum of three bilinear forms:

$$\hat{a}_i(u,v) := a_i(u,v) + s_i(u,v) + p_i(u,v), \tag{5.4}$$

where

$$a_i(u,v) := \int\limits_{\Omega_i} \varrho_i(x) \nabla u_i \nabla v_i dx, \tag{5.5}$$

$$s_i(u,v) := \sum_{E_{ij} \subset \partial\Omega_i} \frac{1}{l_{ij}} \int\limits_{E_{ij}} \varrho_{ij}(x) \left(\frac{\partial u_i}{\partial n_i}(v_j - v_i) + \frac{\partial v_i}{\partial n_i}(u_j - u_i)\right) ds, \tag{5.6}$$

and

$$p_i(u,v) := \sum_{E_{ij} \subset \partial\Omega_i} \frac{\delta}{l_{ij} h_{ij}} \int\limits_{E_{ij}} \varrho_{ij}(x)(u_j - u_i)(v_j - v_i) ds. \tag{5.7}$$

Here, the bilinear form p_i is called the penalty term with a positive penalty parameter δ. In the above equations, we set $l_{ij} = 2$ when $E_{ij} := \partial\Omega_i \cap \partial\Omega_j$ is a common edge (or part of an edge) of $\partial\Omega_i$ and $\partial\Omega_j$. On E_{ij} we define $\varrho_{ij}(x) = 2\varrho_i(x)\varrho_j(x)/(\varrho_i(x) + \varrho_j(x))$, i.e., as the harmonic average of $\varrho_i(x)$ and $\varrho_j(x)$ on E_{ij}. Similarly, we define $h_{ij} = 2h_i h_j/(h_i + h_j)$. In order to simplify notation we include the index $j = \partial$ when $E_{i\partial} := \partial\Omega_i \cap \partial\Omega$ is an edge of $\partial\Omega$ and set $l_{i\partial} = 1$, $v_\partial = 0$ for all $v \in X_h(\Omega)$, $\varrho_{i\partial}(x) = \varrho_i(x)$ and $h_{i\delta} = h_i$. The outward normal derivative on $\partial\Omega_i$ is denoted by $\partial/\partial n_i$. Note that when $\varrho_{ij}(x)$ is given by the harmonic average, then $\min\{\varrho_i, \varrho_j\} \leqslant \varrho_{ij} \leqslant 2\min\{\varrho_i, \varrho_j\}$.

We also define the positive local bilinear form d_i with weights $\varrho_i(x)$ and $\delta\varrho_{ij}(x)/(l_{ij}h_{ij})$ as

$$d_i(u,v) = a_i(u,v) + p_i(u,v), \tag{5.8}$$

and introduce the global bilinear form $d_h(\cdot,\cdot)$ on $X_h(\Omega)$ defined by

$$d_h(u,v) = \sum_{i=1}^{N} d_i(u,v). \tag{5.9}$$

For $u = \{u_i\}_{i=1}^N \in X_h(\Omega)$ the associated broken norm is then defined by

$$\| u_h \|_h^2 := d_h(u,u) = \sum_{i=1}^N \{\| \varrho_i^{1/2} \nabla u_i \|_{L^2(\Omega_i)}^2 + \sum_{E_{ij}\subset\partial\Omega_i} \frac{\delta}{l_{ij}h_{ij}} \int_{E_{ij}} \varrho_{ij}(x)(u_i - u_j)^2 ds\}. \quad (5.10)$$

The discrete problem (5.2) has a unique solution for a sufficiently large penalty parameter δ. This follows from the following lemma:

Lemma 5.1. *There exists $\delta_0 > 0$ such that for $\delta \geqslant \delta_0$ and for all $u \in X_h(\Omega)$,*

$$\gamma_0 d_i(u,u) \leqslant \hat{a}_i(u,u) \leqslant \gamma_1 d_i(u,u), \quad (5.11)$$

and

$$\gamma_0 d_h(u,u) \leqslant \hat{a}_h(u,u) \leqslant \gamma_1 d_h(u,u), \quad (5.12)$$

hold, where γ_0 and γ_1 are positive constants independent of ϱ_i, h_i and H_i.

Proof. This proof is a slight modification of the proof given in [4, 5], therefore, it is omitted here.

□

We will assume below that $\delta \geqslant \delta_0$; i.e., that (5.11) and (5.12) are valid. The a priori error estimates for the discussed method are optimal for regular coefficients and when h_i and h_j are of the same order (see, e.g., [1], [10]). For piecewise constant coefficients ϱ_i and/or when the mesh sizes h_i and h_j are not of the same order, the error estimates depend on the ratio h_i/h_j. There is also the question of regularity of the solution of (2.1). Assuming the regularity of the solution, we have the following result:

Lemma 5.2. *Let u^* and u_h^* be the solutions of* (2.1) *and* (5.2)*. For $u^* \in H_0^1(\Omega)$ and $u^*_{|\Omega_i} \in H^{1+r}(\Omega_i)), i = 1, \cdots, N$, we have*

$$\| u^* - u_h^* \|_h^2 \leqslant C \sum_{i=1}^N (h_i^{1+r} + \frac{h_j^{2+r}}{h_i}) |u^*|_{H^{1+r}(\Omega_i)}^2,$$

with $r \in [1/2, 1]$ and C which is independent of h_i, H_i and u^*.

For the proof see [1, 10] and [4, 5].

6. Additive average Schwarz method for (5.2)

In this section, we design and analyze an additive average Schwarz method for the discrete problem (5.2). To this end, we use the general theory of additive Schwarz methods (ASMs) described in [11].

6.1. Decomposition of $X_h(\Omega)$

Let us decompose

$$X_h(\Omega) = V^{(0)}(\Omega) + V^{(1)}(\Omega) + \cdots + V^N(\Omega), \tag{6.1}$$

where for $i = 1, \cdots, N$

$$V^{(i)}(\Omega) := \{v = \{v_k\}_{k=1}^N \in X_h(\Omega) : v_k = 0 \quad \text{for} \quad k \neq i\}. \tag{6.2}$$

This means that $V^{(i)}(\Omega)$ is the zero extension of $X_i(\Omega_i)$ to $\overline{\Omega}_j$ for $j \neq i$. The coarse space $V^{(0)}$ is defined as

$$V^{(0)}(\Omega) = span\{\phi^{(i)}\}_{i=1}^N, \tag{6.3}$$

where $\phi^{(i)} = \{\phi_k^{(i)}\}_{k=1}^N \in X_h(\Omega)$ with $\phi_k^{(i)} = 1$ for $k = i$ and $\phi_k^{(i)} = 0$ for $k \neq i$. This is a space of piecewise constant functions with respect to $\Omega_i, i = 1, \cdots, N$. Note that the introduced spaces $V^{(i)}(\Omega)$ satisfy (6.1).

6.2. Inexact solver

For $u^{(i)} = \{u_k^{(i)}\}_{k=1}^N$ and $v^{(i)} = \{v_k^{(i)}\}_{k=1}^N$ belonging to $V^{(i)}(\Omega)$, $i = 1, \cdots N$, we set

$$b_i(u^{(i)}, v^{(i)}) = d_i(u^{(i)}, v^{(i)}), \tag{6.4}$$

where in this case (see (5.8)),

$$d_i(u^{(i)}, v^{(i)}) = (\varrho_i(\cdot)\nabla u_i^{(i)}, \nabla v_i^{(i)})_{L^2(\Omega_i)} + \sum_{E_{ij}\subset\partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot) u_i^{(i)}, v_i^{(i)})_{L^2(E_{ij})}. \tag{6.5}$$

For the coarse space $V^{(0)}$ and $u^{(0)} = \{u_i^{(0)}\}_{i=1}^N$ and $v^{(0)} = \{v_i^{(0)}\}_{i=1}^N$ belonging to $V^{(0)}(\Omega)$ we set

$$b_0(u^{(0)}, v^{(0)}) = d_h(u^{(0)}, v^{(0)}). \tag{6.6}$$

Note that in this case

$$b_0(u^{(0)}, v^{(0)}) = \sum_{i=1}^N \sum_{E_{ij}\subset\partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot)(u_j^{(0)} - v_i^{(0)}), (u_j^{(0)} - v_i^{(0)}))_{L^2(E_{ij})}, \tag{6.7}$$

since $u^{(0)}$ and $v^{(0)}$ are piecewise constant functions with respect to $\Omega_i, i = 1, \cdots, N$.

6.3. The operator equation

For $i = 0, \cdots, N$, let us define the operators $T_i^{(DG)} : X_h(\Omega) \to V^{(i)}(\Omega)$ by

$$b_i(T^{(DG)}u, v) = \hat{a}_h(u, v), \qquad v \in V^{(i)}(\Omega). \tag{6.8}$$

Of course, each of these problems has a unique solution. Let us define

$$T_{DG} = T_0^{(DG)} + T_1^{(DG)} + \cdots + T_N^{(DG)}. \tag{6.9}$$

We replace (5.2) by the following operator equation:

$$T_{DG} u_h^* = g_h, \tag{6.10}$$

where

$$g_h = \sum_{i=0}^{N} g_i, \qquad g_i = T_i^{(DG)} u_h^*, \tag{6.11}$$

and u_h^* is the solution of (5.2). Note that to compute g_i we do not need to know u_h^* (see (6.8)). The solutions (5.2) and (6.10) are the same. This follows from the following theorem, the second main result of this paper.

Theorem 6.1. *For any $u \in X_h(\Omega)$ the following holds:*

$$C_3 \beta_2^{-1} \hat{a}_h(u,u) \leqslant \hat{a}_h(T_{DG}u,u) \leqslant C_4 \hat{a}_h(u,u), \tag{6.12}$$

where $\beta_2 = \max_i \max_{j \in \mathcal{I}_i} (\overline{\alpha}_i / \underline{\alpha}_i)(\frac{H_i^2}{h_i h_{ij}})$ and the positive constants C_3 and C_4 do not depend on ρ_i, $\overline{\alpha}_i/\underline{\alpha}_i$, H_i, and h_i, $i = 1, \cdots, N$.

Proof. We need to check three key assumptions of the general theory of ASMs (see Theorem 2.7 of [11]). □

Assumption(i) We check it in the same way as Assumption(i) in the proof of Theorem 4.1. Thus $\eta(\varepsilon) = 1$.

Assumption(ii) We need to prove that for $i = 0, 1, \cdots, N$

$$\hat{a}_h(u,u) \leqslant \omega_i b_i(u,u), \qquad u \in V^{(i)}(\Omega), \tag{6.13}$$

with $\omega_i \leqslant C$, where C is independent of the jumps of $\varrho_i(x), \varrho_{ij}(x), H_i$ and h_i. Using Lemma 5.1 it is enough to prove (6.13) for $d_h(\cdot,)$ (see (5.9)). For $i = 1, \cdots, N$ and $u^{(i)} \in V^{(i)}(\Omega)$ we have

$$d_h(u^{(i)}, u^{(i)}) = (\varrho_i(\cdot) \nabla u_i^{(i)}, \nabla u_i^{(i)})_{L^2(\Omega_i)} + \sum_{E_{ij} \subset \partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot) u_i^{(i)}, u_i^{(i)})_{L^2(E_{ij})} = b_i(u^{(i)}, u^{(i)}). \tag{6.14}$$

For the coarse space $V^{(0)}(\Omega)$ and $u^{(0)} \in V^{(0)}(\Omega)$

$$d_h(u^{(0)}, u^{(0)}) = \sum_{i=0}^{N} \sum_{E_{ij} \subset \partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot)(u_i^{(0)} - u_j^{(0)}), (u_i^{(0)} - u_j^{(0)}))_{L^2(E_{ij})} = b_0(u^{(0)}, u^{(0)}).$$

Thus $\omega_i \leqslant C$ for $i = 0, \cdots, N$ in view of Lemma 5.1.

Assumption(iii) We need to show that for $u \in X_h(\Omega)$ there exist $u^{(i)} \in V^{(i)}(\Omega), i = 0, \cdots, N$, such that $u = \sum_{i=0}^{N} u^{(i)}$ and

$$\sum_{i=0}^{N} b_i(u^{(i)}, u^{(i)}) \leqslant C \beta_2 \hat{a}_h(u,u). \tag{6.15}$$

Using Lemma 5.1, it is enough to prove (6.15) for $d_h(\cdot,\cdot)$.

For $u = \{u_i\}_{i=1}^N \in X_h(\Omega)$, let

$$u^{(0)} = \{\bar{u}_i\}_{i=1}^N, \qquad \bar{u}_i := \frac{1}{|\partial\Omega_i|} \int_{\partial\Omega_i} u_i(x) ds, \tag{6.16}$$

and set

$$u = u^{(0)} + (u - u^{(0)}) = u^{(0)} + \sum_{i=1}^{N} u^{(i)},$$

where $u^{(i)} := \{u_k^{(i)}\}_{k=1}^N$ with $u_k^{(i)} := u_i - \bar{u}_i$ for $k = i$ and $u_k^{(i)} = 0$ for $k \neq i$. Of course, $u^{(i)} \in V^{(i)}(\Omega)$ and $u = \sum_{i=0}^N u^{(i)}$.

We now check (6.15) for $d_h(.,.)$. For $i = 0$, see (6.7), we have

$$b_0(u^{(0)}, u^{(0)}) = \sum_{i=1}^{N} \sum_{E_{ij} \subset \partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot)(\bar{u}_j - \bar{u}_i), \bar{u}_j - \bar{u}_i)_{L^2(E_{ij})}. \tag{6.17}$$

Note that

$$\begin{aligned}(\varrho_{ij}(\cdot)(\bar{u}_j - \bar{u}_i), \bar{u}_j - \bar{u}_i)_{L^2(E_{ij})} \leqslant C \Big\{ & \parallel \varrho_{ij}^{1/2}(\cdot)(\bar{u}_j - u_j) \parallel_{L^2(E_{ji})}^2 \\ & + \parallel \varrho_{ij}^{1/2}(\cdot)(\bar{u}_i - u_i) \parallel_{L^2(E_{ij})}^2 \\ & + \parallel \varrho_{ij}^{1/2}(\cdot)(u_i - u_j) \parallel_{L^2(E_{ij})}^2 \Big\},\end{aligned} \tag{6.18}$$

where $E_{ij} = E_{ji}, E_{ij} \subset \partial\Omega_i, E_{ji} \subset \partial\Omega_j$. By Friedrich's inequality we have

$$\begin{aligned}\frac{1}{h_{ij}} \parallel \varrho_{ij}^{1/2}(\cdot)(\bar{u}_i - u_i) \parallel_{L^2(E_{ij})}^2 &\leqslant C \frac{\overline{\alpha}_i}{h_{ij}} \parallel u_i - \bar{u}_i \parallel_{L^2(\partial\Omega_i)}^2 \\ &\leqslant C \frac{\overline{\alpha}_i H_i^2}{h_{ij}} |u_i|_{H^1(\partial\Omega_i)}^2 \\ &\leqslant C \frac{\overline{\alpha}_i}{\underline{\alpha}_i} \frac{H_i^2}{h_i h_{ij}} \parallel \varrho_i^{1/2} \nabla u_i \parallel_{L^2(\Omega_i^h)}^2 \\ &\leqslant C \frac{\overline{\alpha}_i}{\underline{\alpha}_i} (\frac{H_i^2}{h_i h_{ij}}) (\varrho_i(\cdot)\nabla u_i, \nabla u_i)_{L^2(\Omega_i)},\end{aligned} \tag{6.19}$$

since $\varrho_{ij}(x) \leqslant 2\varrho_i(x) \leqslant 2\overline{\alpha}_i$ on $\partial\Omega_i$. In the same way we show that

$$\frac{1}{h_{ij}} \parallel \varrho_{ij}^{1/2}(\cdot)(\bar{u}_j - u_j) \parallel_{L^2(E_{ij})}^2 \leqslant C \frac{\overline{\alpha}_j}{\underline{\alpha}_j} (\frac{H_j^2}{h_j h_{ji}}) (\varrho_j(\cdot)\nabla u_j, \nabla u_j))_{L^2(\Omega_j)}. \tag{6.20}$$

Substituting (6.19), (6.20) into (6.18) and the resulting inequality into (6.17), we obtain

$$b_0(u^{(0)}, u^{(0)}) \leqslant C\beta_2 d_h(u,u) \leqslant C\beta_2 \hat{a}_h(u,u). \tag{6.21}$$

We have by (6.5) that

$$\sum_{i=1}^{N} b_i(u^{(i)}, u^{(i)}) = \sum_{i=1}^{N} (\varrho_i(\cdot)\nabla u_i, \nabla u_i)_{L^2(\Omega_i)}^2 + \sum_{i=1}^{N} \sum_{E_{ij} \subset \partial\Omega_i} \frac{\delta}{l_{ij}} \frac{1}{h_{ij}} (\varrho_{ij}(\cdot)(u_i - \bar{u}_i), (u_i - \bar{u}_i)_{L^2(E_{ij})}. \tag{6.22}$$

Using (6.19) and Lemma 5.1, we obtain

$$\sum_{i=1}^{N} b_i(u^{(i)}, u^{(i)}) \leqslant C\beta_2 d_h(u,u) \leqslant C\beta_2 \hat{a}_h(u,u). \tag{6.23}$$

Adding estimates (6.21) and (6.23), we get (6.15). This completes the proof of Theorem 6.1.

Remark 6.1. Estimate (6.12) can be improved when $\overline{\alpha}_i$ and $\underline{\alpha}_i$ are of the same order and $\underline{\alpha}_i \leqslant \varrho_i(x)$ on $\overline{\Omega}_i \backslash \Omega_i^h$. In this case, $\beta_2 = \max_i \max_{j \in \mathcal{I}_i}(H_i/h_{ij})$.

Remark 6.2. The layer Ω_i^h can be replaced by Ω_i^δ, the layer around $\partial\Omega_i$ with width δ_i. In this case, $\beta_2 = \max_i \max_{j \in \mathcal{I}_i}(\frac{\overline{\alpha}_i}{\underline{\alpha}_i}\frac{H_i^2}{h_{ij}\delta_i})$ where $\overline{\alpha}_i$ and $\underline{\alpha}_i$ here are defined on Ω_i^δ (see (2.3)).

7. Implementation

To find the solution of (3.1) for the first discretization and (5.2) for the second discretization, we need to solve Eqs. (4.9) and (6.10), respectively. The operators T_A and T_{DG} are symmetric positive definite and relatively well conditioned (see Theorem 4.1 and Theorem 6.1). To solve these equations a conjugate gradient method is used. We next discuss the implementation of the method for Eq. (6.10) (for (4.9) it is similar). For simplicity of the presentation, we discuss only the Richardson method.

Problem (6.10) is solved by the method

$$u^{n+1} = u^n - \tau(T_{DG}u^n - g_h) = u^n - \tau T_{DG}(u^n - u_h^*),$$

where the relaxation parameter τ can be chosen using the estimates of Theorem 6.1.

To compute

$$r^n := T_{DG}(u^n - u_h^*) = \sum_{i=0}^{n} T_i^{(DG)}(u^n - u_h^*),$$

we need to find $r_i^n := T_i^{(DG)}(u^n - u_h^*)$ by solving the following equations (see (6.8)):

$$b_i(T_i^{(DG)}r_i^n, v) = \hat{a}_h(r_i^n, v) = \hat{a}_h(u^n, v) - f(v), \quad v \in V^{(i)}(\Omega),$$

for $i = 0, \cdots, N$. Note that these problems are independent of each other, therefore, they can be solved in parallel. The problems for $i = 1, \cdots, N$ are local and defined on $\overline{\Omega}_i$ and reduce to discrete problems with continuous FEM and piecewise linear functions. The problem for $i = 0$ has local and global component, where the local problem involves a diagonal preconditioning while the global problem has the number of unknowns equal to the number of subregions Ω_i and reduces to a system with a mass matrix.

The above implementation shows that the proposed algorithm is very well suited for parallel computations.

Acknowledgments. Part of this work was supported by the Polish Sciences Foundation under grant NN201006933.

References

1. D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., **39** (2001), pp. 1749–1779.

2. P. E. Bjorstad, M. Dryja, and E. Vainikko, *Parallel implementation of a Schwarz domain decomposition algorithms*, in *Applied Parallel Computing in Industrial Problems and Optimization* (J. Wasniewski, J. Dongara, K. Madsen, and D. Olsem, eds.) Lecture Notes in Computer Science, Springer, **1184** (1996), pp.141–157.

3. P. G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, North-Holland, Amsterdam, 1978.

4. M. Dryja, *On discontinuous Galerkin methods for elliptic problems with discontinuous coefficients*, Comput. Methods App. Math., **3** (2003), pp. 76–85.

5. M. Dryja, J. Galvis, and M. Sarkis, *BDDC methods for discontinuous Galerkin discretization of elliptic problems,* J. Complexity, **23** (2007), pp. 715–739.

6. M. Dryja and M. Sarkis, *Technical tools for boundary layers and application on heterogeneous coefficients*, 2009, submitted.

7. X. Feng and O. A. Karakashion, *Two-level additive Schwarz methods for discontinuous approximation of second order elliptic problems*, SIAM J. Numer. Anal., **39** (2001), pp. 1343–1365.

8. I. G. Graham, P. O. Lechner, and R. Scheichl, *Domain decomposition for multiscale PDEs*, Numer. Math., **106** (2007), pp. 589–626.

9. C. Pechstein and R. Scheichl, *Analysis of FETI methods for multiscale PDEs*, Numer. Math., **111** (2008), pp. 293–333.

10. B. Riviere, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations - Theory and Implementation*, Frontiers in Applied Mathematics, SIAM, 2008.

11. A. Toselli and O. Widlund, *Domain Decomposition Methods - Algorithms and Theory,* v. 34, Springer Series in Computational Mathematics, Springer-Verlag, 2005.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 177–194

A REGULARIZING PARAMETER FOR SOME FREDHOLM INTEGRAL EQUATIONS

L. FERMO[1]

Abstract — The regularizing parameter appearing in some Fredholm integral equations of the second kind is discussed. Theoretical estimates and the results of numerical tests confirming the theoretical expectations are given.

2000 Mathematics Subject Classification: 65R20; 45E05.

Keywords: Fredholm Integral Equations, Nyström Method.

1. Introduction

In [5], the author introduced a particular procedure to regularize the following Fredholm integral equation of the second kind:

$$f(y) - \frac{\mu}{G_2(y)} \int_0^\infty k(x,y) f(x) w_\alpha(x) dx = \frac{g(y)}{G_1(y)}, \tag{1.1}$$

where $w_\alpha(x) = x^\alpha e^{-x^\beta}$, $\alpha > -1$, $\beta > \frac{1}{2}$ is a generalized Laguerre weight, f is the unknown, $\mu \in \mathbb{R}$, g and k are given smooth functions, and G_1 and G_2 are functions with zeros at the origin of the type y^λ with $0 < \lambda < 1$. The suggested approach consists in "moving" the singularities into the kernel and then regularizing the equation by applying a smoothing transformation depending on the parameter $q \in \mathbb{N}$. Hence, the Nyström method is used to approximate the solution of the equation in a suitable Banach weighted space C_v equipped with a uniform norm.

In this paper, we discuss the choice of the parameter q. Indeed, the approximate solution F_m^* tends to the exact solution F^* with an error of the type

$$\|F^* - F_m^*\|_{C_v} = \mathcal{O}\left(\frac{1}{m}\right)^\sigma,$$

where σ depends on q and increases with increasing q. Consequently, it would appear natural to take q very large to have a good order of convergence. But when the parameter q increases, the speed of convergence slows down compromising the numerical results. Then, the aim of this paper is to propose a suitable choice of the parameter q in order to approximate the solution of the considered equation with a satisfactory theoretical order of convergence and with positive numerical results.

The paper is organized as follows. In Section 2, the regularizing procedure proposed in [5] is described. Section 3 presents the main results including some numerical tests. Section 4 gives proofs to conclude the paper.

[1] *Department of Mathematics and Computer Science, University of Basilicata, v.le dell'Ateneo Lucano, 10 85100 Potenza, Italy* E-mail: luisa.fermo@unibas.it

2. Preliminaries: a regularizing procedure

Let us consider Eq. (1.1) in the weighted space C_u, $u(x) = (1+x)^\rho x^\gamma e^{-x^\beta/2}$, $x, \rho, \gamma \geqslant 0$, defined as

$$C_u = \left\{ f \in C((0,\infty)) : \lim_{\substack{x\to\infty \\ x\to 0}} (fu)(x) = 0 \right\}, \tag{2.1}$$

where $C(J)$ denotes the collection of all continuous functions on $J \subseteq [0,\infty)$. If $\gamma = 0$, then the space C_u consists of all continuous functions on $[0,\infty)$ such that $\lim\limits_{x\to\infty}(fu)(x) = 0$. This space equipped with the following norm

$$\|f\|_{C_u} = \|fu\|_\infty = \sup_{x\geqslant 0} |(fu)(x)|$$

is a Banach space.

In order to approximate the solution of Eq. (1.1) in C_u (if it exists), we could apply the Nyström method or the projection method based on orthogonal polynomials with respect to the weight w_α appearing in the integral (see, for instance, [12]). Nevertheless, it is possible to see (see, for instance, [5], [7], [6]) that, in virtue of the low smoothness of the given functions, these methods lead to very poor numerical results.

Hence the necessity arises to introduce a regularizing procedure that would allow us to improve the smoothness properties of the given functions in order to approximate the solution of (1.1) with a satisfactory order of convergence. In [5], an alternative numerical approach was proposed in this direction. The suggested procedure consists mainly of three steps which we now summarize.

The aim of the first step is to reduce the given equation to a *regularized equation.* To this end we consider (1.1) and for the sake of simplicity, but without loss of generality, we assume

$$G_1(y) = y^\delta, \quad G_2(y) = y^\epsilon, \quad 0 < \epsilon < \delta < 1.$$

We multiply both sides of (1.1) by y^δ and setting $\lambda = \delta - \epsilon$ we get

$$(y^\delta f)(y) - \mu y^\lambda \int_0^\infty k(x,y) x^\delta f(x) x^{\alpha-\delta} e^{-x^\beta} dx = g(y), \quad \delta < \alpha + 1. \tag{2.2}$$

Now, in order to improve the smoothness of the kernel, we introduce the following one-to-one map $\gamma_q : [0,\infty) \to [0,\infty)$ defined as

$$\gamma_q(t) = t^{q/\lambda}, \quad 1 \leqslant q \in \mathbb{N} \tag{2.3}$$

and we change the variables $x = \gamma_q(t)$ and $y = \gamma_q(s)$ in (2.2).

In this way we obtain

$$F(s) - \mu \int_0^\infty h(t,s)\ F(t)\ w_\eta(t) dt = G(s), \tag{2.4}$$

where $F(s) = f(\gamma_q(s)) s^{\frac{q\delta}{\lambda}}$ is the new unknown,

$$G(s) = g(\gamma_q(s)) \quad \text{and} \quad h(t,s) = \frac{q}{\lambda}\, k(\gamma_q(t), \gamma_q(s))\ t^{[\eta]}\ s^q, \tag{2.5}$$

are the new given functions, and $w_\eta(t) = t^{\eta-[\eta]}e^{-t^{q\beta/\lambda}}$ is the new Laguerre weight with $\eta = \frac{q}{\lambda}(1+\alpha-\delta)-1$, $\delta < \alpha+1$, $\beta > \frac{1}{2}$ and $[\eta]$ is its integer part.

We immediately note that the new equation, which we will call the *regularized equation*, has smooth given functions.

Now we have to fix the space in which we will study the *regularized equation* to assure its being unisolvent. This is the second step.

To this end, we introduce the weighted space C_v, with

$$v(s) = u(\gamma_q(s))s^{-\frac{q\delta}{\lambda}} = (1+s^{\frac{q}{\lambda}})^\rho s^{\frac{q}{\lambda}(\gamma-\delta)}e^{-\frac{s^{\frac{q\beta}{\lambda}}}{2}} \tag{2.6}$$

and study the smoothness properties of the new functions and the characteristics of the integral operator associated with the regularized equation

$$(\mathcal{K}F)(s) = \mu\int_0^\infty h(t,s)\ F(t)\ w_\eta(t)dt. \tag{2.7}$$

We denote by k_x (respectively by k_y) the function $k(x,y)$ as a function of the only variable y (respectively x).

Moreover, we define the Zygmund type space

$$Z_{s,r}(v) = \left\{ f \in C_v : \sup_{\tau>0}\frac{\Omega_\varphi^r(f,\tau)_v}{\tau^s} < \infty,\ r > s > 0 \right\}, \tag{2.8}$$

equipped with the norm

$$\|f\|_{Z_{s,r}(v)} = \|fv\|_\infty + \sup_{\tau>0}\frac{\Omega_\varphi^r(f,\tau)_v}{\tau^s},$$

where [17]

$$\Omega_\varphi^r(f,\tau)_v = \sup_{0<h\leqslant\tau}\|(\Delta_{h\varphi}^r f)v\|_{I_{rh}}$$

denotes the main part of modulus of smoothness with $r \geqslant 1$, $\varphi(x) = \sqrt{x}$, $\|\cdot\|_{I_{rh}}$ is the uniform norm on the interval $I_{rh} = [8r^2h^2, \mathcal{C}h^*]$, $h^* = h^{-2/(2\beta-1)}$, and $\mathcal{C}$ is a fixed constant and

$$\Delta_{h\varphi}^r f(x) = \sum_{i=0}^r (-1)^i \begin{pmatrix} r \\ i \end{pmatrix} f\left(x + \left(\frac{r}{2}-i\right)h\varphi(x)\right).$$

For the sake of brevity, we will set $Z_{s,r}(v) := Z_s(v)$.

The following two propositions hold true.

Proposition 2.1. *Let $u(x) = (1+x)^\rho x^\gamma e^{-x^\beta/2}$ and v as in (2.6) with $\beta > \frac{1}{2}$, $\rho > 0$ and $\gamma > \delta$. If the known functions of the original equation are such that*

$$g \in Z_r(u), \tag{2.9}$$

$$\sup_{x\geqslant 0} u(x)\|k_x\|_{Z_r(u)} < \infty, \tag{2.10}$$

$$\sup_{y\geqslant 0} u(y)\|k_y\|_{Z_r(u)} < \infty, \tag{2.11}$$

with $r > 2\delta$, then the known functions of the regularized equation are such that

$$G \in Z_{\sigma_1}(v), \tag{2.12}$$

$$\sup_{t\geqslant 0} v(t)\|h_t\|_{Z_{\sigma_2}(v)} < \mathcal{C}\ \mathcal{D}, \tag{2.13}$$

$$\sup_{s\geqslant 0} v(s)\|h_s\|_{Z_{\sigma_3}(v)} < \mathcal{C}\ \mathcal{A}\ , \tag{2.14}$$

where $\sigma_1 = \frac{q}{\lambda}(r-2\delta)$, $\sigma_2 = \frac{q}{\lambda}(r-2\delta)+2q$, $\sigma_3 = \frac{q}{\lambda}(r-2\delta)+2[\eta]$, q and λ are the parameters appearing in the transformation γ_q defined in (2.3)*, $\mathcal{C}$ is a positive constant independent of the given functions and of q and λ, while $\mathcal{A}$ and $\mathcal{D}$ are constants depending on q and λ.*

Remark 2.1. We note that the choice of q as a natural number is closely related to the smoothness properties of the given functions (on which the order of convergence depends, as we will see in Section 3). Indeed, if q is not necessarily a natural number, the kernel h_t has a worse smoothness because of the factor s^q. In fact, in this case we have $h_t \in Z_{2\frac{q}{\lambda}(\gamma-\delta)+2q}(v)$. As an example, consider $k(x,y) = (x^{2/3}+y^{7/2})$, $\gamma_q(t) = t^{\frac{3}{2}q}$, $\lambda = \delta = 2/3$, $\gamma = 0.7$. Then, if $q \in \mathbb{N}$, we have $h_t(s) \in Z_{12.6q}(v)$, otherwise $h_t(s) \in Z_{2.1q}(v)$.

Proposition A [5]. *Let $u(s) = (1+s)^\rho s^\gamma e^{-\frac{s^\beta}{2}}$, $\beta > \frac{1}{2}$, v as in* (2.6) *with ρ and γ such that*

$$\rho > \frac{1}{2}, \qquad \max\left\{\delta, \frac{\delta+\alpha}{2}\right\} < \gamma < \frac{\alpha+1}{2}. \tag{2.15}$$

Then, if the kernel k_x satisfies (2.10)*, the operator $\mathcal{K} : C_v \to C_v$ defined in* (2.7) *is compact and for* (2.4) *the Fredholm Alternative Theorem holds true in C_v.*

We remark that if $\alpha < \delta$, find the value of γ, it is essential that $0 < \delta < \frac{\alpha+1}{2}$.

Now by means of the previous propositions it is possible to determine the conditions under which the *regularized equation* (2.4) is unisovent in C_v.

Proposition B [5]. *Let u and v be as in Proposition A and let* (2.10) *be satisfied. Then the original equation* (1.1) *has a unique solution $f^* \in C_u$ for each given right-hand side in C_u if and only if the regularized equation* (2.4) *has a unique solution $F^* \in C_v$ for each given right-hand side $G \in C_v$. Moreover the following relation holds true:*

$$(f^*u)(t) = (F^*v)(\gamma_q^{-1}(t)) \tag{2.16}$$

for each point $t \in [0,\infty)$.

The last step consists in applying the Nyström method to the *regularized equation* (2.4) in order to approximate its solution.

To this end, we first approximate the integral $\mathcal{K}F$ by using the following truncated Gaussian rule (see, e.g., [14],[15],[4]):

$$\int_0^\infty h(t,s)F(t)w_\eta(t)dt = \sum_{k=1}^{j} \lambda_k(w_\eta)h(x_k,s)F(x_k) + e_m^*(h_yF), \tag{2.17}$$

where $x_k = x_{m,k}(w_\eta)$, $k = 1, ..., j$, are the zeros of the polynomial $p_m(w_\eta)$ which is orthonormal with respect to the weight w_η, λ_k, $k = 1, ..., j$ are the Christoffel numbers corresponding to w_η,

$$x_j = \min_{1\leqslant k\leqslant m} \{x_k : \ x_k \geqslant \theta a_m\}, \quad 0 < \theta < 1, \tag{2.18}$$

$$a_m = 4\frac{\Gamma(\frac{q}{\lambda}\beta)^{\frac{2\lambda}{q\beta}}}{\Gamma(2\frac{q}{\lambda}\beta)^{\frac{\lambda}{q\beta}}} m^{\frac{\lambda}{q\beta}}, \tag{2.19}$$

denotes the Mhaskar-Rahmanov-Saff number (see, e.g., [11]) and $e^*_m(h_y F)$ is the remainder term.

Thus, setting

$$(\mathcal{K}_m F)(s) = \mu \sum_{k=1}^{j} \lambda_k(w_\eta) h(x_k, s) F(x_k),$$

we go to consider the operator equation

$$(I - \mathcal{K}_m)F_m = G,$$

where F_m is unknown.

Then, multiplying this equation by the weight v chosen as in Proposition A and collocating on the zeros x_i, $i = 1, ..., j$, we obtain the following linear system:

$$\sum_{k=1}^{j} \left[\delta_{i,k} - \mu\lambda_k(w_\eta)\frac{v(x_i)}{v(x_k)} h(x_k, x_i) \right] b_k = (Gv)(x_i), \quad i = 1, \ldots, j, \tag{2.20}$$

where $\delta_{i,k}$ is a Kronecker symbol and $b_k = F_m(x_k)v(x_k)$, $k = 1, \ldots, j$ are the unknowns. Now, if the above system has a unique solution $[b^*_1, \ldots, b^*_j]^T$, then we can construct the following weighted Nyström interpolant:

$$F^*_m(s)v(s) = \mu \sum_{k=1}^{j} \lambda_k(w_\eta) \frac{v(s)}{v(x_k)} h(x_k, s) b^*_k + G(s)v(s). \tag{2.21}$$

Hence, in order to obtain an approximate solution of (2.4), we have to solve a linear system of j equations in j unknowns rather than a system of m equations in m unknowns and this implies a significant economy in computations. Moreover, we remark that system (2.20) can easily be constructed because it only requires the computation of the zeros x_k, $k = 1, ..., j$ and of the Christoffel Numbers $\lambda_k(w_\eta)$, $k = 1, ..., j$. To this end, one can use, in the Laguerre case, the routine *gaussq* (see [8]) or routines *recur* and *gauss* (see [9] and [10]), and in the general case, the Mathematica Package “OrthogonalPolynomials” (see [3]).

The stability and the convergence of the proposed method is stated in the following theorem proved in [5].

Theorem A [5]. *Assume that Eq.* (1.1) *has a unique solution f^* in C_u and that the hypotheses of Proposition* 2.1 *are satisfied. Then for m sufficiently large, system* (2.20) *is unisolvent and its matrix $\mathbf{B}_j$ is well conditioned holding*

$$cond(\mathbf{B}_j) \leqslant \mathcal{C}, \tag{2.22}$$

where $\mathcal{C}$ does not depend on m and $cond(\mathbf{B}_j) = \|\mathbf{B}_j\|_\infty \|\mathbf{B}_j^{-1}\|_\infty$.

3. Main Results

3.1. Why the choice of q? The error estimate.

The regularizing procedure and the Nyström method summarized in the previous section do not impose any restriction on the parameter q. Indeed, until now we have seen that, for each value of q, the given functions of Eq. (2.4) are smooth , the *regularized equation* is unisolvent in the space C_v, and system (2.20) has a unique solution and is well conditioned. Nevertheless, we need an optimal choice of q. In order to understand the reason of this necessity, let us estimate the error.

To this end, we denote by F^* the unique solution of (2.4) in C_v and by F^*_m the Nyström interpolant defined in (2.21).

By the well-known argument (see, e.g., [1])

$$\begin{aligned}\|[F^*-F^*_m]v\|_\infty &\sim \|[\mathcal{K}F^*-\mathcal{K}_mF^*]v\|_\infty\\ &=\sup_{s\geqslant 0} v(s)\left|\int\limits_0^\infty h(x,y)F^*(x)w_\eta(x)dx-\sum_{k=1}^{j}\lambda_k(w_\eta)h(x_k,s)F^*(x_k)\right|\\ &=\sup_{s\geqslant 0} v(s)|e^*_M(h_sF^*)|, \qquad (3.1)\end{aligned}$$

where $e^*_M(h_sF^*)$ is the remainder term of the Gaussian rule (2.17).

Now, since in virtue of the assumptions about the parameters of the weight v it results in $\int\limits_0^\infty \frac{w_\eta(t)}{v^2(t)}dt<\infty$, we have [12]

$$|e^*_m(h_yF)|\leqslant \mathcal{C}[E_M(h_yF)_{v^2}+e^{-Am}\|h_yFv^2\|_\infty], \qquad (3.2)$$

where the constants $\mathcal{C}$ and A are independent of m and F, $M=[(\frac{\theta}{1+\theta})^\beta m]$ and $E_n(f)_v=\inf\limits_{P_n\in\mathbb{P}_n}\|(f-P_n)v\|_\infty$ denotes the error of the best approximation of $f\in C_v$ by polynomials of degree n at most ($P_n\in\mathbb{P}_n$).

Hence, choosing $M=am$, $0<a<1$ and taking into account that for all $f,g\in C_v$, we get

$$E_m(fg)_{v^2}\leqslant \mathcal{C}[\ \|fv\|E_m(g)_v+2\|gv\|_\infty E_m(f)_v\ ], \qquad (3.3)$$

by (3.1) we have

$$\|[F^*-F^*_m]v\|_\infty\leqslant \mathcal{C}\left[\|F^*v\|_\infty\sup_{s\geqslant 0}v(s)E_{[\frac{M}{2}]}(h_s)_v+\sup_{s\geqslant 0}v(s)\|h_sv\|_\infty E_{[\frac{M}{2}]}(F^*)_v\right].$$

By Proposition 2.1 we deduce that $h,g\in Z_\sigma(v)$, with $\sigma=\frac{q}{\lambda}(r-2\delta)$ and then $F^*\in Z_\sigma(v)$, too. Moreover, since $\forall f\in Z_s(v)$ (see, e.g., [17])

$$E_m(f)_v\leqslant \mathcal{C}\left(\frac{\sqrt{a_m}}{m}\right)^s\|f\|_{Z_s(v)},\quad m>s\quad \mathcal{C}\neq\mathcal{C}(m,f), \qquad (3.4)$$

we have

$$\|[F^*-F^*_m]v\|_\infty\leqslant \mathcal{C}\left(\frac{\sqrt{a_m}}{m}\right)^\sigma\|F^*\|_{Z_{\sigma(v)}}\sup_{s\geqslant 0}v(s)\|h_s\|_{Z_{\sigma 3}(v)}. \qquad (3.5)$$

We have proved the following result.

Theorem 3.1. *Assume that the assumptions of Theorem A are satisfied. Then if F^* denotes the unique solution of Eq.* (2.4) *and F^*_m is the Nyström interpolant defined in* (2.21)*, then*

$$\|[F^* - F^*_m]v\|_\infty \leqslant \mathcal{C}\left(\frac{\sqrt{a_m}}{m}\right)^\sigma \|F^*\|_{Z_{\sigma(v)}} \sup_{s\geqslant 0} v(s)\|h_s\|_{Z_{\sigma_3}(v)}, \tag{3.6}$$

where $\mathcal{C} \neq \mathcal{C}(m, F^*)$ and $\sigma = \frac{q}{\lambda}(r - 2\delta)$.

Hence, the theoretical order of convergence depends on the smoothness properties of the given functions. Consequently, we emphasize again the importance of choosing q as a natural number. And if it is not natural, then, by Remark 2.1, we obtain that the theoretical order of convergence is worse $\mathcal{O}\left(\left(\frac{\sqrt{a_m}}{m}\right)^{\frac{q}{\lambda}(\gamma-\delta)+2q}\right)$.

From estimate (3.6) it follows that for any constant $\mathcal{C}$ independent of m the error tends to zero as $\left(\frac{\sqrt{a_m}}{m}\right)^\sigma$, since theoretically we can choose m sufficiently large. Moreover, the rate of convergence increases with increasing q. Consequently, we tempt to take q very large to have a good order of convergence. But now we linger over the Zygmund norms appearing on the right hand side of (3.6). By Proposition 2.1 it follows that

$$\sup_{s\geqslant 0} v(s)\|h_s\|_{Z_{\sigma_3}(v)} \leqslant \mathcal{C}\ \mathcal{A},$$

where $\mathcal{A}$ is a constant depending on q. Moreover, using the same argument we can see that $\|F^*\|_{Z_{\sigma(v)}}$ also has the same behavior.

The error estimate is of the following type:

$$\|[F^* - F^*_m]v\|_\infty = \mathcal{C}\mathcal{A}^2\mathcal{O}\left(\left(\frac{\sqrt{a_m}}{m}\right)^\sigma\right), \tag{3.7}$$

i.e., a constant $\mathcal{A}$ depending on a parameter (which can be changed) appears. Consequently, it is necessary to analyze the behavior of this constant when the parameter varies. Indeed, if it becomes large as q does, the numerical convergence can be compromised even if the theoretical one is ensured.

In the following subsection we will make an evaluation of this constant and study its behavior. Here we only observe that in the approximation theory an error estimate in which a parameter-dependent constant frequently appears. For instance, in $[-1, 1]$, if we consider the function $f(x) = \log(1 + x)$, it is possible to prove that there exists a polynomial P (see, e.g., [16]) such that

$$\|[f - P]v^{\gamma,\delta}\|_\infty \leqslant \mathcal{C}(r-1)!\frac{\log m}{m^r}, \quad v^{\gamma,\delta}(x) = (1-x)^\gamma(1+x)^\delta.$$

Then also in this case a constant $\mathcal{A} = (r-1)!$ appears. Moreover, here the numerical problem we have is evident: if r increases, then the order of convergence becomes large but the speed of convergence slows down because the constant $\mathcal{A}$ increases. This will be our problem.

3.2. Constant $\mathcal{A}$ and the crucial problem of choosing the regularizing parameter

In [5], to give an idea of the constant $\mathcal{A}$, the following estimate was proved in the case where the parameter η appearing in (2.5) is equal to zero.

Proposition C [5]. *Let $q \geqslant 1$ and $0 < \lambda < 1$. Then*

$$\mathcal{A} \leqslant \left(\left[\frac{q}{\lambda}\right]+1\right)^{[\frac{q}{\lambda}]} \mathcal{W}\left(\left[\frac{q}{\lambda}\right]\right),$$

where $\mathcal{W}\left(\left[\frac{q}{\lambda}\right]\right)$ denotes the $\left[\frac{q}{\lambda}\right]$th Bell number.

Now we will make an evaluation for $\mathcal{A}$.

By the proof of Proposition 2.1, setting $\ell = \min\{[\frac{q}{\lambda}], [r]\}$ it follows that

$$\mathcal{A} = \begin{cases} \displaystyle\sum_{i=0}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} \\ \qquad \displaystyle\sum_{m=0}^{i} \mathcal{B}_{i,m}\left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right), ..., \frac{q}{\lambda} \cdot ... \cdot \left(\frac{q}{\lambda}-i+m\right)\right), \quad \ell \leqslant [\eta]; \\ \displaystyle\sum_{i=\ell-[\eta]}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} \\ \qquad \displaystyle\sum_{m=0}^{i} \mathcal{B}_{i,m}\left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right), ..., \frac{q}{\lambda} \cdot ... \cdot \left(\frac{q}{\lambda}-i+m\right)\right), \quad \ell > [\eta] \end{cases} \tag{3.8}$$

where $\mathcal{B}_{i,m}$ denotes the partial Bell polynomials defined in (4.1) with $\mathcal{B}_{0,m} = 1$ for all $m = 0, \cdots, i$ and $\mathcal{B}_{i,0} = 0$ for all $i = 1, \cdots, \ell$.

Now, from the theory of Bell's polynomials it is known that

$$\sum_{k=1}^{n} B_{n,k}(x_1, x_2, ..., x_{n-k+1}) = B_n(x_1, x_2, ..., x_n),$$

where $B_n(x_1, x_2, ..., x_n)$ are the so-called complete Bell polynomials which satisfy the following property:

$$\begin{aligned} & B_n(x_1, x_2, ..., x_n) \\ & := det \begin{pmatrix} x_1 & \begin{pmatrix} n-1 \\ 1 \end{pmatrix} x_2 & \begin{pmatrix} n-1 \\ 2 \end{pmatrix} x_3 & \dots & \begin{pmatrix} n-1 \\ n-2 \end{pmatrix} x_{n-1} & x_n \\ -1 & x_1 & \begin{pmatrix} n-2 \\ 1 \end{pmatrix} x_2 & \dots & \begin{pmatrix} n-2 \\ n-3 \end{pmatrix} x_{n-2} & x_{n-1} \\ 0 & -1 & x_1 & \dots & \begin{pmatrix} n-3 \\ n-4 \end{pmatrix} x_{n-3} & x_{n-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & x_1 & x_2 \\ 0 & 0 & 0 & \dots & -1 & x_1 \end{pmatrix}. \end{aligned} \tag{3.9}$$

Then, in virtue of this relation, in order to compute the constant $\mathcal{A}$ we have only to compute special sums of determinants of a particular matrix. Indeed, since $\mathcal{B}_{0,m} = 1$ for all $m = 0, \cdots, i$ and $\mathcal{B}_{i,0} = 0$ for all $i = 1, \cdots, \ell$, by (3.8) and (3.9), the constant $\mathcal{A}$ can be rewritten as

$$\mathcal{A} = \begin{cases} \dfrac{[\eta]!}{([\eta]-\ell)!} + \displaystyle\sum_{i=1}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} det(\mathbf{A}_i), & \ell \leqslant [\eta]; \\ \displaystyle\sum_{i=\ell-[\eta]}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} det(\mathbf{A}_i), & \ell > [\eta] \end{cases} \tag{3.10}$$

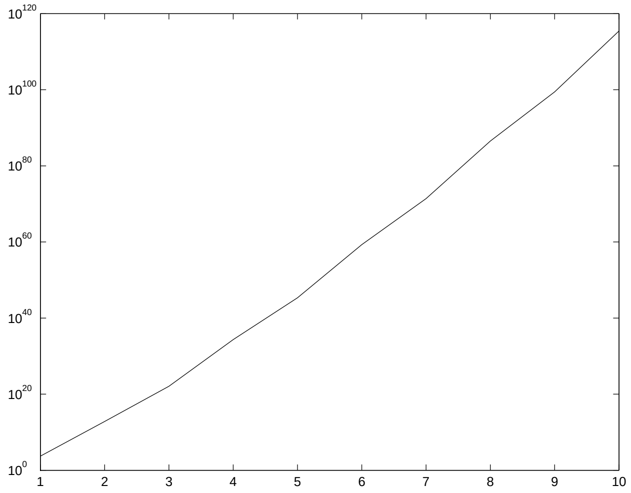


Fig. 3.1. $\mathcal{A}$

where $\det(\mathbf{A}_i)$ denotes the determinant of the matrix defined in (3.9) with $n = i$ and with

$$(x_1, x_2, ..., x_n) = \left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda} - 1\right), ..., \frac{q}{\lambda} \cdot ... \cdot \left(\frac{q}{\lambda} - i + 1\right)\right).$$

Note that in the simple case where $\eta = 0$ we have in $\mathcal{A} = det(\mathbf{A}_\ell)$.

We also underline that in order to compute the determinant of the matrix $\mathbf{A}_i$, one can use the following formula:

$$det(\mathbf{A}_i) = \sum_{k=0}^{i-1} \left(\begin{array}{c} i-1 \\ k \end{array} \right) x_{k+1}\, det(\mathbf{A}_{i-k-1}), \quad i \geqslant 2,$$

with $det(\mathbf{A}_0) = 1$ and $det(\mathbf{A}_1) = x_1$.

Now by (3.10) the behavior of the constant is evident: as q increases, it becomes very large. Moreover, we note that since the constant depends on the ratio $\frac{q}{\lambda}$, when λ is close to zero, this constant becomes large even when q is small (see Fig. 3.1). On the contrary, if λ is close to one, it becomes large when q is large (see Table 3.4). Figure 3.1 shows the trend of the constant when q changes in the case where $\ell = [\frac{q}{\lambda}]$, $\lambda = 2/9$, $\delta = 2/9$ and $\alpha = -1/3$.

The problem announced in the previous subsection is confirmed: when q becomes large the numerical convergence is compromised even if the theoretical one is assured. Indeed, by the error estimate (3.7)

$$\|[F^* - F_m^*]v\|_\infty = \mathcal{C}\mathcal{A}^2\mathcal{O}\left(\left(\frac{\sqrt{a_m}}{m}\right)^{\frac{q}{\lambda}(r-2\delta)}\right),$$

we deduce that if q becomes large, then the order of convergence increases but the speed of convergence slows down because of the presence of the constant $\mathcal{A}$. Consequently, we need a very large number of points m to obtain the required convergence. For instance, assume $\lambda = \delta = 2/9$, $r = 2$. According to (3.10), if $q = 8$, then $\mathcal{A} = 3.351200611656362e + 086$. Therefore, to have the approximate solution with, e.g., 7 correct digits, we need a number of points $m > 1899$. But this is not realistic. In fact, in order to construct the Nyström interpolant F_m^* defined in (2.21), we have to solve system (2.20). Thus, we have to compute the zeros x_k and the Christoffel numbers λ_k of a polynomial of degree $m > 1899$. And this requires a computational effort.

Then, for this reason, an optimal choice of the parameter q is necessary. To this end, we suggest to proceed in the following way:

1. Regularize the given equation as shown in Section 2.

2. Compute the order of convergence according to (3.6).

3. Compute the constant $\mathcal{A}$ according to (3.10) for different values of q. Now, as mentioned above, the constant $\mathcal{A}$ becomes very large as q increases. Consequently, after a certain value q_0 of q, the constant $\mathcal{A}^2$ (we need it later to compute the optimal parameter q) cannot be computed numerically. Among the highest values $\mathcal{A}^2 \sim 10^{292}$. After this value it is impossible to know $\mathcal{A}^2$. Because of this, in this phase we fix the range $[1, q_0]$ in which we can choose our optimal parameter q.

4. Fix the correct digits we want to be exact in the approximate solution and then compute the number of points m we need to obtain it. For instance, if we want to have an approximate solution with a correct digits, taking into account (3.7) and (2.19), it has to be
$$m \geqslant \left(\mathcal{A}^2 2^\sigma \left(\frac{\Gamma(\frac{q}{\lambda}\beta)}{\sqrt{\Gamma(2\frac{q}{\lambda}\beta)}} \right)^{\frac{\lambda\sigma}{q\beta}} 10^{a+1} \right)^{\frac{1}{(1-\frac{\lambda}{2q\beta})\sigma}}, \quad 1 \leqslant q \leqslant q_0. \tag{3.11}$$

5. Choose the optimal parameter $q \in [1, q_0]$, that is the natural number which minimize the right-hand side of (3.11).

6. Solve system (2.20) and construct the Nyström interpolant (2.21).

7. Compute the solution of the original equation according to (2.16).

Proceeding in this way, we will approximate the solution of the considered equations with a satisfactory theoretical order of convergence and with positive numerical results.

We note that theoretically the parameter q_0 can be large (it depends on the other parameters involved in the computation of $\mathcal{A}$). Consequently, the optimal parameter q can be large. On the other hand, it is very difficult to give an analytical expression of the minimal point of the right-hand side of (3.11). In any case, if $q \in [1, q_0]$ is large, then we have no numerical problem: system (2.20) is well conditioned for each value of q. However, we underline that in all the examples tested the optimal parameter q has always been small.

In the following subsection we will carry out some numerical tests confirming our theoretical expectations.

3.3. Numerical Tests

In this subsection, we will give the numerical results obtained for some Fredholm integral equations.

To this end, we will follow the procedure suggested in the previous subsection. Thus, first of all, we will regularize the given equation as shown in Section 2. Subsequently, we will choose the optimal parameter q to avoid a compromise of the numerical convergence.

Then with fixed q, we will construct the Nyström interpolant (2.21) of Eq. (2.4) and finally we will compute the approximate solution f_m^* of the original equation according to (2.16).

In each numerical test, we take, as a reference solution, the approximated solution obtained at $m = 256$ and in all the tables we will give $e_m = \max_i |(f_{256}u)(y_i) - (f_m u)(y_i)|$, where $\{y_i\}_{i=1}^{20}$ denotes 20 equispaced points on the interval $(0, \infty)$ and the condition number in the infinity norm of system (2.20). All computations were performed in 16-digit arithmetics.

Example 3.1. Consider the equation

$$f(y) - \frac{1}{5y^{2/3}} \int_0^\infty (x^2 + y^2 + 8)\ f(x)\ x^{4/5} e^{-x^{3/4}} dx = \frac{(y^{3/2} + 2)}{y^{7/9}}.$$

It has a unique solution in the weighted space C_u with $u(x) = (1+x)^{0.6} x^{0.85} e^{-\frac{x^{3/4}}{2}}$. Note that the function $k(x, y) = x^2 + y^2 + 8$ is an analytical function while $g(y) = (y^{3/2} + 2) \in Z_{4.7}(u)$. Now, applying the regularizing procedure shown in Section 2, we obtain

$$F(s) - \frac{9}{5} q s^q \int_0^\infty (t^{18q} + s^{18q} + 8)\ t^{[\eta]}\ F(t)\ w_\eta(t) dt = (s^{\frac{27q}{2}} + 2),$$

with $\eta = \frac{46}{5} q - 1$ and $w_\eta(t) = t^{\eta - [\eta]} e^{-t^{\frac{27q}{4}}}$. The new equation has a unique solution in C_v with v as in (2.6), according to Proposition B. We note that the new kernel is an analytical function while the right-hand side pertains to $Z_{28.30q}(v)$. Consequently, the order of convergence is $\mathcal{O}\left((\frac{\sqrt{a_m}}{m})^\sigma \right)$ with $\sigma = 28.30q$ according to (3.6). Now, we choose the optimal parameter q. Then, first of all, we compute the constant $\mathcal{A}$ according to (3.10) for different values of q. Thus we fix the greatest value of q, namely q_0, for which we can compute numerically $\mathcal{A}^2$. In this case, we have $q_0 = 5$. Now, we would like to know the approximate solution with 6 correct digits and we compute the optimal parameter q, that is the value of $q \in [0, 5]$ which minimize the right hand side of (3.11). The following graph shows the behavior of (3.11) when q changes.

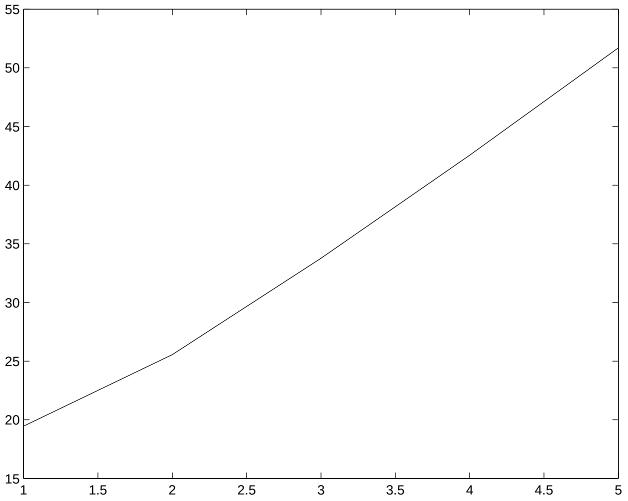


Fig. 3.2. m

Hence we deduce that the optimal parameter is $q = 1$ in accordance with which we have the required convergence with a number of points $m \geqslant 19$. Table 3.3 shows the weighted approximate solution obtained with this optimal parameter ($\theta = 0.9$).

Table 3.1. **q=1**

m	j	e_m	$cond(\mathbf{B}_j)$
16	16	4.31192e-007	24.63689043040089
32	31	1.88633e-007	29.56259271555317
64	60	1.04805e-013	32.56276574794698
128	120	7.10542e-015	34.42444895667656

If the parameter q increases, for instance, $q = 4$, then the numerical results are poor. Indeed, as shown in Table 3.2, in order to have 6 correct digits we have to solve a system of order 63 rather than 16 as done in the case $q = 1$. From the last table we can also see that if the parameter q increases, the condition number in the infinity norm of system (2.20) is still bounded.

Table 3.2. **q=4**

m	j	e_m	$cond(\mathbf{B}_j)$
32	32	4.05913e-006	25.04999125397745
64	63	7.50975e-007	29.87180783730457
128	126	2.27320e-011	32.77334652130129

Example 3.2. We consider the following Fredholm integral equation:

$$f(y) - \frac{1}{2}\int_0^\infty (x^2 + y + 3) f(x) x^{4/3} e^{-x} dx = \frac{y+1}{y^{1/3}} - \frac{3}{2}(13 + y),$$

whose exact solution is $f(y) = \frac{1+y}{y^{1/3}}$.
The considered equation has a unique solution in the weighted space C_u with $u(x) = (1 + x)^{0.7} x^{9/8} e^{-x/2}$. Using the regularizing procedure shown in Section 2, we get

$$F(s) - \frac{3q}{2} s^q \int_0^\infty (t^{6q} + s^{3q} + 3) t^{6q-1} F(t) e^{-t^{3q}} dt = s^{3q} + 1 - \frac{3}{2}(13 + s^{3q}) s^q,$$

which has a unique solution in C_v according to Proposition B. We immediately notice that all given functions are polynomials for each q and the convergence is very fast. Table 3.3 shows the numerical results obtained at $q = 1$ ($\theta = 0.7$). Note that in this case $\mathcal{A} = 132$. If the parameter q increases the given functions are still polynomials and we expect the same numerical results but they are poor. Indeed, since the constant $\mathcal{A}$ increases, the speed of convergence slows down compromising the numerical results. Table 3.4 shows what happens in the case where $q = 8$ ($\theta = 0.96$). Note that in this case $\mathcal{A} = 8.321415742355469e + 050$.

Table 3.3. **q=1**

m	j	e_m	$cond(\mathbf{B}_j)$
8	8	5.68434e-014	21.78681961756113

Table 3.4. **q=8**

m	j	e_m	$cond(\mathbf{B}_j)$
16	16	1.16761e-001	16.75283784307782
32	31	2.57434e-002	25.56426459953316
64	61	6.02116e-005	31.11032195144892
128	122	9.02389e-013	33.82723852547484
256	242	4.26325e-014	37.34898825173112

Example 3.3. Consider the equation

$$f(y) - \frac{1}{7}\int_0^\infty (x^{7/2} + y^{2/3} + 7)\ f(x)\ \sqrt{x}e^{-x}dx = \frac{2}{y^{2/3}e^{(1+y^{4/3})}}.$$

It has a unique solution in the weighted space C_u with $u(x) = (1+x)^{0.6}x^{0.7}e^{-x/2}$. Applying the procedure shown in Section 2, the given equation is equivalent to

$$F(s) - \frac{3q}{14}s^q\int_0^\infty (t^{\frac{21q}{4}} + s^q + 7)t^{[\eta]}\ F(t)\ w_\eta(t)dt = \frac{2}{e^{(1+s^{2q})}}$$

with $\eta = \frac{5}{4}q - 1$ and $w_\eta(t) = t^{\eta-[\eta]}e^{-t^{\frac{3}{2}q}}$. The new equation has a unique solution in C_v with v as in (2.6), in virtue of Proposition B.

Note that the right-hand side and the kernel with respect to the variable s of the new equation are analytical functions while the kernel with respect to the variable t pertains to $Z_{6.45q}(v)$. Consequently, the order of convergence is $\mathcal{O}(\frac{1}{m^{5.37q}})$, according to (3.6) and (2.19).

Now we choose the optimal parameter q to obtain an approximate solution with 7 correct digits. Computing expression (3.10), we can see that we can determine numerically $\mathcal{A}^2$ if $q \in [1, q_0]$ with $q_0 = 32$. Then, taking into account (3.11), we can construct Table 3.5.

Hence we deduce that the optimal parameter is $q = 5$. Table 3.6 shows the obtained numerical results ($\theta = 0.9$).

Example 3.4. Consider the equation

$$f(y) - \frac{1}{12}\int_0^\infty \sin{(xy)}e^{-xy}\ f(x)\ x^{-1/5}e^{-x^{3/2}}dx = \frac{\log{(1+y)}}{\sqrt{y}(y^2+4)}.$$

Table 3.5

q	$\mathcal{A}$	$m \geqslant$
1	1.5	381
2	132	27
3	7012.6875	13
4	13610520	14
5	1.769577199804688e+009	11
6	1.190512570851900e+013	13
$\vdots$	$\vdots$	$\vdots$
25	2.134477512167769e+090	26
$\vdots$	$\vdots$	$\vdots$
32	7.851241025230311e+125	33

Table 3.6. **q=5**

m	j	e_m	$cond(\mathbf{B}_j)$
16	16	3.41507e-006	28.78177615020002
32	31	2.94802e-010	34.21321407988217
64	62	2.02615e-015	37.37669660028023

It is unisovent in the weighted space C_u with $u(x) = (1+x)^{0.8}x^{0.35}e^{-x^{3/2}/2}$. Using the regularizing procedure described in Section 2, we find that it is equivalent to

$$F(s) - \frac{q}{6}s^q \int_0^\infty \sin{(ts)^{2q}}\ e^{-(st)^{2q}}\ F(t)\ t^{\frac{3}{5}q-1}e^{-t^{3q}}dt = \frac{\log{(1+s^{2q})}}{(s^{4q}+4)},$$

which has a unique solution in C_v with v as in (2.6) according to Proposition B.

We immediately notice that all given functions are analytical for each value of q. Consequently, the convergence is very fast as shown in Table 3.7 in which the results were obtained with $q = 1$ and $\theta = 0.7$. Note that in this case the constant $\mathcal{A} = 6$.

Table 3.7. **q=1**

m	j	e_m	$cond(\mathbf{B}_j)$
8	8	8.87958e-007	1.045102472850763
16	14	1.03388e-010	1.046930825888115
32	27	6.92534e-018	1.048257667797889

If the parameter q increases, the order of convergence remains the same because the functions are still analytic but the speed of convergence slows down because the constant increases. Indeed, if, for instance, $q = 7$, we have $\mathcal{A} = 1.257542760359232e+024$. Table 3.8

Table 3.8. **q=7**

m	j	e_m	$cond(\mathbf{B}_j)$
16	15	2.16707e-005	1.044237627930557
32	28	1.00269e-005	1.046535459802606
64	56	3.66678e-008	1.047805484514525
128	110	4.82443e-011	1.048941221806483

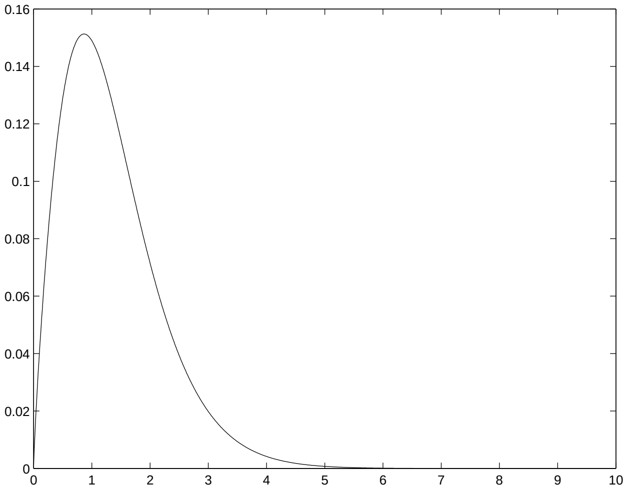


Fig. 3.3. $(f^*_{32}u)(y)$

shows the results obtained at $q = 7$ ($\theta = 0.9$). Note that the condition number of system (2.20) is also small in the case where the parameter q increases.

Figure 3.3 shows the graph of the weighted approximate solution $f^*_{32}u$.

4. Proofs

Proof of Proposition 2.1.
We begin by proving (2.12). Let $\ell = \min\{[\frac{q}{\lambda}], [r]\}$. By the Faá di Bruno Formula we have

$$G^{(\ell)}(s) = \sum_{k=1}^{\ell} g^{(k)}(\gamma_q(s))\ \mathcal{B}_{\ell,k}(\gamma_q^{(1)}(s), \gamma_q^{(2)}(s), ..., \gamma_q^{(\ell-k+1)}(s)),$$

where $\mathcal{B}_{\ell,k}$ denotes the partial Bell polynomials defined as (see, e.g., [2, p. 134])

$$\mathcal{B}_{\ell,k}(x_1, x_2, ..., x_{\ell-k+1}) = \sum \frac{\ell!}{k_1!k_2!...k_{\ell-k+1}!} \left(\frac{x_1}{1!}\right)^{k_1} \left(\frac{x_2}{2!}\right)^{k_2} \cdot ... \cdot \left(\frac{x_{\ell-k+1}}{(\ell-k+1)!}\right)^{k_{\ell-k+1}}, \quad (4.1)$$

where the sum is extended to all positive integers $k_1, k_2, ..., k_{\ell-k+1}$ such that $k = k_1 + k_2 + ... + k_{\ell-k+1}$ and $k_1 + 2k_2 + ... + (\ell - k + 1)k_{\ell-k+1} = \ell$.
Developing $\mathcal{B}_{\ell,k}(\gamma_q^{(1)}(s), \gamma_q^{(2)}(s), ..., \gamma_q^{(\ell-k+1)}(s))$ leads to

$$G^{(\ell)}(s) = \sum_{k=1}^{\ell} g^{(k)}(\gamma_q(s)) s^{\frac{q}{\lambda}k-\ell} \mathcal{B}_{\ell,k}\left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right), ..., \frac{q}{\lambda} \cdot ... \cdot \left(\frac{q}{\lambda} - \ell + k\right)\right).$$

Then denoting by $\varphi(s)=\sqrt{s}$, $u(s)=(1+s)^{\rho}s^{\gamma}e^{-s^{\beta}/2}$ and $v(s)=u(\gamma_q(s))s^{-\frac{q}{\lambda}\delta}$, we deduce

$$\begin{aligned}&|(G^{(\ell)}\varphi^{\ell}v)(s)|\\&\leqslant\sum_{k=1}^{\ell}|(g^{(k)}\varphi^{k}u)(\gamma_q(s))|s^{\frac{q}{\lambda}(\frac{k}{2}-\delta)-\frac{\ell}{2}}\mathcal{B}_{\ell,k}\left(\frac{q}{\lambda},\frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right),...,\frac{q}{\lambda}\cdot...\cdot\left(\frac{q}{\lambda}-\ell+k\right)\right).\end{aligned}\tag{4.2}$$

Now, by the assumption $g\in Z_r(u)$. Therefore, taking into account that

$$\Omega_{\varphi}^{k}(g,t)_u\leqslant\mathcal{C}\sup_{0<h\leqslant t}h^{k}\|g^{(k)}\varphi^{k}u\|_{I_{hk}},\tag{4.3}$$

with $I_{hk}=[8k^2h^2,\mathcal{C}h^{-2}]$ by some computations we have

$$|(g^{(k)}\varphi^{k}u)(\gamma_q(s))|<\mathcal{C}s^{\frac{q}{\lambda}(\frac{r}{2}-\frac{k}{2})}(1+s^{q/\lambda})^{\rho}\ e^{-s^{q/\lambda}/2}M(s),$$

where M is a smooth function.

Thus, by (4.2) taking the supremum on $I_{h\ell}$, we have

$$\|G^{(\ell)}\varphi^{\ell}v\|_{I_{h\ell}}\leqslant\mathcal{C}h^{\frac{q}{\lambda}(r-2\delta)-\ell}\sum_{k=1}^{\ell}\mathcal{B}_{\ell,k}\left(\frac{q}{\lambda},\frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right),...,\frac{q}{\lambda}\cdot...\cdot\left(\frac{q}{\lambda}-\ell+k\right)\right),$$

from which by using (4.3) and some properties of the main part of modulus of smoothness (see, e.g., [13]) we deduce

$$\sup_{\tau>0}\frac{\Omega_{\varphi}^{n}(g,\tau)_u}{\tau^{\frac{q}{\lambda}(r-2\delta)}}\leqslant\mathcal{C}\ \sup_{\tau>0}\frac{\Omega_{\varphi}^{\ell}(g,\tau)_u}{\tau^{\frac{q}{\lambda}(r-2\delta)}}<\infty,\quad n>\frac{q}{\lambda}(r-2\delta).$$

Now we prove (2.13). As for the uniform norm, it is easy to see that

$$\begin{aligned}\sup_{t}v(t)\|h_tv\|_{\infty}&=\sup_{t}v(t)\sup_{s\geqslant0}|k(\gamma_q(t),k(\gamma_q(s))s^{q}v(s)|\\&\leqslant\sup_{x\geqslant0}u(x)\|k_xu\|_{\infty}<\sup_{x\geqslant0}u(x)\|k_x\|_{Z_r(u)},\end{aligned}\tag{4.4}$$

which is bounded by the assumptions. Moreover, by applying the Leibnitz formula we have

$$h_t^{(\ell)}(s)=\begin{cases}\displaystyle\sum_{i=0}^{\ell}\begin{pmatrix}\ell\\i\end{pmatrix}\frac{q!}{(q-\ell+i)!}s^{q+i-\ell}[k(\gamma_q(t),\gamma_q(s))]^{(i)}, & \ell\leqslant q;\\ \displaystyle\sum_{i=\ell-q}^{\ell}\begin{pmatrix}\ell\\i\end{pmatrix}\frac{q!}{(q-\ell+i)!}s^{q+i-\ell}[k(\gamma_q(t),\gamma_q(s))]^{(i)}, & \ell>q.\end{cases}$$

Hence, by using the Bruno di Fáa formula for computing $[k(\gamma_q(t),\gamma_q(s))]^{(i)}$ according to which we have

$$[k(\gamma_q(t),\gamma_q(s))]^{(i)}=\sum_{m=0}^{i}s^{\frac{q}{\lambda}m-i}\mathcal{B}_{i,m}\left(\frac{q}{\lambda},\frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right),...,\frac{q}{\lambda}\cdot...\cdot\left(\frac{q}{\lambda}-i+m\right)\right)k^{(m)}(\gamma_q(t),\gamma_q(s)),$$

with $\mathcal{B}_{0,m}=1$ for all $m=0,...,i$ and $\mathcal{B}_{i,0}=0$ for all $i=1,...,\ell$ and proceeding as already done for the function G, we get

$$\sup_{t}v(t)\sup_{\tau>0}\frac{\Omega_{\varphi}^{j}(h_t,\tau)_v}{\tau^{\frac{q}{\lambda}(r-2\delta)+2q}}<\ \mathcal{C}\ \mathcal{D},$$

where $\mathcal{D}$ is a constant depending on q and λ. Then (2.13) is proved. Proceeding in the same way, it is possible to prove (2.14), i.e.,

$$\sup_s v(s) \sup_{\tau>0} \|h_s\|_{Z_{\frac{q}{\lambda}(r-2\delta)+2[\eta]}(v)} < \ \mathcal{C} \ \ \mathcal{A},$$

where

$$\mathcal{A} = \begin{cases} \displaystyle\sum_{i=0}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} \\ \qquad \displaystyle\sum_{m=0}^{i} \mathcal{B}_{i,m}\left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right), \ldots, \frac{q}{\lambda}\cdot \ldots \cdot \left(\frac{q}{\lambda}-i+m\right)\right), & \ell \leqslant [\eta]; \\ \displaystyle\sum_{i=\ell-[\eta]}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \frac{[\eta]!}{([\eta]-\ell+i)!} \\ \qquad \displaystyle\sum_{m=0}^{i} \mathcal{B}_{i,m}\left(\frac{q}{\lambda}, \frac{q}{\lambda}\left(\frac{q}{\lambda}-1\right), \ldots, \frac{q}{\lambda}\cdot \ldots \cdot \left(\frac{q}{\lambda}-i+m\right)\right), & \ell > [\eta]. \end{cases}$$

□

Acknowledgments. The work was supported by University of Basilicata, research project "Equazioni Integrali con nuclei fortemente oscillanti su intevalli limitati e non". The author is very grateful to Professor Giuseppe Mastroianni for his stimulant and constructive discussions and thanks the referee for his pertinent remarks which have improved the first version of the paper.

References

1. K. E. Atkinson, *The Numerical Solution of Integral Equations of the second kind*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 1997.

2. L. Comtet, *Advanced combinatorics*, D. Reidel Publishing Co., Dordrecht, 1974, the art of finite and infinite expansions. Revised and enlarged edition.

3. A. S. Cvetković and G. V. Milovanović, *The Mathematica package "OrthogonalPolynomials"*, Facta Univ. Ser. Math. Inform., (2004), no. 19, pp. 17–36.

4. B. Della Vecchia and G. Mastroianni, *Gaussian rules on unbounded intervals*, J. Complexity, **19** (2003), no. 3, pp. 247–258.

5. L. Fermo, *A Nyström method for a Class of Fredholm integral equations of the third kind on unbounded domains*, Applied Numerical Mathematics, **59** (2009), pp. 2970–2989.

6. L. Fermo and M. G. Russo, *A Nyström method for Fredholm integral equations with right-hand sides having isolated singularities*, Calcolo, **46** (2009), pp. 61–93.

7. L. Fermo and M. G. Russo, *Numerical Methods for Fredholm integral equations with singular right-hand sides*, to appear on Adv. Comput. Math., doi:10.1007/s10444-009-9137-4, (2009).

8. W. Gauschy, *Algorithm 726:ORTHPOL–a package of routines for generating orthogonal polynomials and Gauss -type quadrature rules*, ACM Trans. Math. Softw., **20** (1994), pp. 21–62.

9. G. H. Golub, *Some modified matrix eigenvalue problems*, Siam Rev., **15** (1973), pp. 318–334.

10. G. H. Golub and J. H. Welsch, *Calculation of Gaussian quadrature rules*, Math. Comput., **23** (1969), pp. 221–230.

11. A. L. Levin and D. S. Lubinsky, *Christoffel functions, orthogonal polynomials and Nevai's conjecture for Freud weights*, Constr. Approx., **8** (1992), pp. 463–535.

12. G. Mastroianni and G. V. Milovanovic, *Some numerical methods for second kind Fredholm integral equation on the real semiaxis*, IMA J. Numer. Anal., **29** (2009), pp. 1046–1066.

13. M. Mastroianni and G. V. Milovanovic, *Interpolation Processes. Basic Theory and Applications.*, Springer, 2008.

14. G. Mastroianni and G. Monegato, *Truncated Gauss-Laguerre quadrature rules,*, Recent trends in numerical analysis,Adv. Theory Comput. Math., Nova Sci. Publ., HUNTINGTON,NY, **3** (2001), pp. 213–221.

15. G. Mastroianni and G. Monegato, *Truncated quadrature rules over* $(0, \infty)$ *and Nyström type methods*, SIAM J. Numer. Anal., **41** (2003), pp. 1870–1892.

16. G. Mastroianni and M. G. Russo, *Lagrange interpolation in some weighted uniform spaces*, Facta Univ. Ser. Math. Inform., (1997), no. 12, pp. 185–201, dedicated to Professor Dragoslav S. Mitrinović (1908–1995) (Niš, 1996).

17. G. Mastroianni and J. Szabados, *Polynomial approximation on the real semiaxis with generalized Laguerre weights*, Stud. Univ. Babeş-Bolyai Math., **52** (2007), no. 4, pp. 105–128.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 195–203

RUNGE-KUTTA NYSTROM METHOD OF ORDER THREE FOR SOLVING FUZZY DIFFERENTIAL EQUATIONS

K. KANAGARAJAN[1] AND M. SAMBATH[1]

Abstract — In this paper we present a numerical algorithm for solving fuzzy differential equations based on Seikkala's derivative of a fuzzy process. We discuss in detail a numerical method based on a Runge-Kutta Nystrom method of order three. The algorithm is illustrated by solving some fuzzy differential equations.

2000 Mathematics Subject Classification: 34A12, 34K28, 65L05.

Keywords: numerical solution, fuzzy differential equation, Runge-Kutta Nystrom method of order 3.

1. Introduction

The fuzzy set theory is a tool that makes it possible to describe vague and uncertain notions. The concept of the fuzzy derivative was first introduced by Chang and Zadeh [4]. Later Dubois and Prade [5] defined and used the extension principle. Other methods have been discussed by Puri and Ralescu [12]. Fuzzy differential equations have been suggested as a way of modelling uncertain and incompletly specified systems and were studied by many researchers [7, 8, 9]. The existence of solutions of fuzzy differential equations has been studied by several authors [2, 3]. It is difficult to obtain an exact solution for fuzzy differential equations and, therefore, several numerical methods were proposed [10, 11]. Abbasbandy and Allahviranloo [1] developed numerical algorithms for solving fuzzy differential equations based on Seikkala's derivative of the fuzzy process introduced in [14]. In this paper, we apply the Runge-Kutta Nystrom method of order three to solve fuzzy differential equations and have established that this method is better than the Euler method. The structure of the paper is organized as follows:

In Section 2, we give some basic definitions and results. In Section 3, we define the initial value problem and discuss the Runge-Kutta Nystrom method of order three. In Section 4, we apply the third order Runge-Kutta Nystrom method to solve the initial value problem and give the convergence result. Finally, in Section 5, we give some examples to illustrate our results.

[1] *Department of Mathematics, Sri Ramakrishna Mission Vidyalaya, College of Arts and Science, Coimbatore 641 020, India.* E-mail: kanagarajank@gmail.com

2. Preliminaries

Consider the initial value problem

$$\begin{cases} y'(t) = f(t, y(t)); \quad a \leqslant t \leqslant b, \\ y(a) = \alpha. \end{cases} \tag{2.1}$$

The point of all Runge-Kutta method is to express the difference between the value of y at t_{n+1} and t_n as

$$y_{n+1} - y_n = \sum_{i=1}^{m} w_i k_i, \tag{2.2}$$

where w_i's are constants and for $i = 1, 2, \cdots m$,

$$k_i = hf\left(t_n + c_i h,\ y_n + h\sum_{j=1}^{i-1} a_{ij} k_j\right). \tag{2.3}$$

Equation (2.2) must be exact for powers of h through h^m, because it must be coincident with Taylor series of order m. Therefore, the truncation error T_m, can be writtern as

$$T_m = \gamma_m h^{m+1} + O(h^{m+2}).$$

The true value of γ_m will generally be much less than the bound of Theorem 2.1. Thus, if the $O(h^{m+2})$ term is small compared to $\gamma_m h^{m+1}$ for small h, then the bound on $\gamma_m h^{m+1}$ will usually be a bound on the error as a whole. The famous nonzero constants c_i, a_{ij} in the Runge-Kutta Nystrom method of order three are

$$c_1 = 0, \quad c_2 = 2/3, \quad c_3 = 2/3, \quad a_{21} = 2/3, \quad a_{32} = 2/3,$$

where $m = 3$. Hence we have (see [6])

$$\begin{aligned} k_1 &= hf\Big(t_i, y_i\Big), \\ k_2 &= hf\left(t_i + \tfrac{2h}{3}, y_i + \tfrac{2}{3}k_1\right), \\ k_3 &= hf\left(t_i + \tfrac{2h}{3}, y_i + \tfrac{2}{3}k_2\right), \\ y_{i+1} &= y_i + \tfrac{1}{8}(2k_1 + 3k_2 + 3k_3), \end{aligned} \tag{2.4}$$

where

$$a = t_0 \leqslant t_1 \leqslant \cdots \leqslant t_N = b \ \text{ and } \ h = \frac{(b-a)}{N} = t_{i+1} - t_i. \tag{2.5}$$

Theorem 2.1. *Let $f(t, y)$ belong to $C^3[a, b]$ and its partial derivatives be bounded and let us assume that there exist positive constants L, M, such that*

$$|f(t, y)| < M, \quad \left|\frac{\partial^{i+j} f}{\partial t^i \partial y^j}\right| < \frac{L^{i+j}}{M^{j-1}}, \quad i + j \leqslant m,$$

then in the Runge-Kutta Nystrom method of order three, we have (see [13])

$$\begin{aligned} y(t_{i+1}) - y_{i+1} &\approx \gamma_3 h^4 + O(h^5), \\ y(t_{i+1}) - y_{i+1} &\approx \frac{25}{108} h^4 M L^3 + O(h^5). \end{aligned}$$

The triangular fuzzy number v is defined by three numbers $a_1 < a_2 < a_3$, where the graph of $v(x)$ (a membership function of the fuzzy number v) is a triangle with the base on the interval $[a_1, a_3]$ and vertex at $x = a_2$. We specify v as $(a_1/a_2/a_3)$. We will write (2.1) $v > 0$ if $a_1 > 0$; (2.2) $v \geqslant 0$ if $a_1 \geqslant 0$; (2.3) $v < 0$ if $a_3 < 0$; and (2.4) $v \leqslant 0$ if $a_3 \leqslant 0$.

Let E be the set of all upper semicontinuous normal convex fuzzy numbers with bounded $r-$level intervals. This means that if $v \in E$, then the $r-$level set

$$[v]_r = \{s \mid v(s) \geqslant r\}, \quad 0 < r \leqslant 1,$$

is a closed bounded interval denoted by

$$[v]_r = [v_1(r), v_2(r)].$$

Let I be a real interval. The mapping $x : I \to E$ is called a fuzzy process and its $r-$level set is denoted by

$$[x(t)]_r = [x_1(t;r),\ x_2(t;r)], \quad t \in I, \quad r \in (0,1].$$

The derivative $x'(t)$ of the fuzzy process $x(t)$ is defined by

$$[x'(t)]_r = [x'_1(t;r),\ x'_2(t;r)], \quad t \in I, \quad r \in (0,1],$$

provided that this equation defines a fuzzy number, as in [14].

Lemma 2.1. *Let $v, w \in E$ and s be a scalar, then for $r \in (0,1]$*

$$\begin{aligned}
[v+w]_r &= [v_1(r)+w_1(r), v_2(r)+w_2(r)],\\
[v-w]_r &= [v_1(r)-w_1(r), v_2(r)-w_2(r)],\\
[v\cdot w]_r &= [\min\{v_1(r)\cdot w_1(r), v_1(r)\cdot w_2(r), v_2(r)\cdot w_1(r), v_2(r)\cdot w_2(r)\},\\
&\quad \max\{v_1(r)\cdot w_1(r), v_1(r)\cdot w_2(r), v_2(r)\cdot w_1(r), v_2(r)\cdot w_2(r)\}],\\
[sv]_r &= s[v]_r.
\end{aligned}$$

3. Fuzzy Cauchy Problem

Consder the fuzzy initial value problem

$$\begin{cases} y'(t) = f(t, y(t)); \quad t \in I = [0,T],\\ y(a) = y_0, \end{cases} \tag{3.1}$$

where f is a continuous mapping from $R_+ \times R$ onto R and $y_0 \in E$ with r-level sets

$$[y_0]_r = [y_1(0;r), y_2(0;r)], \quad r \in (0,1].$$

The extension principle of Zadeh leads to the following definition of $f(t,y)$ when $y = y(t)$ is a fuzzy number:

$$f(t,y)(s) = \sup\{y(\tau) | s = f(t,r)\}, \quad s \in R.$$

It follows that

$$[f(t,y)]_r = [f_1(t,y;r),\ f_2(t,y;r)], \quad r \in (0,1],$$

where

$$\begin{aligned} f_1(t,y;r) &= \min\{f(t,u)| \quad u \in [y_1(r), y_2(r)]\},\\ f_2(t,y;r) &= \max\{f(t,u)| \quad u \in [y_1(r), y_2(r)]\}. \end{aligned} \tag{3.2}$$

Theorem 3.1. *[14] Let f satisfy*

$$|f(t,v) - f(t,\overline{v})| \leqslant g(t, |v - \overline{v}|), \quad t \geqslant 0, \quad v, \overline{v} \in R,$$

where $g : R_+ \times R_+$ is a continuous mapping such that $r \to g(t,r)$ is nondecreasing and the initial value problem

$$u'(t) = g(t, u(t)), \quad u(0) = u_0, \tag{3.3}$$

has a solution on R_+ for $u_0 > 0$ and that $u(t) = 0$ is the only solution of (3.3) *for $u_0 = 0$. Then the fuzzy initial value problem* (3.1) *has a unique solution.*

4. Third-order Runge-Kutta Nystrom method

Let the exact solution $[Y(t)]_r = [Y_1(t;r), Y_2(t;r)]$ be approximated by some $[y(t)]_r = [y_1(t;r), y_2(t,r)]$. From (2.2),(2.3) we define

$$\begin{aligned} y_1(t_{n+1};r) - y_1(t_n;r) &= \sum_{i=1}^{3} w_i k_{i,1}(t_n, y(t_n;r)), \\ y_2(t_{n+1};r) - y_2(t_n;r) &= \sum_{i=1}^{3} w_i k_{i,2}(t_n, y(t_n;r)), \end{aligned} \tag{4.1}$$

where w_i's are constants and

$$[k_i(t, y(t;r))]_r = [k_{i,1}(t, y(t;r), k_{i,2}(t, y(t;r))], \quad i = 1,2,3$$

$$\begin{aligned} k_{i,1}(t_n, y(t_n;r)) &= hf\left(t_n + c_i h \ , \ y_1(t_n) + \sum_{j=1}^{i-1} a_{ij} k_{j,1}(t_n, y(t_n;r))\right), \\ k_{i,2}(t_n, y(t_n;r)) &= hf\left(t_n + c_i h \ , \ y_2(t_n) + \sum_{j=1}^{i-1} a_{ij} k_{j,2}(t_n, y(t_n;r))\right), \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} k_{1,1}(t, y(t;r)) &= \min\Big\{hf\,(t,u)\,|u \in [y_1(t;r), y_2(t;r)]\Big\}, \\ k_{1,2}(t, y(t;r)) &= \max\Big\{hf\,(t,u)\,|u \in [y_1(t;r), y_2(t;r)]\Big\}, \\ k_{2,1}(t, y(t;r)) &= \min\left\{hf\left(t + \tfrac{2}{3}h, u\right)\Big|u \in [z_{1,1}(t, y(t;r)), z_{1,2}(t, y(t;r)]\right\}, \\ k_{2,2}(t, y(t;r)) &= \max\left\{hf\left(t + \tfrac{2}{3}h, u\right)\Big|u \in [z_{1,1}(t, y(t;r)), z_{1,2}(t, y(t;r)]\right\}, \\ k_{3,1}(t, y(t;r)) &= \min\left\{hf\left(t + \tfrac{2}{3}h, u\right)\Big|u \in [z_{2,1}(t, y(t;r)), z_{2,2}(t, y(t;r)]\right\}, \\ k_{3,2}(t, y(t;r)) &= \max\left\{hf\left(t + \tfrac{2}{3}h, u\right)\Big|u \in [z_{2,1}(t, y(t;r)), z_{2,2}(t, y(t;r)]\right\}, \end{aligned} \tag{4.3}$$

where in the third-order Runge-Kutta method

$$\begin{aligned} z_{1,1}(t,y(t;r)) &= y_1(t;r)+\tfrac{2}{3}k_{1,1}(t,y(t;r)),\\ z_{1,2}(t,y(t;r)) &= y_2(t;r)+\tfrac{2}{3}k_{1,2}(t,y(t;r)),\\ z_{2,1}(t,y(t;r)) &= y_1(t;r)+\tfrac{2}{3}k_{2,1}(t,y(t;r)),\\ z_{2,2}(t,y(t;r)) &= y_2(t;r)+\tfrac{2}{3}k_{2,2}(t,y(t;r)). \end{aligned} \tag{4.4}$$

Define

$$\begin{aligned} F[t,y(t;r)] &= 2k_{1,1}(t,y(t;r)+3k_{3,1}(t,y(t;r))+3k_{3,1}(t,y(t;r)),\\ G[t,y(t;r)] &= 2k_{1,2}(t,y(t;r)+3k_{3,2}(t,y(t;r))+3k_{3,1}(t,y(t;r)). \end{aligned} \tag{4.5}$$

The exact and approximate solutions at $t_n,\ 0\leqslant n\leqslant N$ are denoted by $[Y(t_n)]_r=[Y_1(t_n;r),Y_2(t_n;r)]$ and $[y(t_n)]_r=[y_1(t_n;r),y_2(t_n;r)]$, respectively. The solution is calculated by the grid points (2.5). By (4.1),(4.5) we have

$$\begin{aligned} Y_1(t_{n+1};r) &\approx Y_1(t_n;r)+\frac{1}{8}F\left[t_n,Y(t_n;r)\right],\\ Y_2(t_{n+1};r) &\approx Y_2(t_n;r)+\frac{1}{8}G\left[t_n,Y(t_n;r))\right]. \end{aligned} \tag{4.6}$$

We define

$$\begin{aligned} y_1(t_{n+1};r) &= y_1(t_n;r)+\tfrac{1}{8}F\left[t_n,y(t_n;r)\right],\\ y_2(t_{n+1};r) &= y_2(t_n;r)+\tfrac{1}{8}G\left[t_n,y(t_n;r)\right]. \end{aligned} \tag{4.7}$$

The following lemmas will be applied to show the convergence of these approximations. That is

$$\lim_{h\to 0} y_1(t;r)=Y_1(t;r),$$
$$\lim_{h\to 0} y_2(t;r)=Y_2(t;r).$$

Lemma 4.1. *[10] Let the sequence of numbers $\{W_n\}_{n=0}^N$ satisfy*

$$|W_{n+1}|\leqslant A|W_n|+B,\ 0\leqslant n\leqslant N-1,$$

for some given positive constants A and B. Then

$$|W_n|\leqslant A^n|W_0|+B\frac{A^n-1}{A-1},\ 0\leqslant n\leqslant N.$$

Lemma 4.2. *[10] Let the sequence of numbers$\{W_n\}_{n=0}^N,\ \{V_n\}_{n=0}^N$ satisfy*

$$|W_{n+1}|\leqslant |W_n|+A\max\{|W_n|,|V_n|\}+B,$$
$$|V_{n+1}|\leqslant |V_n|+A\max\{|W_n|,|V_n|\}+B,$$

for some given positive constants A and B, and denote

$$U_n=|W_n|+|V_n|,\quad 0\leqslant n\leqslant N.$$

Then

$$U_n\leqslant \overline{A}^nU_0+\overline{B}\frac{\overline{A}^n-1}{\overline{A}-1},\ 0\leqslant n\leqslant N,$$

where $\overline{A}=1+2A$ and $\overline{B}=2B$.

Let $F(t,u,v)$ and $G(t,u,v)$ be obtained by substituting $[y(t)]_r = [u,v]$ into (4.5),

$$F[t,y(t;r)] = 2k_{1,1}(t,y(t;r) + 3k_{3,1}(t,y(t;r)) + 3k_{3,1}(t,y(t;r)),$$
$$G[t,y(t;r)] = 2k_{1,2}(t,y(t;r) + 3k_{3,2}(t,y(t;r)) + 3k_{3,1}(t,y(t;r)).$$

The domain where F and G are defind is therefore

$$K = \{(t,u,v)|0 \leqslant t \leqslant T, \quad -\infty < v < \infty\ , \quad -\infty < u \leqslant v\}.$$

Theorem 4.1. *Let $F(t,u,v)$ and $G(t,u,v)$ belong to $C^3(k)$ and let the partial derivatives of F and G be bounded over K. Then, for arbitrary fixed $r, 0 \leqslant r \leqslant 1$, the approximate solutions (4.6) converge to the exact solutions $Y_1(t;r)$ and $Y_2(t;r)$ uniformly in t.*

Proof. It suffices to show

$$\lim_{h\to 0} y_1(t_N;r) = Y_1(t_N;r),$$
$$\lim_{h\to 0} y_2(t_N;r) = Y_2(t_N;r),$$

where $t_N = T$. For $n = 0,1,\cdots,N-1$, by using the Taylor theorem we get

$$\begin{aligned} Y_1(t_{n+1};r) &= Y_1(t_n;r) + \tfrac{1}{8}F\left[t_n, Y(t_n;r)\right] + \tfrac{25}{108}h^4ML^3 + O(h^5), \\ Y_2(t_{n+1};r) &= Y_2(t_n;r) + \tfrac{1}{8}G\left[t_n, Y(t_n;r))\right] + \tfrac{25}{108}h^4ML^3 + O(h^5), \end{aligned} \tag{4.8}$$

$$W_n = Y_1(t_n;r) - y_1(t_n;r),$$
$$V_n = Y_2(t_n;r) - y_2(t_n;r).$$

Hence from (4.7) and (4.8)

$$W_{n+1} = W_n + \frac{1}{8}\left\{F\left[t_n, Y_1(t_n;r), Y_2(t_n;r)\right] - F\left[t_n, y_1(t_n;r), y_2(t_n;r)\right]\right\} + \frac{25}{108}h^4ML^3 + O(h^5),$$
$$V_{n+1} = V_n + \frac{1}{8}\left\{G\left[t_n, Y_1(t_n;r), Y_2(t_n;r)\right] - G\left[t_n, y_1(t_n;r), y_2(t_n;r)\right]\right\} + \frac{25}{108}h^4ML^3 + O(h^5).$$

Then

$$|W_{n+1}| \leqslant |W_n| + \frac{1}{4}Ph \cdot \max\{|W_n|, |V_n|\} + \frac{25}{108}h^4ML^3 + O(h^5),$$
$$|V_{n+1}| \leqslant |V_n| + \frac{1}{4}Ph \cdot \max\{|W_n|, |V_n|\} + \frac{25}{108}h^4ML^3 + O(h^5),$$

for $t \in [0,T]$ and $P > 0$ is a bound for the partial derivatives of F and G. Thus, by Lemma 4.2

$$|W_n| \leqslant (1 + \frac{1}{2}Ph)^n|U_0| + \left(\frac{25}{54}h^4ML^3 + O(h^5)\right)\frac{(1+\frac{1}{2}Ph)^n - 1}{\frac{1}{2}Ph},$$
$$|V_n| \leqslant (1 + \frac{1}{2}Ph)^n|U_0| + \left(\frac{25}{54}h^4ML^3 + O(h^5)\right)\frac{(1+\frac{1}{2}Ph)^n - 1}{\frac{1}{2}Ph},$$

where $|U_0| = |W_0| + |V_0|$. In particular

$$|W_N| \leqslant (1 + \frac{1}{2}Ph)^N |U_0| + \left(\frac{25}{28}h^3 ML^3 + O(h^4)\right) \frac{(1+\frac{1}{2}Ph)^{\frac{T}{h}} - 1}{P},$$
$$|V_N| \leqslant (1 + \frac{1}{2}Ph)^N |U_0| + \left(\frac{25}{28}h^3 ML^3 + O(h^4)\right) \frac{(1+\frac{1}{2}Ph)^{\frac{T}{h}} - 1}{P}.$$

Since $W_0 = V_0 = 0$, we obtain

$$|W_N| \leqslant \frac{25}{28} ML^3 \left(\frac{e^{\frac{1}{2}Ph} - 1}{P}\right) h^3 + O(h^4),$$
$$|V_N| \leqslant \frac{25}{28} ML^3 \left(\frac{e^{\frac{1}{2}Ph} - 1}{P}\right) h^3 + O(h^4),$$

and if $h \to 0$, we get $W_N \to 0$ and $V_N \to 0$ which completes the proof. □

5. Numerical Examples

Example 5.1. Consider the fuzzy differential equation

$$\begin{cases} y'(t) = -y(t), & t \geqslant 0, \\ y(0) = [0.96 + 0.04r,\ 1.01 - 0.01r]. \end{cases} \tag{5.1}$$

The exact solution is given by

$$Y(t;r) = \left[(0.96 + 0.04r)e^{-t}, (1.01 - 0.01r)e^{-t}\right].$$

At $t = 0.1$ we get

$$Y(0.1;r) = \left[(0.96 + 0.04r)e^{-0.1}, (1.01 - 0.01r)e^{-0.1}\right].$$

With the use of the third-order Runge-kutta Nystrom method the approximate solution is

$$\begin{aligned} y_1(t_{n+1};r) &= y_1(t_n;r)\left(1 - h + \frac{h^2}{2!} - \frac{h^3}{3!}\right), \\ y_2(t_{n+1};r) &= y_2(t_n;r)\left(1 - h + \frac{h^2}{2!} - \frac{h^3}{3!}\right). \end{aligned}$$

The exact and approximate solutions obtained by the Euler method and by the third-orde Runge-Kutta Nystrom method are compared and plotted in Fig. 5.1.

Example 5.2. Consider the fuzzy differential equation

$$\begin{cases} y'(t) = ty(t), & t \in [0,1] \\ y(0) = \left[\sqrt{e} - 0.5(1-r), \sqrt{e} + 0.5(1-r)\right]. \end{cases} \tag{5.2}$$

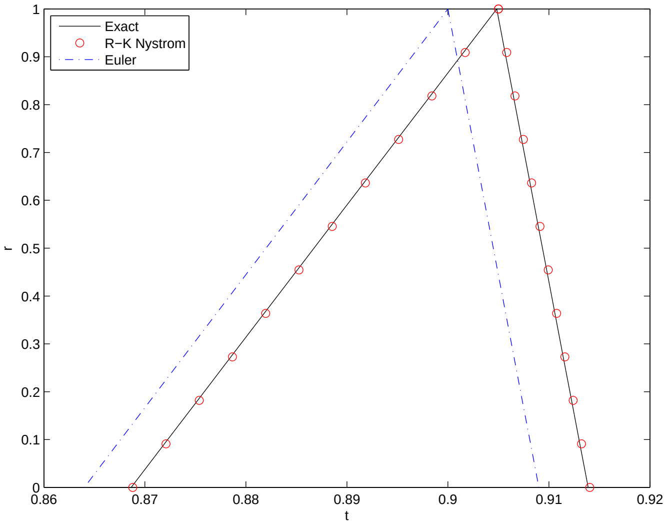


Fig. 5.1. (h=0.2)

The exact solution is given by $Y(t;r) = \left[(\sqrt{e} - 0.5(1-r))e^{\frac{t^2}{2}},\ (\sqrt{e} + 0.5(1-r))e^{\frac{t^2}{2}}\right]$. At $t = 0.1$ we get $Y(0.1;r) = [(\sqrt{e} - 0.5(1-r))e^{0.005},\ (\sqrt{e} + 0.5(1-r))e^{0.005}]$. The exact and approximate solutions obtained by the third-order Runge-Kutta Nystrom method (Eqs. (4.3) - (4.7)), are compared and plotted in Fig. 5.2.

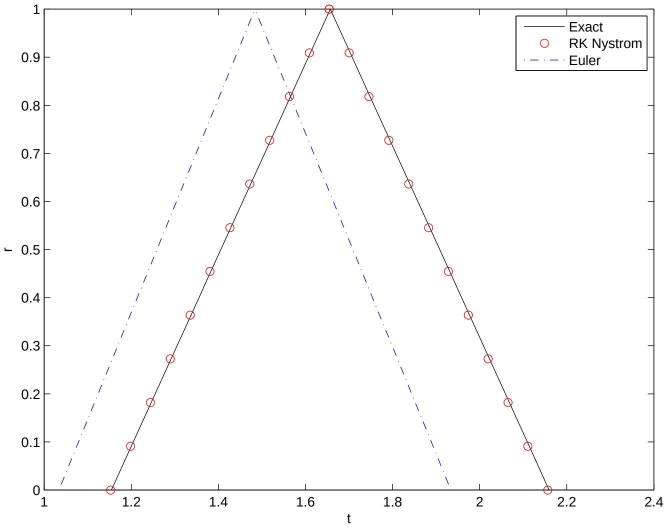


Fig. 5.2. (h=0.5)

6. Conclusions

In this work, we have used the third-order Runge-Kutta Nystrom method to find a numerical solution of fuzzy differential equations. Taking into account the convergence order of the Euler method is $O(h)$ (as given in [10]), a higher order of convergence $O(h^3)$ is obtained by the proposed method. Comparison of the solutions of examples 5.1 and 5.2 shows that the proposed method gives a better solution than the Euler method does.

References

1. S. Abbasbandy and T. Allahviranloo, *Numerical solution of fuzzy differential equations by Taylor method*, Journal of Computational Methods in Applied Mathematics, **2** (2002), pp. 113–124.

2. K. Balachandran and P. Prakash, *Existence of solutions of fuzzy delay differential equations with nonlocal condition*, Journal of the Korea Society for Industrial and Applied Mathematics, **6** (2002), pp. 81–89.

3. K. Balachandran and K. Kanagarajan, *Existence of solutions of fuzzy delay integrodifferential equations with nonlocal condition*, Journal of the Korea Society for Industrial and Applied Mathematics, **9** (2005), pp. 65–74.

4. S.L. Chang and L.A. Zadeh, *On fuzzy mapping and control*, IEEE Trans, Systems Man Cybernet, **2** (1972), pp. 30–34.

5. D. Dubois and H. Prade, *Towards fuzzy differential calculus part 3: Differentiation*, Fuzzy Sets and Systems, **8** (1982), pp. 225-233.

6. M.K. Jain, *Numerical Solution of Differential Equations*, Wiley Eastern Limited, 1984.

7. O. Kaleva, *Fuzzy differential equations*, Fuzzy Sets and Systems, bf 24 (1987), pp. 301-317.

8. O. Kaleva, *The Cauchy problem for fuzzy differential equations*, Fuzzy Sets and Systems, **35** (1990), pp. 389-396.

9. V. Lakshmikantham and R. Mohapatra, *Theory of Fuzzy Differential Equations and Inclutions*, Taylor and Francis, London,(2005).

10. M. Ma, M. Friedman, and A. Kandel, *Numerical solutions of fuzzy differential equations*, Fuzzy Sets and Systems, bf 105 (1999), pp. 133-138.

11. S.Ch. Palligkinis, G. Papageorgiou, and I.Th. Famelis, *Runge-Kutta methods for fuzzy differential equations*, Applied Mathematics and Computation, **209** (2009), pp. 97-105.

12. M.L. Puri and D.A. Ralescu, *Differentials of fuzzy functions*, Journal of Mathematical Analysis and Applications, **91** (1983), pp. 552-558.

13. A. Ralston and P. Rabinowitz,*First Course in Numerical Analysis*, McGraw Hill International Edition, 1978.

14. S. Seikkala, *On the fuzzy initial value problem*, Fuzzy Sets and Systems, **24** (1987), pp. 319-330.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 204–218

COMPUTATION OF THE HARTREE-FOCK EXCHANGE BY THE TENSOR-STRUCTURED METHODS

V. KHOROMSKAIA[1]

Abstract — We propose a novel numerical method for fast and accurate evaluation of the exchange part of the Fock operator in the Hartree-Fock equation which is a (nonlocal) integral operator in $\mathbb{R}^3 \times \mathbb{R}^3$. Usually, this challenging computational problem is solved by analytical evaluation of two-electron integrals using the "analytically separable" Galerkin basis functions, like Gaussians. Instead, we employ the agglomerated "grey-box" numerical computation of the corresponding six-dimensional integrals in the tensor-structured format which does not require analytical separability of the basis set. The point of our method is a low-rank tensor representation of arising functions and operators on an $n \times n \times n$ Cartesian grid and the implementation of the corresponding multi-linear algebraic operations in the tensor product format. Linear scaling of the tensor operations, including the 3D convolution product, with respect to the one-dimension grid size n enables computations on huge 3D Cartesian grids thus providing the required high accuracy. The presented algorithm for evaluation of the exchange operator and a recent tensor method for the computation of the Coulomb matrix are the main building blocks in the numerical solution of the Hartree-Fock equation by the tensor-structured methods. These methods provide a new tool for algebraic optimization of the Galerkin basis in the case of large molecules.

2000 Mathematics Subject Classification: 65F30, 65F50, 65N35, 65F10.

Keywords: Hartree-Fock operator, exchange matrix, canonical model, discrete tensor convolution, tensor-structured methods, tensor-product basis functions.

1. Introduction

In recent decades great progress has been made in the development of canonical and Tucker-type decomposition algorithms as applied to problems of independent component analysis, signal processing and higher order statistics (see [3, 4]) and a comprehensive survey on tensor decomposition methods [18].

Theoretical analysis of the multilinear tensor product approaches for the treatment of some multivariate operators and functions arising in scientific computing was performed in [5, 6, 8, 12]. The application of tensor decomposition algorithms to discretized multivariate functions and operators [10, 14, 13, 11] showed that methods of multi-way analysis can be applied to the numerical solution of basic equations of mathematical physics placing stringent requirements upon the accuracy of results. In particular, the Tucker and canonical tensor product approximations allow to reduce dramatically the complexity of accurate function and operator calculus in $\mathbb{R}^d$, $d \geqslant 3$, realized on large Cartesian grids [15]. Tensor-structured

[1] *Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany.* E-mail: vekh@mis.mpg.de

algorithms acting as "grey-box" schemes, appear to be efficient in electronic structure calculations [11, 15, 16, 17].

Here, we develop a grid-based tensor-structured method for computing the exchange operator in the Hartree-Fock equation using the low-rank representation of the functions and operators involved on an $n \times n \times n$ Cartesian grid. Numerical complexity of the corresponding algorithm scales linearly in the univariate grid size n, $O(n)$.

The Hartree-Fock model provides a meanfield approximation for the ground state of many-electron systems. This implies the solution of a nonlinear eigenvalue problem in $\mathbb{R}^3$

$$\left(-\frac{1}{2}\Delta + V_{nuc} + V_H - V_x\right)\varphi_a(x) = \lambda_a\,\varphi_a(x), \quad x \in \mathbb{R}^3, \tag{1.1}$$

for N_{orb} lowest eigenvalues λ_a and spatial eigenfunctions φ_a $(a = 1, ..., N_{orb})$; in the case of a closed-shell N electron system, $N = 2N_{orb}$. Equation (1.1) corresponds to a nonlinear single-particle Schrödinger equation in $\mathbb{R}^3$, where the potentials V_H and V_x represent a meanfield acting on a single electron generated by the remaining $N - 1$ electrons in the system. Here, the external potential V_{nuc} contains bare Coulomb- or pseudopotentials of the nuclei.

The tensor-structured (TS) methods developed in [14, 10, 13, 15] have been successfully used for highly accurate *grid-based* numerical computations of the Hartree potential and the Coulomb matrix in the Hartree-Fock equation [15, 17]. For efficient computation of the Hartree potential in (1.1),

$$V_H(x) := \int_{\mathbb{R}^3} \frac{\rho(y)}{\|x-y\|}\,dy, \quad x \in \mathbb{R}^3, \tag{1.2}$$

which corresponds to the convolution of the Coulomb potential with the electron density,

$$\rho(y) = 2\sum_{a=1}^{N_{orb}} \varphi_a(y)\varphi_a^*(y), \tag{1.3}$$

we used the low-rank tensor product representation of the electron density ρ and the convolving kernel on an $n \times n \times n$ Cartesian grid and performed multilinear operations in the tensor-product format.

In the present paper, we consider the tensor product approximation of the nonlocal (integral) exchange operator V_x in the Hartree-Fock equation. Note that the calculation of the exchange Galerkin matrix in the Hartree-Fock equation is a challenging problem due to the nonlocal character of the exchange operator

$$\left(V_x\psi\right)(x) := \sum_{b=1}^{N_{orb}} \int_{\mathbb{R}^3} \frac{\varphi_b(x)\varphi_b^*(y)}{\|x-y\|}\psi(y)\,dy, \quad x \in \mathbb{R}^3, \tag{1.4}$$

leading to the integration in six dimensions (see (3.1)). This problem is usually solved analytically by evaluating the so-called two-electron integrals using separable basis sets like Gaussians (see [27, 20] and the references therein).

Here, we propose and implement an agglomerated grid-based method for evaluating the Hartree-Fock exchange (1.4). We apply the tensor product approximation of arising operators and functions on an $n \times n \times n$ Cartesian grid and use multilinear tensor operations providing linear scaling with respect to the one-dimension grid size n, $O(n)^2$. We use the fast tensor product convolution for the multivariate functions in $\mathbb{R}^d$, $d \geqslant 3$, already employed in [15] for evaluating V_H, which provides the complexity $O(d\,n \log n)$; in our case, $d = 3$. The tensor product convolution developed in [13] considerably outperforms the benchmark algorithm based on the 3D Fast Fourier Transform (FFT) having the cost $O(n^3 \log n)$.

To cover the general case of molecular geometries, in the numerical examples, we use equal grid sizes n for three spatial dimensions (a cubic computational box) and do not employ information on molecular symmetry. Therefore, our scheme works as a "grey-box" algebraic algorithm, where as the input data only the discrete representation of the Galerkin basis functions is used. However, the algorithm works as well with arbitrary $n_1 \times n_2 \times n_3$ grids.

Our initial algorithm for evaluating (1.4) has the complexity $O(n \log n R_0^2 + n_{ef}\, R_0^4\, N_{orb})$, where $n_{ef} \ll n$ is the "effective" univariate grid size, and R_0 is the number of Galerkin basis functions. Here we reduce the constant in the linear complexity scaling in n by truncating the regions of computation intervals, where the values of rapidly decaying basis functions (in particular, Gaussians) are less than the threshold controlling the accuracy of computations. Thus, we have for the number of grid points in effective support of the interacting vectors, $n_{ef} = \alpha n$, with α much less than 1.

To reduce the R_0-asymptotics to $O(R_0^3)$, we further apply the canonical-to-Tucker algorithm for decreasing the ranks of intermediate results after every convolution step. The corresponding rank reduction algorithms are considered in [15].

The main advantage of the proposed computational scheme is the ability to avoid "analytically separable" rank-1 basis sets like Gaussians, which are obligatory for the standard approaches. It is well known that the sizes of Gaussian basis sets grow significantly for larger molecules, which makes the related Hartree-Fock problem with the complexity scaling as R_0^3 computationally unfeasible. Here, we use the discretized Gaussians mostly for the sake of convenient comparison of the accuracy of computations with the benchmark results of the standard MOLPRO package [26]. Indeed, we can employ, as the Galerkin basis any appropriate set of functions which are separable algebraically (say, using the Tucker decomposition), with ranks larger or equal to 1 and complying with the approximation requirements. Therefore, the tensor-structured method proposed in this paper provides a new means for algebraic optimization of the Galerkin basis in the case of large molecules.

The accuracy of the computation on a particular grid is estimated by $O(h^2)$, where $h = O(n^{-1})$ is the stepsize of the grid. We achieve $O(h^3)$ accuracy in our evaluation of the exchange matrix by using the Richardson extrapolation on a couple of consequent grids. The univariate size of the computational box for small organic molecules is in the range of $14 \div 20$ $\overset{\circ}{A}$. Since the TS methods enable computations on huge 3D Cartesian grids, the univariate stepsizes of applied grids range from $h \approx 2 \cdot 10^{-2}$ $\overset{\circ}{A}$ for $n = 1024$, up to $h \approx 8 \cdot 10^{-4}$ $\overset{\circ}{A}$ for the benchmark grids with the number of entries $n^3 = 16384^3$.

The rest of the paper is organized as follows. In Section 2, we recall the definitions of the basic rank-structured formats and describe the multilinear tensor-product operations in the rank-R canonical format. In Section 3, we discuss the representation of the exchange operator in the particular Galerkin basis and the discrete computational scheme. The latter does not depend on the character of the basis finctions and allows arbitrary vectors of the

[2] Note that the commonly used attribute "linear in the problem size" for the problems in three spatial dimensions often means linear complexity with respect to the volume size which is $V = n^3$.

canonical agglomerated representation of a given 3D tensor. We give a detailed description of the algorithm and provide a complexity estimate. Section 4 describes the numerical results of computations of the Hartree-Fock exchange matrix for the pseudopotential case of some organic molecules and the all electron case of water molecule using huge 3D Cartesian grids. The figures illustrate the accuracy $O(h^3)$ and the linear scaling of the computation time in the univariate grid size n. Numerical experiments were performed in Matlab 7.6 on a standard SUN station. The results of computations are given in comparison with the output of the standard quantum chemistry package MOLPRO [26].

The tensor-structured computations of the Hartree-Fock exchange, along with the tensor-based algorithms for calculating the Coulomb matrix considered in [15], are the main building blocks in the recent grid-based numerical solution of the Hartree-Fock equation by the TS methods (see the 3D nonlinear EVP solver [16]).

2. Tensor-structured representation of multivariate functions and operators

2.1. Rank-structured tensor approximation

A tensor of order d is a multidimensional array of real/complex data whose elements are referred by using a tensor-product index set $\mathcal{I} = I_1 \times \ldots \times I_d$. We use the common notation

$$A = [a_{i_1,...,i_d} : i_\ell \in I_\ell] \in \mathbb{R}^{\mathcal{I}}, \quad I_\ell = \{1, ..., n_\ell\}, \quad \ell = 1, ..., d,$$

to denote the dth order tensor, and $\mathbf{n}$ for the d-tuple $(n_1, ..., n_d)$. The tensor A is an element of the tensor-product linear space $\mathbb{V}_\mathbf{n} = \otimes_{\ell=1}^d \mathbb{V}_\ell$ with $\mathbb{V}_\ell = \mathbb{R}^{I_\ell}$ equipped with the Euclidean *scalar product* $\langle \cdot, \cdot \rangle : \mathbb{V}_\mathbf{n} \times \mathbb{V}_\mathbf{n} \to \mathbb{R}$, defined as

$$\langle A, B \rangle := \sum_{(i_1,...,i_d)\in\mathcal{I}} a_{i_1,...,i_d} b_{i_1,...,i_d} \quad \text{for } A\,, B \in \mathbb{V}_\mathbf{n}. \tag{2.1}$$

Assume for simplicity that $dim\,\mathbb{V}_\ell = \#I_\ell = n$ for all $\ell = 1, ..., d$, then the number of entries in V amounts to n^d, hence growing exponentially in d.

To get rid of exponential scaling in the dimension, approximate data-sparse "rank structured" representations of tensors in $\mathbb{V}_\mathbf{n}$ can be applied. As the simplest rank structured ansatz, we make use of rank-1 tensors. Specifically, the *tensor product* of vectors $u_\ell = \{u_{\ell,i_\ell}\}_{i_\ell \in I_\ell} \in \mathbb{V}_\ell$ $(\ell = 1, ..., d)$ forms the canonical rank-1 tensor

$$A \equiv [u_\mathbf{i}]_{\mathbf{i}\in\mathcal{I}} = u_1 \otimes ... \otimes u_d \in \mathbb{V}_\mathbf{n} \quad \text{with entries} \quad u_\mathbf{i} = u_{1,i_1} \cdots u_{d,i_d},$$

which requires only dn numbers to store it.

We define a tensor in *the canonical format*

$$A_{(R)} = \sum_{k=1}^{R} c_k u_k^{(1)} \otimes \ldots \otimes u_k^{(d)}, \quad c_k \in \mathbb{R}, \tag{2.2}$$

with normalised vectors $u_k^{(\ell)} \in \mathbb{V}_\ell$ $(\ell = 1, ..., d)$, where the minimal parameter $R \in \mathbb{N}$ in (2.2) is called the rank (or canonical rank) of the tensor. In our tensor-structured computations, we use the rank-R canonical representation for multilinear operations.

Given the rank parameter $\mathbf{r} = (r_1, ..., r_d)$, we can represent the initial tensor A in the so-called Tucker format

$$A \approx A_{(\mathbf{r})} = \sum\nolimits_{\nu_1=1}^{r_1} \cdots \sum\nolimits_{\nu_d=1}^{r_d} \beta_{\nu_1,...,\nu_d}\, v^{(1)}_{\nu_1} \otimes \ldots \otimes v^{(d)}_{\nu_d}, \tag{2.3}$$

with some vectors $v^{(\ell)}_{\nu_\ell} \in \mathbb{V}_\ell = \mathbb{R}^{I_\ell}$ $(1 \leqslant \nu_\ell \leqslant r_\ell)$, which form the orthonormal basis of $\operatorname{span}\{v^{(\ell)}_\nu\}_{\nu=1}^{r_\ell}$ $(\ell = 1, ..., d)$. Here we call the parameter

$$r = \max_\ell \{r_\ell\}$$

the *maximal Tucker rank*. For classes of function related tensors, the choice $r = O(\log n)$ ensures the approximation order $O(1/n)$ [5, 6, 10]. The coefficients tensor $\boldsymbol{\beta} = [\beta_{\nu_1,...,\nu_d}]$, that is an element of the tensor space

$$\mathbb{B}_\mathbf{r} = \mathbb{R}^{J_1 \times ... \times J_d}, \quad J_\ell = \{1, \ldots r_\ell\},\ \ell = 1, \ldots d, \tag{2.4}$$

is called the *core tensor*.

Introducing the (orthogonal) side matrices $V^{(\ell)} = [v^{(\ell)}_1 ... v^{(\ell)}_{r_\ell}]$, we then use a tensor-by-matrix contracted product to represent the Tucker decomposition of $A_{(\mathbf{r})}$,

$$A_{(\mathbf{r})} = \boldsymbol{\beta} \times_1 V^{(1)} \times_2 V^{(2)} ... \times_d V^{(d)}. \tag{2.5}$$

In the present computations, we also use the mixed Tucker-canonical format,

$$A_{(\mathbf{r})} = \left(\sum_{k=1}^{R} b_k u^{(1)}_k \otimes \ldots \otimes u^{(d)}_k \right) \times_1 V^{(1)} \times_2 V^{(2)} \times_3 \ldots \times_d V^{(d)},$$

that is visualized in Fig. 2.1. In this case, the Tucker core is represented by a rank-R canonical tensor. A more detailed description of the tensor decomposition algorithms and of the multigrid rank reduction scheme based on the canonical-to-Tucker approximation is given in [15].

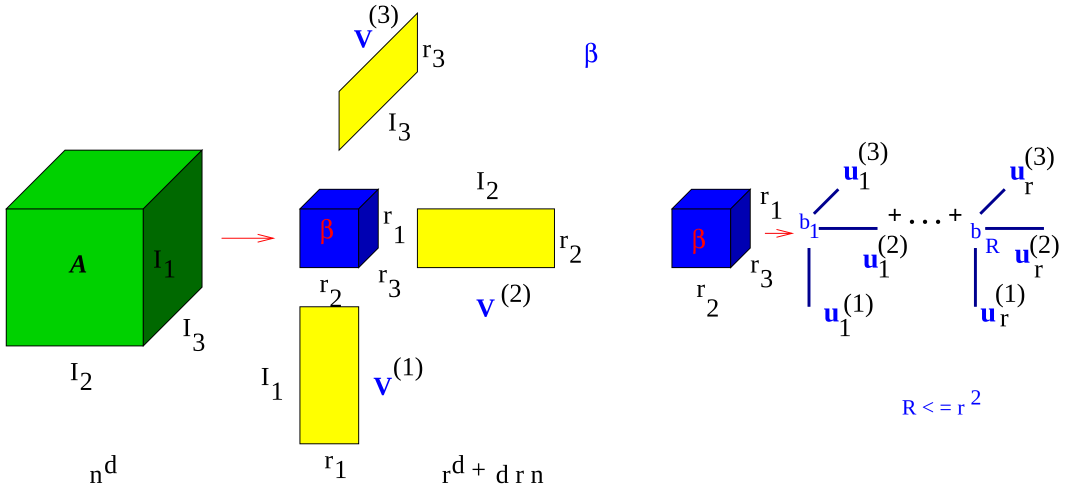


Fig. 2.1. Mixed Tucker-canonical format

2.2. Multilinear operations in the tensor product format

In our numerical scheme, we apply the following linear operations with dth order tensors:

1. summation of tensors;
2. scalar product of tensors;
3. Hadamard product of tensors;
4. convolution product of tensors.

A comprehensive description of the multi-linear tensor-product operations for the multidimensional tensors is presented in the survey [18] (see also [14, 13, 22] for details on function related tensors).

Let us consider tensors A_1, A_2, represented in the rank-R canonical format, (2.2),

$$A_1 = \sum_{k=1}^{R_1} c_k u_k^{(1)} \otimes \ldots \otimes u_k^{(d)}, \quad A_2 = \sum_{m=1}^{R_2} b_m v_m^{(1)} \otimes \ldots \otimes v_m^{(d)}, \tag{2.6}$$

with normalized vectors $u_k^{(\ell)}, v_m^{(\ell)} \in \mathbb{R}^{I_\ell}$. (For simplicity of notation, we consider $n_\ell = n$.)

1. The sum of two canonical tensors given by (2.6) can be written as

$$A_1 + A_2 = \sum_{k=1}^{R_1} c_k u_k^{(1)} \otimes \ldots \otimes u_k^{(d)} + \sum_{m=1}^{R_2} b_m v_m^{(1)} \otimes \ldots \otimes v_m^{(d)}, \tag{2.7}$$

resulting in a canonical tensor with the rank $R_S = R_1 + R_2$. This operation has no cost since it is simply a concatenation of two tensors.

2. For given canonical tensors A_1, A_2, the *scalar product* (2.1) can be computed by

$$\langle A_1, A_2 \rangle := \sum_{k=1}^{R_1} \sum_{m=1}^{R_2} c_k b_m \prod_{\ell=1}^{d} \left\langle u_k^{(\ell)}, v_m^{(\ell)} \right\rangle. \tag{2.8}$$

The calculation of (2.8) includes $R_1 R_2$ scalar products of vectors in $\mathbb{R}^n$, leading to the overall complexity

$$\mathcal{N}_{\langle \cdot,\cdot \rangle} = O(dnR_1R_2).$$

3. The *Hadamard product* $A \odot B \in \mathbb{R}^{\mathcal{I}}$ of two tensors $A, B \in \mathbb{R}^{\mathcal{I}}$, $A = [a_{\mathbf{i}}]$, $B = [b_{\mathbf{i}}]$, of the same size $\mathcal{I}$ is defined componentwise

$$(A \odot B)_{\mathbf{i}} = a_{\mathbf{i}}\, b_{\mathbf{i}}, \quad \mathbf{i} \in \mathcal{I}.$$

Hence, for A_1, A_2 given by (2.6) we tensorize the Hadamard product by

$$A_1 \odot A_2 := \sum_{k=1}^{R_1} \sum_{m=1}^{R_2} c_k b_m \left(u_k^{(1)} \odot v_m^{(1)} \right) \otimes \ldots \otimes \left(u_k^{(d)} \odot v_m^{(d)} \right). \tag{2.9}$$

This leads to the complexity $O(dnR_1R_2)$.

4. In electronic structure calculations, the three-dimensional convolution transform with the Newton convolving kernel, $p(x-y)=\frac{1}{\|x-y\|}$, is the most computationally expensive operation. We employ the discrete version of the multidimensional convolution transform [13]

$$w(x)=\int\limits_{R^3} f(y)p(x-y)dy, \quad x\in\mathbb{R}^3,\ \operatorname{supp} f\subset[-b,b]^3,$$

by applying the standard collocation scheme to discretise the convolution product on the tensor grid

$$\omega_{\mathbf{3},n}:=\omega_1\times\omega_2\times\omega_3, \quad \omega_\ell:=\{-b+(m-1)h:m=1,...,n+1\},\ \ell=1,...,3, \tag{2.10}$$

with a mesh-size $h=2b/n$, with n being an even number. We denote the grid points by $\{x_{\mathbf{m}}\}$, $\mathbf{m}\in\mathcal{M}:=\{1,...,n+1\}^3$. For given piecewise constant basis functions $\{\phi_{\mathbf{i}}\}$, $\mathbf{i}\in\mathcal{I}:=\{1,...,n\}^3$, associated with $\omega_{\mathbf{3},n}$, and a given continuous density function f, let $f_{\mathbf{i}}=f(y_{\mathbf{i}})$ be the representation coefficients of f in $\{\phi_{\mathbf{i}}\}$,

$$f(y)\approx\sum_{\mathbf{i}\in\mathcal{I}} f_{\mathbf{i}}\phi_{\mathbf{i}}(y), \tag{2.11}$$

where $y_{\mathbf{i}}$ is the midpoint of the grid-cell (voxel) $\delta_{\mathbf{i}}:=\delta_{i_1}\times\delta_{i_2}\times\delta_{i_3}$ numbered by $\mathbf{i}\in\mathcal{I}$, with $\delta_{i_\ell}:=[-b+(i_\ell-1)h,-b+i_\ell h]$ $(\ell=1,...,3)$. Now the collocation scheme reads as

$$f*p\approx\{W_{\mathbf{m}}\}_{\mathbf{m}\in\mathcal{M}}, \quad W_{\mathbf{m}}:=\sum_{\mathbf{i}\in\mathcal{I}} f_{\mathbf{i}}\int\limits_{\mathbb{R}^3}\phi_{\mathbf{i}}(y)p(x_{\mathbf{m}}-y)dy, \quad x_{\mathbf{m}}\in\omega_{\mathbf{3},n}.$$

As a first step, we precompute the coefficients

$$p_{\mathbf{i}}=\int\limits_{\mathbb{R}^3} p(y)\phi_{\mathbf{i}}(y)dy, \quad \mathbf{i}\in\mathcal{I}.$$

The coefficient tensor $P=[p_{\mathbf{i}}]\in\mathbb{R}^{\mathcal{I}}$ for the Coulomb potential $p(x-y)=\frac{1}{\|x-y\|}$ is approximated in the rank-R_N canonical tensor format using the optimised *sinc*-quadratures [2], where the rank parameter $R_N=O(|\log\varepsilon|\log n)$ depends logarithmically on both the required accuracy $\varepsilon>0$ and the grid size n. The 3rd order coefficient tensor $F=[f_{\mathbf{i}}]\in\mathbb{R}^{\mathcal{I}}$ is approximated either in the rank $\mathbf{r}=(r,r,r)$ Tucker format or via the canonical model with tensor rank R.

Following [13, 14], the resultant discrete convolution tensor $[W_{\mathbf{m}}]$ can be obtained by copying the corresponding portion of the *tensor convolution* in $\mathbb{V}_{\mathbf{n}}$,

$$F*P:=[z_{\mathbf{j}}], \quad z_{\mathbf{j}}:=\sum_{\mathbf{i}\in\mathcal{I}} f_{\mathbf{i}}p_{\mathbf{j}-\mathbf{i}+\mathbf{1}}, \quad \mathbf{j}\in\mathcal{J}:=\{1,...,2n-1\}^3, \tag{2.12}$$

centred at $\mathbf{j}=\mathbf{n}$, where the sum is over all $\mathbf{i},\mathbf{j}\in\mathcal{I}$, which leads to legal subscripts for $v_{\mathbf{j}-\mathbf{i}+\mathbf{1}}$, i.e., $\mathbf{j}-\mathbf{i}+\mathbf{1}\in\mathcal{I}$.

Approximating F in the rank-R canonical format (see (2.2)) enables us to compute $F*P$ in the form (for two canonical tensors as in (2.6))

$$F*P:=A_1*A_2=\sum_{k=1}^{R_N}\sum_{m=1}^{R} c_k b_m\left(u_k^{(1)}*v_m^{(1)}\right)\otimes\left(u_k^{(2)}*v_m^{(2)}\right)\otimes\left(u_k^{(3)}*v_m^{(3)}\right). \tag{2.13}$$

Assuming that one-dimensional convolutions $u_k^{(\ell)} * v_m^{(\ell)} \in \mathbb{R}^{2n-1}$ can be computed in $O(n \log n)$ operations, the complexity estimate takes the form

$$\mathcal{N}_{\cdot * \cdot} = O(n \log n R_N R).$$

As mentioned above, the tensor product convolution considerably outperforms the conventional 3D FFT having the complexity $O(n^3 \log n)$ (see the numerics in [15]).

3. Calculation of the Hartree-Fock exchange

3.1. Agglomerated representation of the exchange operator

The exchange Galerkin matrix K_{ex} with respect to the normalized basis set $\{g_k\}_{k=1,\ldots R_0}$ is given by

$$K_{ex} = \{K_{ij}\}_{i,j=1}^{R_0}, \quad K_{ij} := -\frac{1}{2} \int\limits_{\mathbb{R}^3} \int\limits_{\mathbb{R}^3} g_i(x) \frac{\tau(x,y)}{\|x-y\|} g_j(y) dx dy, \quad i,j = 1, \ldots R_0, \tag{3.1}$$

where the density matrix $\tau(x,y)$ is defined as

$$\tau(x,y) = \sum_{a=1}^{N_{orb}} \varphi_a(x) \varphi_a(y),$$

over all occupied orbitals a.

The low cost of the three-dimensional convolution using the canonical representation of the convolving tensors makes possible the agglomerated numerical evaluation of the exchange matrix in the Fock operator. For this purpose, we divide the integration in (3.1) into the following steps. First, we compute the convolutions of the pointwise products of molecular orbitals with the vectors from the normalized Gaussian basis set

$$W_{aj}(x) = \int\limits_{\mathbb{R}^3} \frac{\varphi_a(y) g_j(y)}{\|x-y\|} dy \quad a = 1, \ldots, N_{orb}, \ j = 1, \ldots, R_0. \tag{3.2}$$

These are further used for the calculation of contributions to the Galerkin matrix elements from every orbital a,

$$V_{ij,a} = \int\limits_{\mathbb{R}^3} \varphi_a(x) g_i(x) W_{aj}(x) dx, \quad i,j = 1, \ldots R_0. \tag{3.3}$$

The entries of the exchange matrix are then the sums of the corresponding values over all orbitals

$$K_{ij} = \sum_{a=1}^{N/2} V_{ij,a}, \quad i,j = 1, \ldots N/2. \tag{3.4}$$

We compute the exchange matrix (3.1) using the discrete tensor product representation of arising functions and operators.

The orbital of the molecule is considered as an expansion over the basis set of well separable continuous functions $g_k(x)$,

$$\varphi_a(x) = \sum_{k=1}^{R_0} c_{a,k} g_k(x), \quad x = (x_1, x_2, x_3) \in \mathbb{R}^3, \tag{3.5}$$

where the basis functions g_k, $k = 1, \ldots, R_0$, are represented as the rank-R canonical tensor products,

$$g_k(x) = \sum_{\nu=1}^{R} g_{k,\nu}^{(1)}(x_1)\, g_{k,\nu}^{(2)}(x_2)\, g_{k,\nu}^{(3)}(x_3), \tag{3.6}$$

with 1, 2, 3 designating spatial dimensions.

3.2. Discrete computational scheme

GTOs are used as conventional basis sets in electronic structure calculations due to their separability in spatial variables which is used in the analytical evaluation of the integrals in the calculation of the Hartree and exchange potentials.

In the following, for numerical illustrations, we choose the discretized Gaussians as vectors in the rank-1 canonical representations of the basis functions, mainly for the sake of convenient verification of the results of computations (the corresponding Galerkin matrix) with the standart MOLPRO output [26].

The rank-1 GTO basis functions $g_k(x)$, $k = 1, \ldots R_0$, are given by (3.6) with $R = 1$, where $g_{k,1}^{(\ell)}(x_\ell)$ denotes the generalized univariate Gaussians. The univariate Gaussians $g_k^{(\ell)}(x_\ell) = g_{k,1}^{(\ell)}(x_\ell)$, $\ell = 1, 2, 3$, are functions with an infinite support given as

$$g_k^{(\ell)}(x_\ell) = (x_\ell - A_{\ell,k})^{p_{\ell,k}} \exp(-\alpha_k (x_\ell - A_{\ell,k})^2), \quad x_\ell \in \mathbb{R},\ \alpha_k > 0,$$

where $p_{\ell,k} = 0, 1, \ldots$ is the polynomial degree, and the points $(A_{1,k}, A_{2,k}, A_{3,k}) \in \mathbb{R}^3$ specify the positions of nuclei in a molecule. In our scheme, we use the *discrete* basis functions (given by vectors of the canonical tensor representation (2.2)) which are constructed by discretizing the Gaussians on the given tensor grid by using the associated piecewise constant basis functions.

Assume that the molecule is embedded in a certain fixed computational box $[-b, b]^3$ with a suitable $b > 0$. For simplicity of notation, we take $n_\ell = n$ equal for all dimensions. We introduce the equidistant tensor grid $\omega_{\mathbf{3},n}$ (see (2.10) in §2). The grid points are denoted by $\{x_{\mathbf{m}}\}$, $\mathbf{m} \in \mathcal{M} := \{1, ..., n+1\}^3$. We use a representation like (2.11) with $f(x) = g_k(x)$, where the rank-1 coefficients tensor G_k is given by the values of ℓ-mode functions $g_k^{(\ell)}$ at the centers $y_{i_\ell}^{(\ell)}$ of intervals of the univariate grid $[x_{i_\ell}^{(\ell)}, x_{i_\ell+1}^{(\ell)}]$, $i_\ell = 1, \ldots, n$. This results in canonical vectors of length n with entries $\{g_k^{(\ell)}(y_{i_\ell}^{(\ell)})\}_{i_\ell=1}^n$,

$$\gamma_k^{(\ell)} = \{g_k^{(\ell)}(y_{i_\ell}^{(\ell)})\}_{i_\ell=1}^n \in \mathbb{R}^n, \quad \text{for } \ell = 1, 2, 3,\ k = 1, \ldots R_0, \tag{3.7}$$

such that $G_k = \gamma_k^{(1)} \otimes \gamma_k^{(2)} \otimes \gamma_k^{(3)}$. By summing the tensor products of the canonical vectors with the corresponding weights $c_{a,k}$ as in (2.2) we obtain the discrete representation of the orbital φ_a, $a = 1, \ldots N_{orb}$, in the rank-R_0 canonical format,

$$U_a = \sum_{k=1}^{R_0} c_{a,k} \gamma_k^{(1)} \otimes \gamma_k^{(2)} \otimes \gamma_k^{(3)}, \quad c_{a,k} \in \mathbb{R}, \tag{3.8}$$

where R_0 is the number of basis functions. This discretization can be considered as a representation in the Galerkin set of basis functions $\{G_k\}$ obtained by representing the initial continuous basis set $\{g_k\}$ via piecewise constant basis functions $\{\phi_{\mathbf{i}}\}$ on the uniform grid (see (3.7)).

We use the rank-R_N canonical tensor product representation of the coefficient tensor P for the Newton potential $\frac{1}{\|x-y\|}$, on the same grid. This tensor is precomputed by using the optimized sinc-quadratures [2, 13], where the rank parameter $R_N = O(|\log \varepsilon| \log n)$ depends logarithmically on both the required accuracy $\varepsilon > 0$ and the univariate grid size n. In particular, for our computations the tensor P, representing the Newton potential has the canonical rank in the range $20 \leqslant R_N \leqslant 30$, depending on the one-dimension grid size n and the accuracy requirements $\varepsilon > 0$.

We present Algorithm 1 describing the computational scheme for evaluating (3.2) - (3.4) in the tensor product format[3].

Algorithm 1 Computation of the Exchange Matrix in Tensor Arithmetics

```
*Input data: rank-$R_0$ canonical tensors $U_a \in \mathbb{V}_{\mathbf{n}}$, $a = 1, \dots, N_{orb}$,* rank $R_N$ tensor $P \in \mathbb{V}_{\mathbf{n}}$,
*rank-*1 *canonical tensors* $G_k = \gamma_k^{(1)} \otimes \gamma_k^{(2)} \otimes \gamma_k^{(3)}$ , $k = 1, \dots R_0$, and the filtering threshold $\varepsilon_F > 0$.
**(A0)** Find effective supports $\sigma_j \subset [-b, b]$ for $\gamma_j$, $j = 1, \dots, R_0$, by $\varepsilon_F$-thresholding,

    $\sigma_j = \sigma_j^{(1)} \times \sigma_j^{(2)} \times \sigma_j^{(3)}$, where $\sigma_j^{(\ell)} = \{i : |\gamma_j^{(\ell)}(x_i)| \geqslant \varepsilon_F\} \subset \{1, \dots, n\}$,  $\ell = 1, 2, 3$.

**for** $a = 1, \dots, N_{orb}$
**for** $k = 1, \dots, R_0$
**(A)** Compute the Hadamard product $\theta_{a,k} = U_a \odot G_k$ of tensors $U_a$ and $G_k$ by using (2.9).
**(B)** Compute the tensor convolution $\Theta_{a,k} = \theta_{a,k} * P$ by using (2.13).
**for** $j = 1, \dots, R_0$
**(C)** Compute the restricted scalar products in the window $\sigma_j$,

    $K_{a,k,j} = \langle \theta_{a,j}, \Theta_{a,k} \rangle_{|\sigma_j}$,

**end for** $j$
**end for** $k$
**end for** $a$.
**(D)** Sum matrix elements over all orbital indices, $K_{kj} = \sum_{a=1}^{N_{orb}} K_{a,k,j}$, for $k, j = 1, \dots, R_0$.
*Output data*: the exchange matrix $K = \{K_{kj}\}_{k,j=1}^{R_0}$.
```

Lemma 3.1. *The complexity of Algorithm 1 for the computation of the exchange Galerkin matrix K_{ex} in the Hartree-Fock equation using the discretized GTO basis is estimated by*

$$W_{K_{ex}} = O(N_{orb} R_N (R_0^2 n \log n + R_0^4 n_{ef})).$$

Proof. This estimate includes the cost of the evaluation of convolutions in (3.2) for every orbital, $O(N_{orb} R_N R_0^2 n \log n)$, and the scalar product (3.3) of the tensor $\Theta_{a,k}$ with the products of the orbitals and Gaussians, $O(N_{orb} R_N R_0^4 n_{ef})$. □

Since the canonical rank R_N of tensor P corresponding to the Coulomb potential depends only logarithmically on n, it can be treated as a constant.

[3] The Hadamard product $\theta_{a,j} = U_a \odot G_j$ in the Algorithm 1 can be either (1) stored for all vectors γ_k at step **(A)** or (2) recomputed before evaluating the scalar products at step **(C)** . Due to the very low cost of this operation, and large storage requirements for the case of large grids, $O(R_0^2 n)$, we prefer the case (2).

Table 3.1. **Rank reduction for $\Theta_{a,k}$ in the computation of the exchange matrix for the pseudopotential case of some molecules**

	CH_4	CH_3OH	C_2H_5OH
R_Θ	1250	1875	2775
$r_T = 12$, $\epsilon_T \leqslant 10^{-7}$, R_{RED}	80	90	110
$coef_R$	15	20	23
$r_T = 10$, $\epsilon_T \leqslant 10^{-6}$, R_{RED}	50	70	100
$coef_R$	25	26	27

Remark 3.1. Notice that the rank reduction of the canonical tensor $\Theta_{a,k}$ after step (3.2) reduces the complexity to

$$W_{K_{ex},red} = O(N_{orb} R_0^3 n_{ef}). \tag{3.9}$$

In the case of large molecules, further optimization up to the $O(N_{orb} R_0^2 n_{ef})$-complexity is possible due to the rank reduction applied to the rank-R_0 orbitals (tensors U_a).

Remark 3.2. The rank-R_0 tensors U_a, $a = 1, \ldots, N_{orb}$ representing the orbitals can be chosen as the Galerkin basis set $\{G_a\}$, $a = 1, \ldots, N_{orb}$, where N_{orb} is usually much smaller than R_0. This may relax the critical dependence $O(R_0^4)$ as in Lemma 3.1 above (see also Lemma 3.1 in [16]).

3.3. Rank reduction

The maximal initial rank of tensor $\Theta_{a,k}$ at the step (B) in Algorithm 1 is given by $R_\Theta = R_N R_0$. We perform the rank reduction for this tensor by the canonical-to-Tucker (C2T) and Tucker-to-canonical (T2C) algorithm introduced and discussed in details in [9, 15]. In particular, it is shown that the multigrid version of the C2T algorithm applied to 3-rd order rank-R canonical tensors has a linear complexity with respect to all parameters of the input tensor: the canonical rank R, the Tucker rank r, and the univariate grid size n. Thus, we can reduce the complexity of Algorithm 1 to (3.9) solely by multilinear algebraic methods which do not take into account any previous knowledge on the molecular structure.

Table 3.1 shows the average rank reduction by the C2T and T2C algorithms applied to the tensor $\Theta_{a,k}$ in the calculations for CH_4, CH_3OH and C_2H_5OH molecules. We present the approximate canonical ranks R_{RED} (and respective Tucker ranks r_T) of the tensors corresponding to the *largest value*, over the parameters $a = 1, \ldots, N_{orb}$, $k = 1, \ldots R_0$, to achieve the prescribed approximation error ϵ_T,

$$R_{RED} = \max_{1 \leqslant a \leqslant N_{orb},\, 1 \leqslant k \leqslant R_0} R_{RED}(a,k),$$

where $R_{RED}(a,k)$ denotes the reduced canonical rank of $\Theta_{a,k}$, for given φ_a and g_k. Table 3.1 gives also the corresponding reduction coefficient, $coef_R = \frac{R_\Theta}{R_{RED}}$.

3.4. Window technique for fast computation of inner products

We compute the algebraic tensor representation of the discrete electron orbitals U_a given by (3.8) using the coefficients of their representation in the discrete Gaussian basis set $\gamma_k^{(\ell)}$.

It turns out by the construction that most of $\gamma_k^{(\ell)}$ have a local character (fast exponential decay) with respect to the size of the whole computation domain $[-b, b]^3$. Therefore, we precompute the effective supports of the canonical vectors $\gamma_k^{(\ell)}$ by truncating their parts that are lower than some predefined threshold $\varepsilon > 0$. We call this the "windowing" procedure for finding an active interval for each Gaussian. In our case, the resulting effective vector size of the canonical vectors is, on average, 3 times smaller than the corresponding grid size n even for small molecules. The resulting "effective" univariate grid size is $n_{ef} = \alpha n$, with $\alpha = \alpha(\varepsilon) < 1$. For example, for small molecules $\alpha \sim 0.2 \div 0.3$ for $\varepsilon = 10^{-5}$. This leads to a reduced cost of the scalar products compared to the univariate grid size n.

We expect a much stronger windowing effect in the case of large molecules, since it can be directly applied to Hadamard products $U_a \odot G_k$.

4. Numerical experiments

We tested the presented tensor-structured method by computating the exchange Galerkin matrix for the following molecules:

- all electron case : H_2O ($N_{orb} = 5,\ R_0 = 41$), CH_4 ($N_{orb} = 5, R_0 = 55$);
- pseudopotential case: CH_4 ($N_{orb} = 4, R_0 = 50$), CH_3OH ($N_{orb} = 7,\ R_0 = 75$) and C_2H_5OH ($N_{orb} = 11,\ R_0 = 111$).

The calculations were performed on a standard SUN station using Matlab 7.6. Figures present the absolute error of our computations compared with the corresponding exchange matrix calculated by the benchmark package MOLPRO [26].

The computational box $[-b, b]^3$ for small molecules is in the range of $2b = 14\ \overset{\circ}{A}$ for H_2O, and $2b = 20\ \overset{\circ}{A}$ for CH_4, C_2H_5OH, and CH_3OH. The developed tensor-structured algorithm enables one to computate the Hartree-Fock exchange on huge $n \times n \times n$ 3D Cartesian grids, with the number of entries up to $n^3 = 16384^3$. This corresponds to the usage of univariate mesh-sizes from $h \approx 2 \cdot 10^{-2}$ for grids with $n = 1024$ to $h \approx 8 \cdot 10^{-4}\ \overset{\circ}{A}$ for grids with $n = 16384$.

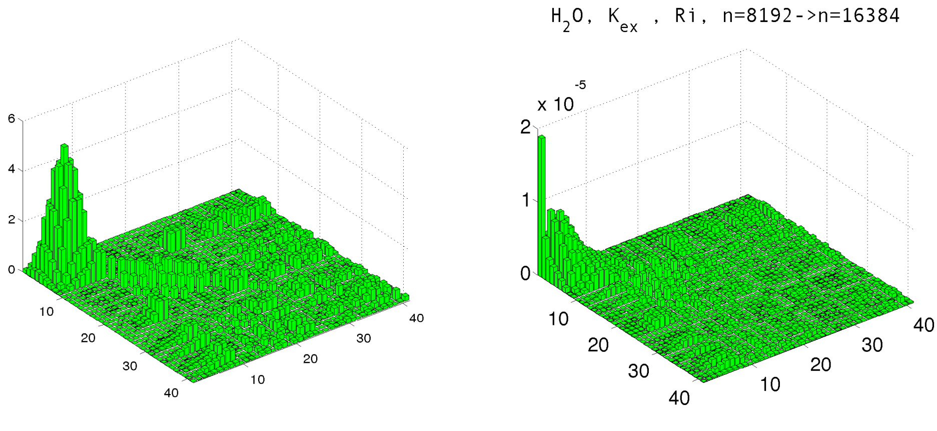


Fig. 4.1. Left: entries of the exchange matrix for the all electron case of H_2O. Right: absolute error in the tensor-structured computation on $n \times n \times n$ 3D Cartesian grids with $n = 8192$ and $n = 16384$

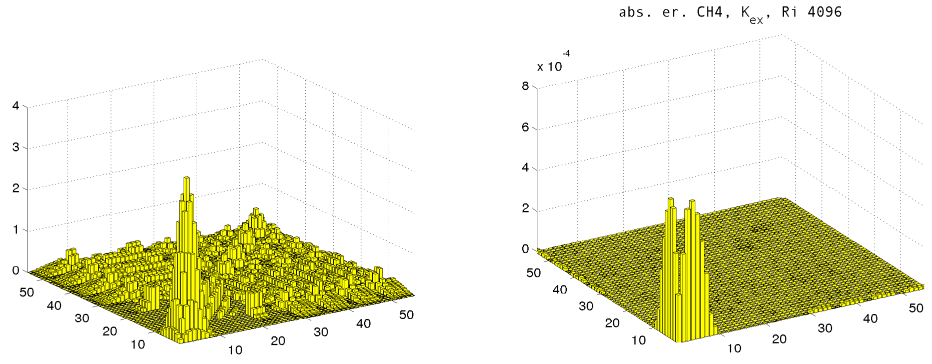


Fig. 4.2. Left: entries of the exchange matrix for CH_4. Right: absolute approximation error in the tensor-structured computations on 3D grids with a one-dimension size $n = 2048$ and $n = 4096$

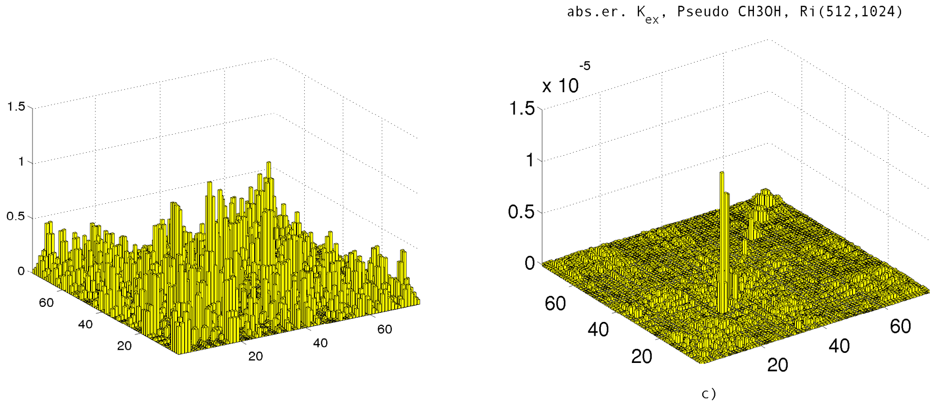


Fig. 4.3. Left: exchange matrix for the *pseudopotential* case of CH_3OH. Right: absolute approximation error in the tensor-product computation on $n \times n \times n$ grids with $n = 512$ and $n = 1024$

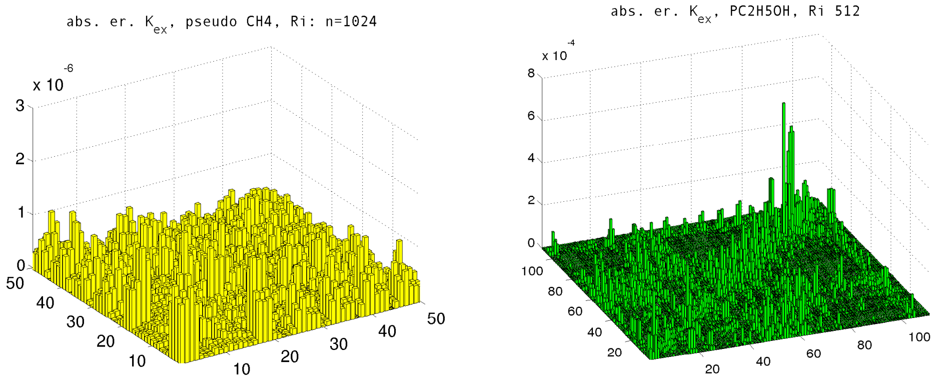


Fig. 4.4. a) Absolute error in the tensor-product computation of the exchange matrix for the pseudopotential case of CH_4 (left) and C_2H_5OH (right) molecules

4.1. All electron case

For a molecules with a moderate size R_0 of basis sets like CH_4 or H_2O grid-sizes up to $n = 16384$ are computationally feasible for MATLAB, which is equivalent to computations with $4.398 \cdot 10^{12}$ nodes in the volume. These grids provide the resolution of the strong cusps in basis functions corresponding to the core electrons in a molecule, thus enabling accurate computations of the exchange matrix for the all electron case.

Computation of the exchange Galerkin matrix for the all electron case of the H_2O molecule is a challenging problem due to the "sharp" Gaussians corresponding to the core electrons of the Oxygen atom. Figure 4.1 (left) shows the absolute values of the exchange matrix entries for H_2O, Figure 4.1 (right) shows the absolute error in tensor-structured computations of this matrix using the Richardson extrapolation on $n \times n \times n$ 3D Cartesian grids with $n = 8192$ and $n = 16384$. We achieve a high accuracy $1.89 \cdot 10^{-5}$ in the "cusp area", the remaining entries are computed with the absolute error in the range of $10^{-6} \div 10^{-8}$.

Figure 4.2 (left) displays the absolute values of the exchange matrix of CH_4 and Fig. 4.2 (right) shows the absolute error in tensor-structured computations reaching an accuracy of 10^{-4} by using the Richardson extrapolation on grids with $n = 2048$ and $n = 4096$. Again, the matrix entries, apart from the "cusp area", are computed with a much higher accuracy.

4.2. Pseudopotential case

We consider the pseudopotential case for larger molecules, achieving an accuracy of up to 10^{-6}, using smaller 3D grids with a one-dimension size $n = 1024$. The Fortran version of the loops including steps (C) – (D) in Algorithm 1 can improve dramatically the CPU computation time.

Figure 4.3 (left) shows the entries of the exchange matrix of the CH_3OH molecule and Fig. 4.3 (right) shows that tensor-structured computations for this molecule using the Richardson extrapolation on grids with $n = 512, 1024$ yield an accuracy of $\sim 10^{-5}$.

Table 4.1. **Comparison of the relative times**

n^3	64^3	128^3	256^3	512^3	1024^3
H_2O	1	1.3	2.0	3.2	8.0
CH_4 (ps)	1	1.3	2.0	3.6	8.9
CH_3OH (ps)	1	1.3	1.9	3.3	5.1

Figure 4.4 shows the absolute error in tensor-structured computations for the exchange matrices in the pseudopotential case of the CH_3OH (left) and C_2H_5OH (right) molecules, respectively. For CH_3OH the Richardson extrapolation on two consequent grids with $n = 512, 1024$ yields an accuracy of $\sim 10^{-5}$, while for C_2H_5OH we obtain $7 \cdot 10^{-4}$, already on small 3D grids with the one-dimension size $n = 256,\ 512$.

Table 4.1 presents the linear scaling of the relative computation times with respect to the one-dimension grid size n, in respective units of the coarsest grid calculations ($n = 64$) for one orbital.

Our calculations demonstrate that the TS methods give promising results for their application in the computation of multivariate integrals in quantum chemistry. They appear to be promising for computations for large molecules by either discretization on fine 3D Cartesian grids or from the viewpoint of post-Hartree-Fock models.

Acknowledgments. The author cordially acknowledges Dr. Heinz-Jürgen Flad (TU Berlin) and PD DrSci. Boris Khoromski (MIS MPI Leipzig) for the formulation of the problem and valuable discussions at all stages of this work. The author gratefully acknowledges Prof. Dr. R. Schneider (TU Berlin) for useful discussions.

References

1. E. Acar, T. G. Kolda, and D. M. Dunlavy, *An Optimization Approach for Fitting Canonical Tensor Decompositions*, Technical Report Number SAND2009-0857, Sandia National Laboratories, Albuquerque,

NM and Livermore, CA, February 2009.

2. C. Bertoglio and B. N. Khoromskij,*Low rank tensor-product approximation of projected Green kernels via sinc-quadratures*, Preprint 79/2008, MPI MIS Leipzig, 2008 (submitted).

3. L. De Lathauwer, B. De Moor, and J. Vandewalle, *On the best rank-1 and rank-$(R_1, ..., R_N)$ approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., **21** (2000), pp. 1324–1342.

4. L. De Lathauwer, B. De Moor, J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., **21** (2000), pp. 1253–1278.

5. I.P. Gavrilyuk, W. Hackbusch, and B.N. Khoromskij, *Hierarchical Tensor-Product Approximation to the Inverse and Related Operators in High-Dimensional Elliptic Problems*, Computing, **74** (2005), pp. 131–157.

6. I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation of a class of operator-valued functions*, Math. Comp., **74** (2005), pp. 681–708.

7. H.-J. Flad, W. Hackbusch, B.N. Khoromskij, and R. Schneider, *Concept of Data-Sparse Tensor-Product Approximation in Many-Particle Modeling*, Matrix Methods: Theory, Algorithms, Applications, World Scientific publishing, Singapoure, (2010), pp. 313–347.

8. W. Hackbusch, B.N. Khoromskij, and E.E. Tyrtyshnikov, *Hierarchical Kronecker tensor-product approximations*, J. Numer. Math. **13** (2005), pp. 119–156.

9. V. Khoromskaia, *Numerical Solution of the Hartree-Fock equation by the Tensor-Structured Methods*, PhD Dissertation, (in preparation), MPI MiS, Leipzig, 2010.

10. B. N. Khoromskij, *Structured Rank-$(r_1, ..., r_d)$ Decomposition of Function-related Tensors in $\mathbb{R}^d$*, Comp. Meth. in Applied Math., **6** (2006), no 2, pp. 194–220.

11. B. N. Khoromskij, *On Tensor Approximation of Green Iterations for Kohn-Sham equations*, Computing and Visualisation in Science, **11** (2008), pp. 259–271.

12. B. N. Khoromskij, *Tensor-structured Preconditioners and Approximate Inverse of Elliptic Operators in $\mathbb{R}^d$* , J. Constructive Approximation, **30** (2009), pp. 599–620.

13. B. N. Khoromskij, *Fast and Accurate Tensor Approximation of Multivariate Convolution with Linear Scaling in Dimension*, Preprint MPI MIS, Leipzig **36** (2008); J. Comp. Appl. Math., to appear.

14. B. N. Khoromskij and V. Khoromskaia, *Low Rank Tucker Tensor Approximation to the Classical Potentials*, Central European J. of Math., **5** (2007), no. 3, pp. 1–28.

15. B.N. Khoromskij and V. Khoromskaia, *Multigrid Tensor Approximation of Function Related Arrays*, SIAM J. on Sci. Comp., **31** (2009), no. 4, pp. 3002–3026.

16. B.N. Khoromskij, V. Khoromskaia, and H.-J. Flad, *Numerical solution of the Hartree-Fock equation in multilevel tensor-structured format*, Preprint 44/2009 MPI MIS, Leipzig, 2009.

17. B. N. Khoromskij, V. Khoromskaia, S. R. Chinnamsetty, and H.-J. Flad, *Tensor Decomposition in Electronic Structure Calculations on 3D Cartesian Grids*, J. of Comput. Phys.**228** (2009), pp. 5749–5762.

18. T.G. Kolda and B. W. Bader, *Tensor Decompositions and Applications*, SIAM Rev., **51** (2009), no. 3, pp. 455–500.

19. C. Le Bris, ed.., *Handbook of Numerical Analysis, Vol. X, Computational Chemistry*, North-Holland, 2003.

20. C. Le Bris. *Computational chemistry from the perspective of numerical analysis*, Acta Numerica, **14** (2005), pp. 363–444.

21. I. Oseledets, D. Savostyanov, and E. E. Tyrtyshnikov, *Cross approximation in electron density computations*, Num. Lin. Alg. Appl., submitted, 2009.

22. I. Oseledets, D. Savostyanov, and E. E. Tyrtyshnikov. *Linear algebra for tensor problems*, Computing, **85** (2009), pp. 169–188.

23. R. Polly, H.-J. Werner, F. R. Manby, and P. J. Knowles. *Fast Hartree-Fock theory using density fitting approximations*, Mol.Phys., **102** (2004), pp. 2311–2321.

24. J. Sielk, H. F. von Horsten, F. Krüger, R. Schneider, and B. Hartke, *Quantum-mechanical wavepacket propagation in a sparse, adaptive basis of interpolating Gaussians with collocation*, Physical Chemistry Chemical Physics, **11** (2009), pp. 463–475.

25. J. VandeVondele, M. Krack, F. Mohamed, M. Parinello, Th. Chassaing, and J. Hutter, *QUICKSTEP: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach*, Comp. Phys. Comm., **167** (2005), pp. 103–128.

26. H.-J. Werner, P.J. Knowles et al, *MOLPRO, version 2002.10, a package of ab initio programs for electronic structure calculations.*

27. T. Yanai, G. Fann, Z. Gan, R. Harrison, and G. Beylkin, *Multiresolution quantum chemistry: Hartree-Fock exchange*, J. Chem. Phys., **121** (2004), no. 14, pp. 6680–6688.

COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 2, pp. 219–232

A FLUX-CORRECTED FINITE ELEMENT METHOD FOR CHEMOTAXIS PROBLEMS

R. STREHL[1] , A. SOKOLOV[1], D. KUZMIN[1], AND S. TUREK[1]

Abstract — An implicit flux-corrected transport (FCT) algorithm has been developed for a class of chemotaxis models. The coefficients of the Galerkin finite element discretization has been adjusted in such a way as to guarantee mass conservation and keep the cell density nonnegative. The numerical behaviour of the proposed high-resolution scheme is tested on the blow-up problem for a minimal chemotaxis model with singularities. It has also been shown that the results for an *Escherichia coli* chemotaxis model are in good agreement with the experimental data reported in the literature.

2000 Mathematics Subject Classification: 35B36; 65N30; 92C15; 92C17.

Keywords: chemotaxis models, pattern formation, flux limiters, finite elements.

1. Introduction

Chemotaxis, an oriented movement towards or away from regions of higher concentrations of certain chemicals, plays a vitally important role in the evolution of many living organisms. The chemotactical response gives numerous creatures, ranging from bacteria and protozoa to tissue cells, a chance to find more favourable locations in their environments. This feature improves their ability to search for food, detect the location of mates or escape danger. Chemotaxis finds many medical and biological applications, including bacteria/cells aggregation and pattern formation processes, tumour growth, etc.

The first mathematical description of chemotactical processes was given by Keller and Segel [14, 15], who modeled the aggregation of the slime mold amoeba *Dictyostelium discoideum.* Their work was followed by the development of sophisticated models for various chemotaxis problems [2, 5, 13, 20, 27]. The numerical treatment of chemotaxis equations has also been addressed by many authors [7, 9, 10, 16, 23, 28]. However, some implementation aspects still call for further research. In particular, it is difficult to design a robust, accurate, and efficient numerical algorithm that does not produce negative densities or concentrations [7]. In the present paper, positivity constraints for the Galerkin finite element discretization are enforced using a generalized flux-corrected transport (FCT) algorithm [4, 17, 19, 29].

A representative class of chemotaxis models based on advection-reaction-diffusion equations is considered in what follows. Following the notation of [13], the nonlinear PDE systems to be solved in a two-dimensional domain $\Omega \subset \mathbb{R}^2$ are written in the unified form

$$u_t = \nabla \cdot (D(u)\nabla u - A(u)\, B(c)\, C(\nabla c)) + q(u) \quad \text{in} \quad \Omega, \tag{1.1}$$

$$c_t = d\Delta c - s(u)\, c + g(u)\, u \quad \text{in} \quad \Omega, \tag{1.2}$$

[1] *Institute of Applied Mathematics, LS III, TU Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany.* E-mail: robert.strehl@math.uni-dortmund.de, asokolow@math.uni-dortmund.de

where $u(\boldsymbol{x}, t)$ denotes the cell density and $c(\boldsymbol{x}, t)$ is the chemoattractant concentration. The functional dependence of the involved coefficients on u and c defines a particular model. A variety of complex chemotactical processes can be modelled in this way [2, 5, 16, 20, 27].

The above transport equation for u and the reaction-diffusion equation for c are endowed with the initial conditions

$$u|_{t=0} = u_0, \qquad c|_{t=0} = c_0 \qquad \text{in} \quad \Omega, \tag{1.3}$$

and homogeneous Neumann boundary conditions are prescribed on the boundary Γ of Ω

$$\boldsymbol{n} \cdot (D(u)\,\nabla u - A(u)\,B(c)\,C(\nabla c)) = \boldsymbol{n} \cdot \nabla c = 0 \qquad \text{on} \quad \Gamma. \tag{1.4}$$

One of the numerical problems to be dealt with is due to the rapid growth of solutions to system (1.1)–(1.2) in a small neighbourhood of certain points or curves. In particular, the blow-up phenomenon, or a singular spiky behaviour of exact solutions, may give rise to nonphysical oscillations if the employed numerical scheme is not guaranteed to satisfy the discrete maximum principle (DMP). The available numerical techniques include various positivity-preserving finite volume and finite element schemes [7, 11, 25], operator-splitting, fractional step algorithms [23, 28], interior penalty discontinuous Galerkin methods [9, 10], and cell-overcrowding prevention models [6, 8, 22]. The flux-corrected transport paradigm described in Section 2 represents a promising new approach to the blow-up problem.

Another interesting application of the proposed methodology is the numerical prediction of bacteria pattern formations. The nonlinear dependence of $B(c)$ on the chemoattractant concentration c can produce travelling waves [3, 24]. Attracting and repulsing substances behave in different ways. As shown by the numerical study of Aida et al. [1, 2] and confirmed experimentally, the pattern for small values of the parameter $\chi = B(c) = \mathit{const}$ resembles a honeycomb, stripe or perforated stripe, while a chaotic spot pattern is observed for large values of χ. In Section 3, the proposed FEM-FCT algorithm is applied to 2D pattern formation problems. The results presented are in good agreement with the available experimental data.

2. Flux-corrected transport

A segregated approach to the numerical solution of the nonlinear model problem (1.1)–(1.2) was adopted. In each time step, the transport equation for the chemoattractant concentration $c(\boldsymbol{x}, t)$ is solved prior to that for the cell density $u(\boldsymbol{x}, t)$. Both equations are written in weak form and discretized in space using (conforming) bilinear finite elements. The discretization in time is performed by the implicit Euler method; Crank-Nicolson and fractional step schemes will be considered in a forthcoming paper. The system of linearized algebraic equations consists of two decoupled subproblems for the unknowns u^{n+1} and c^{n+1} at time t_{n+1}:

$$[\boldsymbol{M}(1) + \Delta t \boldsymbol{L}(D^n) - \Delta t \boldsymbol{K}(c^n)]\, u^{n+1} = \boldsymbol{M}(1) u^n + \Delta t \boldsymbol{q}^n, \tag{2.1}$$

$$[\boldsymbol{M}(1) + \Delta t \boldsymbol{L}(d) - \Delta t \boldsymbol{M}(s^n)]\, c^{n+1} = \boldsymbol{M}(1) c^n + \Delta t \boldsymbol{M}(g^n) u^n, \tag{2.2}$$

where $\boldsymbol{M}(\cdot)$ denotes the (consistent) mass matrix, $\boldsymbol{L}(\cdot)$ is a discrete diffusion operator, and $\boldsymbol{K}(c)$ is a discrete transport operator due to the chemotactical flux $A(u)\,B(c)\,C(\nabla c)$. The entries of $\boldsymbol{M}(\cdot)$, $\boldsymbol{L}(\cdot)$, $\boldsymbol{K}(c)$ and $\boldsymbol{q}^n$ are defined in (2.3)–(2.6). In (2.1)–(2.2) the setting $D^n = D(u^n)$, $s^n = s(u^n)$ and $g^n = g(u^n)$ is used.

Given a set of piecewise-polynomial basis functions $\{\varphi_i\}$, the standard Galerkin discretization yields the following formulae for the coefficients of the matrices $\boldsymbol{M}$, $\boldsymbol{L}$, $\boldsymbol{K}$ and the vector $\boldsymbol{q}^n$:

$$m_{ij}(\psi) = \int\limits_\Omega \varphi_i\varphi_j\psi\,\mathrm{d}\mathbf{x}, \qquad \psi \in \{1, s(u), g(u)\}, \tag{2.3}$$

$$l_{ij}(\psi) = \int\limits_\Omega \nabla\varphi_i \cdot \nabla\varphi_j\psi\,\mathrm{d}\mathbf{x}, \qquad \psi \in \{D(u), d\}, \tag{2.4}$$

$$k_{ij}(c) = \int\limits_\Omega \nabla\varphi_i \cdot A(\varphi_j)\,B(c)\,C(\nabla c)\,\mathrm{d}\mathbf{x}, \tag{2.5}$$

$$q_i^n = \int\limits_\Omega \varphi_i q_j(u^n)\,\mathrm{d}\mathbf{x}. \tag{2.6}$$

In formula (2.5), the discontinuous concentration gradient ∇c can be replaced by a superconvergent approximation constructed using (slope-limited) reconstruction techniques [18].

As was shown by Kuzmin *et al.* [18, 19, 17], positivity constraints can be readily enforced at the discrete level using a conservative manipulation of the matrices $\boldsymbol{M}$ and $\boldsymbol{K}$. The former is approximated by its diagonal counterpart $\boldsymbol{M}_L$ constructed using row-sum mass lumping

$$\boldsymbol{M}_L := \operatorname{diag}\{m_i\}, \qquad m_i = \sum_j m_{ij}(1). \tag{2.7}$$

Next, all negative off-diagonal entries of $\boldsymbol{K}$ are eliminated by adding an artificial diffusion operator $\boldsymbol{D}$. For conservation reasons, this matrix must be symmetric with zero row and column sums. For any pair of neighbouring nodes i and j, the entry d_{ij} is defined as [18, 19]

$$d_{ij} = \max\{-k_{ij}, 0, -k_{ji}\}, \qquad j \neq i. \tag{2.8}$$

Note that $d_{ji} = d_{ij}$, so that the operator $\boldsymbol{D}$ is a symmetric matrix. The diagonal coefficients d_{ii} are defined so that the row and column sums of $\boldsymbol{D}$ are equal to zero

$$d_{ii} = -\sum_{j\neq i} d_{ij}. \tag{2.9}$$

The result is a positivity-preserving discretization of low order. By construction, the added perturbation to the discrete problem admits a conservative decomposition into a sum of internodal fluxes. The mass lumping error and artificial diffusion received by the node i satisfy

$$(M(1)u - M_L u)_i = \sum_j m_{ij}u_j - m_i u_i = \sum_{j\neq i} m_{ij}(u_j - u_i), \tag{2.10}$$

$$(Du)_i = \sum_j d_{ij}u_j = \sum_{j\neq i} d_{ij}u_j + d_{ii}u_i = \sum_{j\neq i} d_{ij}(u_j - u_i). \tag{2.11}$$

Let f denote the difference between the residuals of the low-order scheme and that of the underlying Galerkin approximation. By virtue of the above flux decomposition, we have

$$f_i = \sum_{j \neq i} f_{ij}, \qquad f_{ji} = -f_{ij}, \qquad \forall j \neq i. \tag{2.12}$$

To achieve a high resolution while keeping the scheme positivity-preserving, each flux is multiplied by a solution-dependent correction factor $\alpha_{ij} \in [0,1]$ and inserted into the right-hand side of the nonoscillatory low-order scheme. The original Galerkin discretization corresponds to the setting $\alpha_{ij} := 1$. It may be used in regions where the numerical solution is smooth and well-resolved. The setting $\alpha_{ij} := 0$ is appropriate in the neighborhood of steep fronts.

In essence, the off-diagonal entries of the sparse matrices $\boldsymbol{M}$ and $\boldsymbol{K}$ are replaced by

$$m^*_{ij} := \alpha_{ij} m_{ij}, \qquad k^*_{ij} := k_{ij} + (1-\alpha_{ij}) d_{ij},$$

while the diagonal coefficients of the flux-corrected Galerkin operators are given by

$$m^*_{ii} := m_i - \sum_{j \neq i} \alpha_{ij} m_{ij}, \qquad k^*_{ii} := k_{ii} - \sum_{j \neq i} (1-\alpha_{ij}) d_{ij}.$$

In implicit FEM-FCT schemes [17, 18, 19], the optimal values of α_{ij} are determined using Zalesak's algorithm [29]. The limiting process begins with cancelling all fluxes that are diffusive in nature and tend to flatten the solution profiles. The required modification is

$$f_{ij} := 0 \qquad \text{if} \quad f_{ij}(u_j - u_i) > 0,$$

where u is a positivity-preserving solution of low order [17, 18, 19]. The remaining fluxes are truly antidiffusive, and the computation of α_{ij} involves the following algorithmic steps:

1. Compute the sums of positive/negative antidiffusive fluxes into node i

$$P_i^+ = \sum_{j \neq i} \max\{0, f_{ij}\}, \qquad P_i^- = \sum_{j \neq i} \min\{0, f_{ij}\}.$$

2. Compute the distance to a local extremum of the auxiliary solution u

$$Q_i^+ = \max\{0, \max_{j \neq i}(u_j - u_i)\}, \qquad Q_i^- = \min\{0, \min_{j \neq i}(u_j - u_i)\}.$$

3. Compute the nodal correction factors for the net increment to node i

$$R_i^+ = \min\left\{1, \frac{m_i Q_i^+}{\Delta t P_i^+}\right\}, \qquad R_i^- = \min\left\{1, \frac{m_i Q_i^-}{\Delta t P_i^-}\right\}.$$

4. Check the sign of the antidiffusive flux and apply the correction factor

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{if } f_{ij} > 0, \\ \min\{R_i^-, R_j^+\}, & \text{otherwise.} \end{cases}$$

In the context of chemotaxis problems, the above limiting strategy ensures that the cell density $u(\boldsymbol{x},t)$ and the concentration $c(\boldsymbol{x},t)$ remain nonnegative. However, the resultant algebraic systems are strongly nonlinear and must be solved iteratively. As a remedy, the antidiffusive fluxes f_{ij} for an implicit FCT algorithm can be linearized about a low-order predictor, as proposed by Kuzmin [17]. This linearized version of FEM-FCT is the method that we use to solve our system (1.1)–(1.2) in the present paper. In contrast to nonlinear FCT algorithms, antidiffusive flux correction is done explicitly after calculation of the low-order solution. Therefore, it is readily applicable to linear and nonlinear problems alike. For a detailed presentation of the FEM-FCT methodology, including theoretical analysis (stability, positivity, convergence) and technical implementation details (data structures, matrix assembly), we refer the interested reader to [17, 18, 19] and other publications by Kuzmin *et al.*

3. Numerical results

In this section, the developed FEM-FCT algorithm is applied to chemotaxis models that call for the use of positivity-preserving discretization techniques.

3.1. Blow-up in the center of the domain

The minimal Keller-Segel chemotaxis model

$$u_t = \Delta u - \nabla\cdot(u\nabla c), \tag{3.1}$$

$$c_t = \Delta c - c + u, \tag{3.2}$$

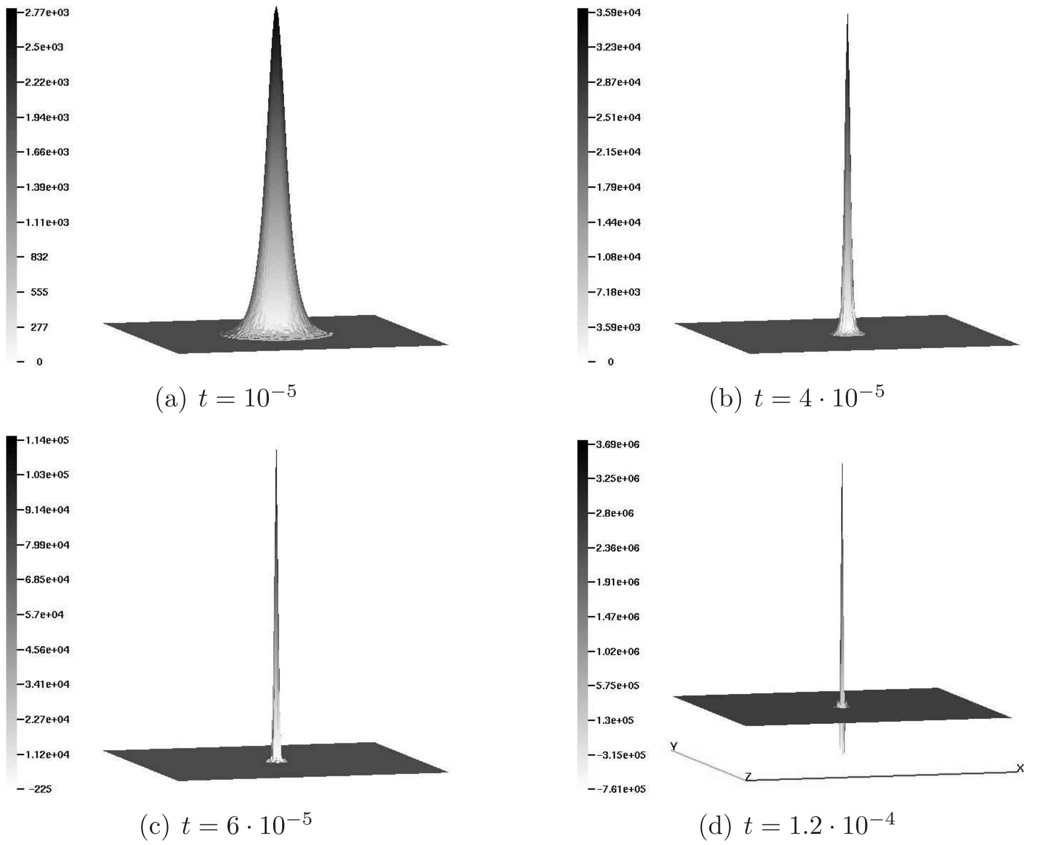
can be written in the form (1.1)–(1.2). The corresponding parameter settings are as follows:

$$\begin{gathered} A(u)=u, \quad B(c)=1, \quad C(\nabla c)=\nabla c, \quad D(u)=1, \\ d=1, \quad s(u)=1, \quad g(u)=1, \quad q(u)=0. \end{gathered}$$

The following bell-shaped initial conditions [7] are prescribed in $\Omega=(0,1)^2$ at $t=0$

$$\begin{aligned} u_0(x,y) &= 1000\, e^{-100((x-0.5)^2+(y-0.5)^2)}, \\ c_0(x,y) &= 500\, e^{-50((x-0.5)^2+(y-0.5)^2)}. \end{aligned} \tag{3.3}$$

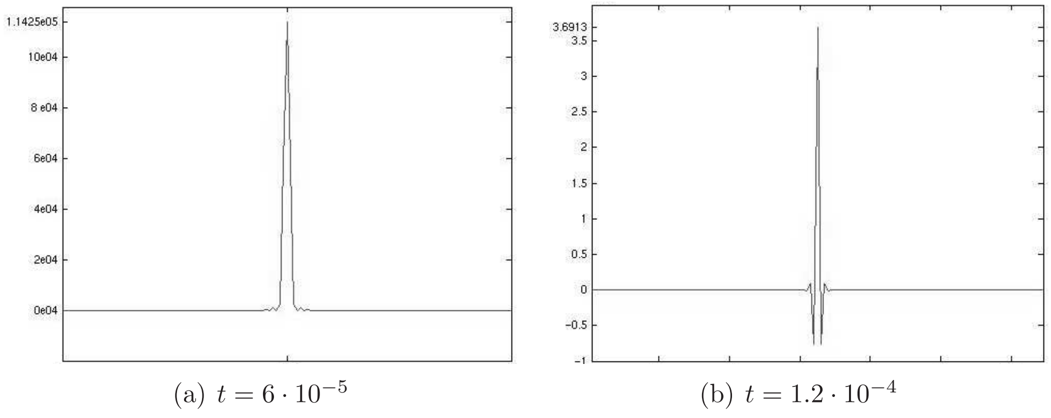
The radially symmetric solution to the initial boundary value problem (3.1)–(3.3) has a peak in the center of the domain Ω, where the blow-up of u and c occurs in finite time [12, 26]. The numerical solutions to the blow-up problem are computed on a uniform grid of bilinear finite elements. The mesh size and time step are given by $h=1/128$ and $\Delta t=10^{-6}$, respectively. Snapshots of the results obtained with the standard Galerkin discretization of system (3.1)–(3.2) are displayed in Fig. 3.1. The two diagrams in Fig. 3.2 show the distribution of the cell density u along the horizontal line $y=0.5$ at two time instants. Note that u becomes negative at a certain intermediate time. The nonphysical negative values grow rapidly as time evolves, which leads to an abnormal termination of the simulation run.



(a) $t = 10^{-5}$ (b) $t = 4 \cdot 10^{-5}$

(c) $t = 6 \cdot 10^{-5}$ (d) $t = 1.2 \cdot 10^{-4}$

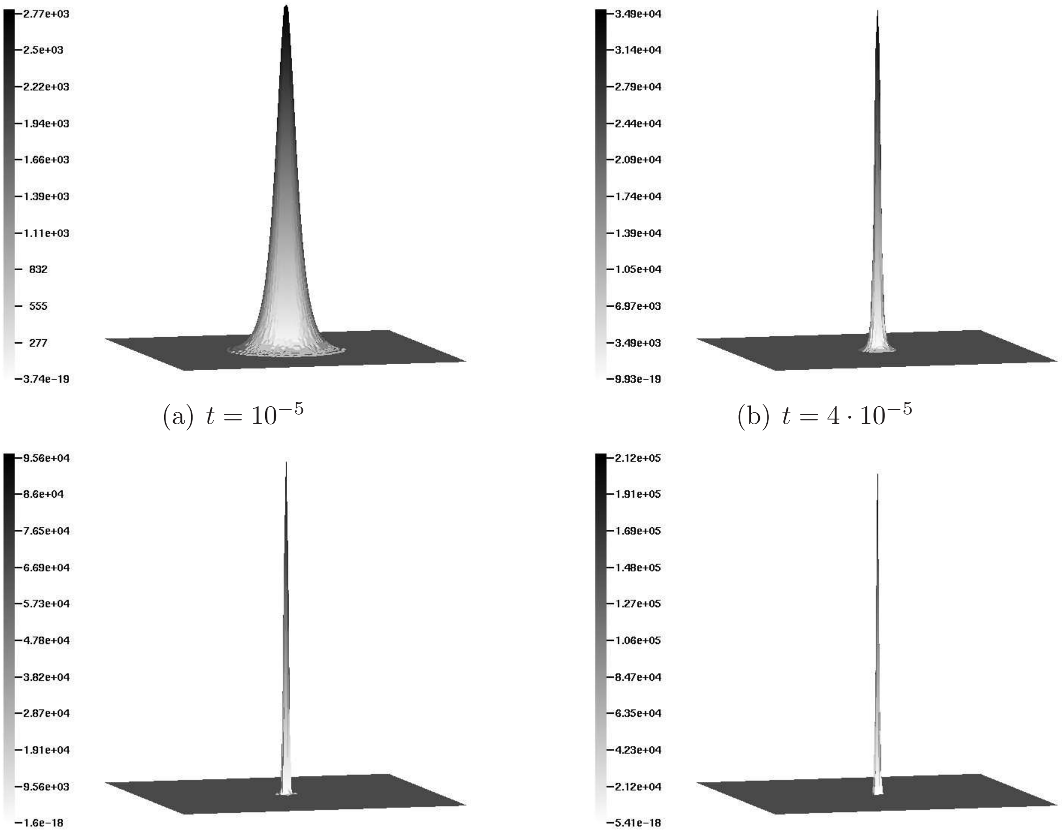
Fig. 3.1. Blow-up in the center, standard Galerkin scheme, $h = \frac{1}{128}$, $\Delta t = 10^{-6}$.



(a) $t = 6 \cdot 10^{-5}$ (b) $t = 1.2 \cdot 10^{-4}$

Fig. 3.2. Blow-up in the center, Galerkin solution at $y = 0.5$, $h = \frac{1}{128}$, $\Delta t = 10^{-6}$.

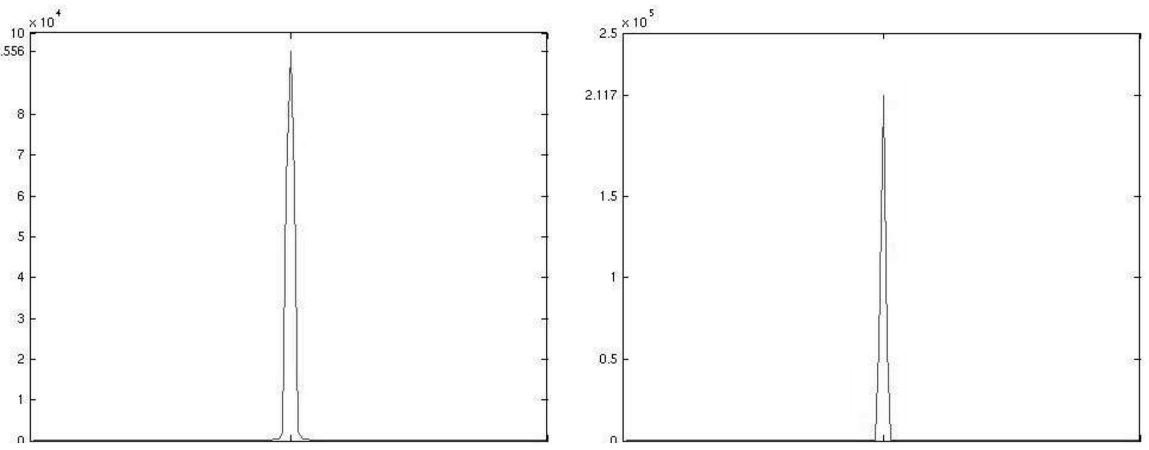
Next, we apply the FCT correction to the discretized form of the minimal chemotaxis system (3.1)–(3.2) and perform simulations with the same parameter settings as before. The numerical solutions presented in Figs. 3.3 and 3.4 are seen to be positive and nonoscillatory.



(a) $t = 10^{-5}$ (b) $t = 4 \cdot 10^{-5}$

(c) $t = 6 \cdot 10^{-5}$ (d) $t = 1.2 \cdot 10^{-4}$

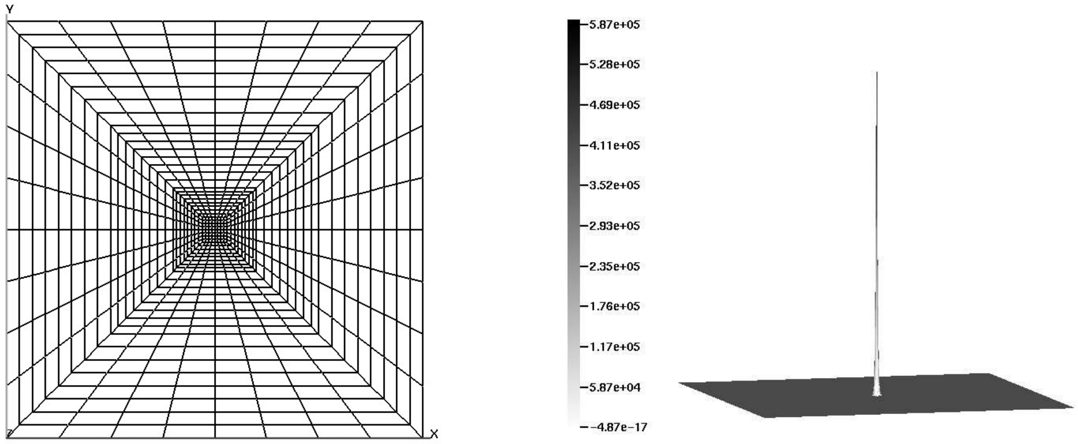
Fig. 3.3. Blow-up in the center, FEM-FCT scheme, $h = \frac{1}{128}$, $\Delta t = 10^{-6}$.



(a) $t = 6 \cdot 10^{-5}$ (b) $t = 1.2 \cdot 10^{-4}$

Fig. 3.4. Blow-up in the center, FEM-FCT solution at $y = 0.5$, $h = \frac{1}{128}$, $\Delta t = 10^{-6}$.

The accuracy of a finite element approximation can be easily improved by means of local mesh refinement in underresolved regions. Since the solution of system (3.1)–(3.2) blows up in the center of the square domain, it is worthwhile to refine the mesh around this point, so as to achieve a higher resolution of the growing peak. For a fair comparison, the number of elements (degrees of freedom) should not exceed that for the uniform grid employed previously. The FEM-FCT solution presented in Fig. 3.5 (b) was computed on a nonuniform mesh constructed from that shown in Fig. 3.5 (a) using 5 levels of global refinement. The total number of elements is $13,312 < 128^2$. Due to the higher mesh density around the point of blow-up, the peak of the cell density is approximately twice as high as that in Fig. 3.3 (d). The peak heights variation with uniform and adaptive mesh refinement is illustrated by the diagram in Fig. 3.6.



(a) adaptive mesh, level 3, $h \leqslant 1/8$

(b) cell density u at $t = 1.2 \cdot 10^{-4}$

Fig. 3.5. Blow-up in the center, adaptive FEM-FCT scheme, 13, 312 elements, $\Delta t = 10^{-6}$.

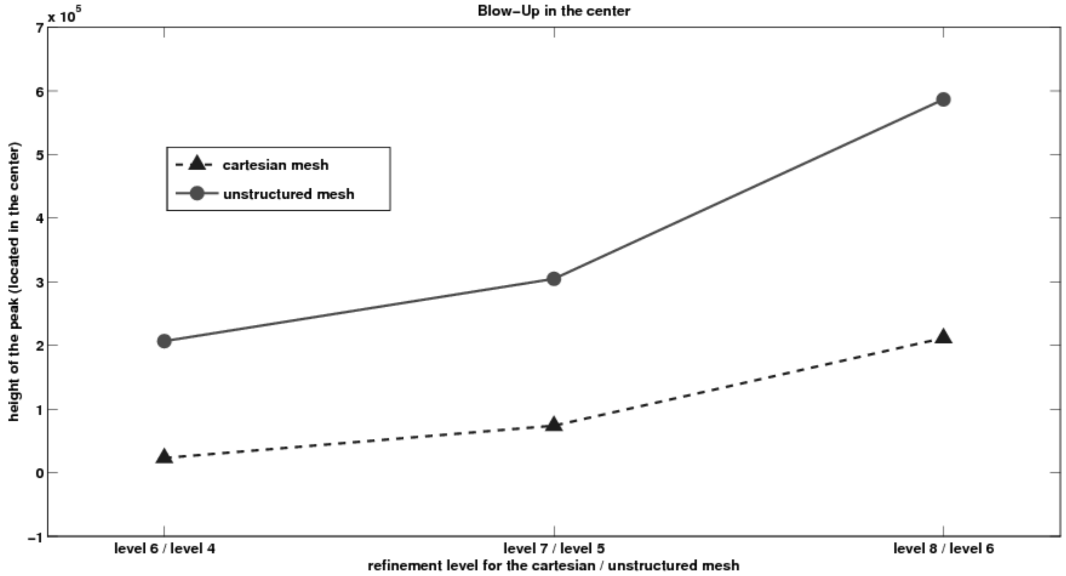


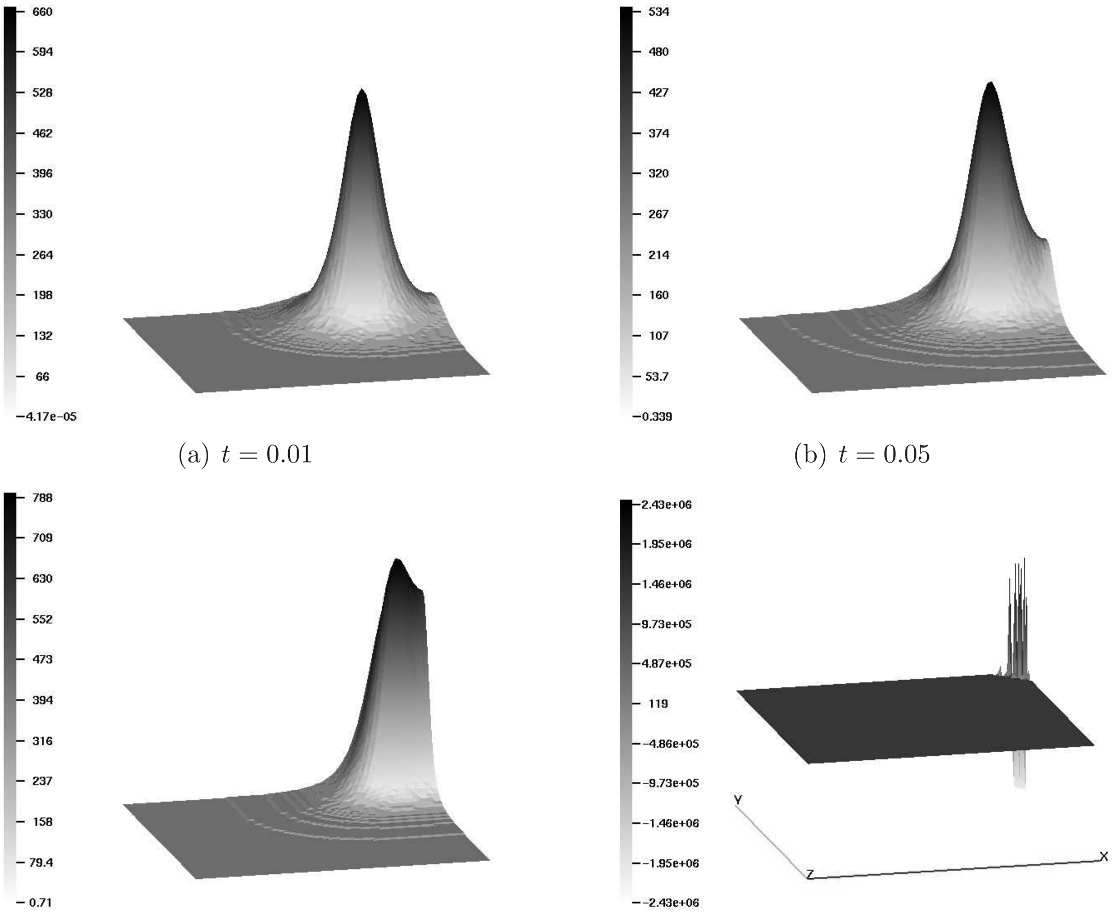
Fig. 3.6. Peak heights variation with mesh refinement.

3.2. Blow-up at the boundary of the domain

In the second example, the system of chemotaxis equations (3.1)–(3.2) is solved subject to the initial conditions

$$\begin{aligned} u_0(x,y) &= 1000\, e^{-100((x-0.75)^2+(y-0.75)^2)}, \\ c_0(x,y) &= 0. \end{aligned} \tag{3.4}$$

Since the initial chemoattractant concentration is zero, the blow-up is expected to occur much later than in the previous example. Therefore, simulations are performed with a larger time step $\Delta t = 10^{-3}$. As time evolves, the solution of system (3.1)–(3.2) assumes a spiky form and moves towards the upper right corner of the domain. The results obtained with the standard Galerkin discretization are displayed in Fig. 3.7. Again, the cell density becomes negative, and nonphysical oscillations are observed in the corner. These problems can be cured using algebraic flux correction of FCT type, as demonstrated by the solutions in Fig. 3.8.



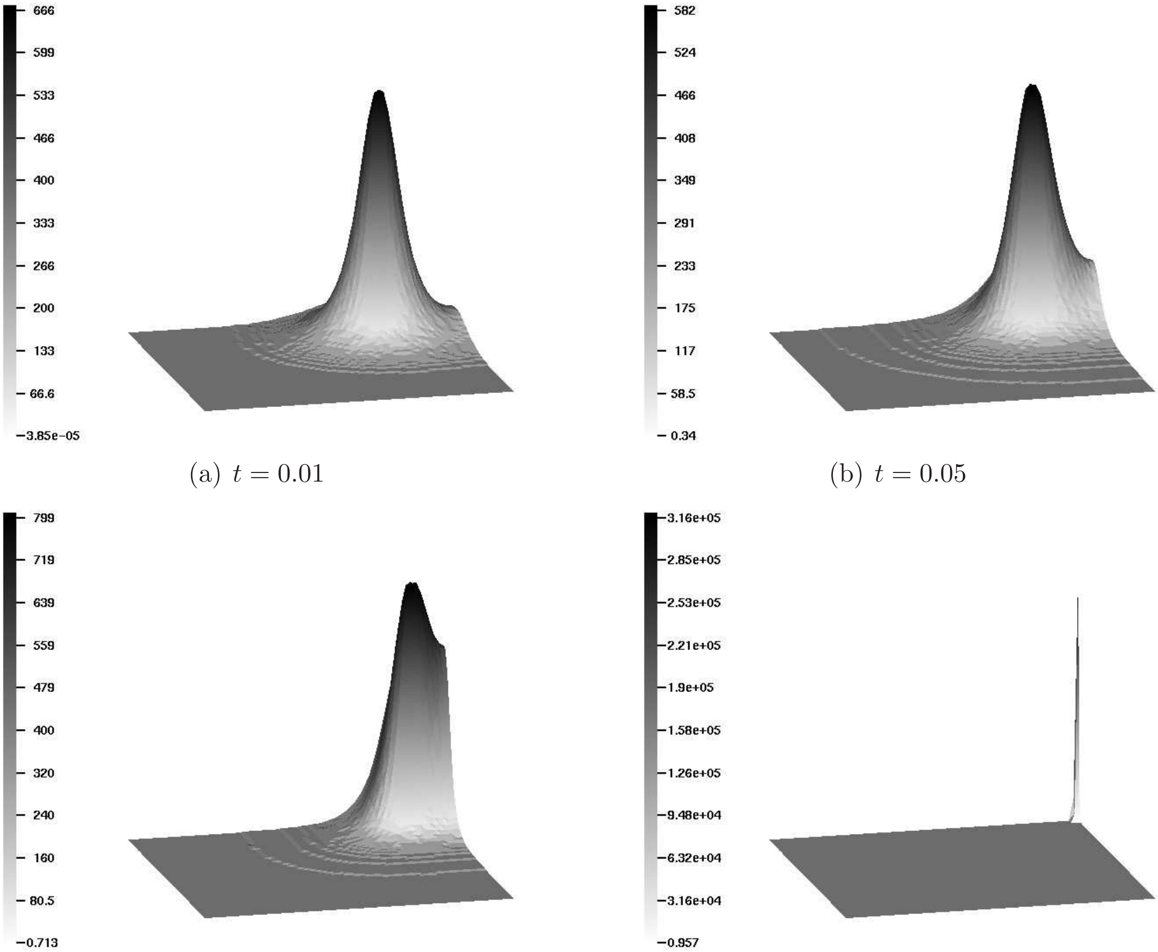
(a) $t = 0.01$

(b) $t = 0.05$

(c) $t = 0.07$

(d) $t = 0.1$

Fig. 3.7. Blow-up in the corner, Galerkin scheme, $h = \frac{1}{128}$, $\Delta t = 10^{-3}$.



(a) $t = 0.01$

(b) $t = 0.05$

(c) $t = 0.07$

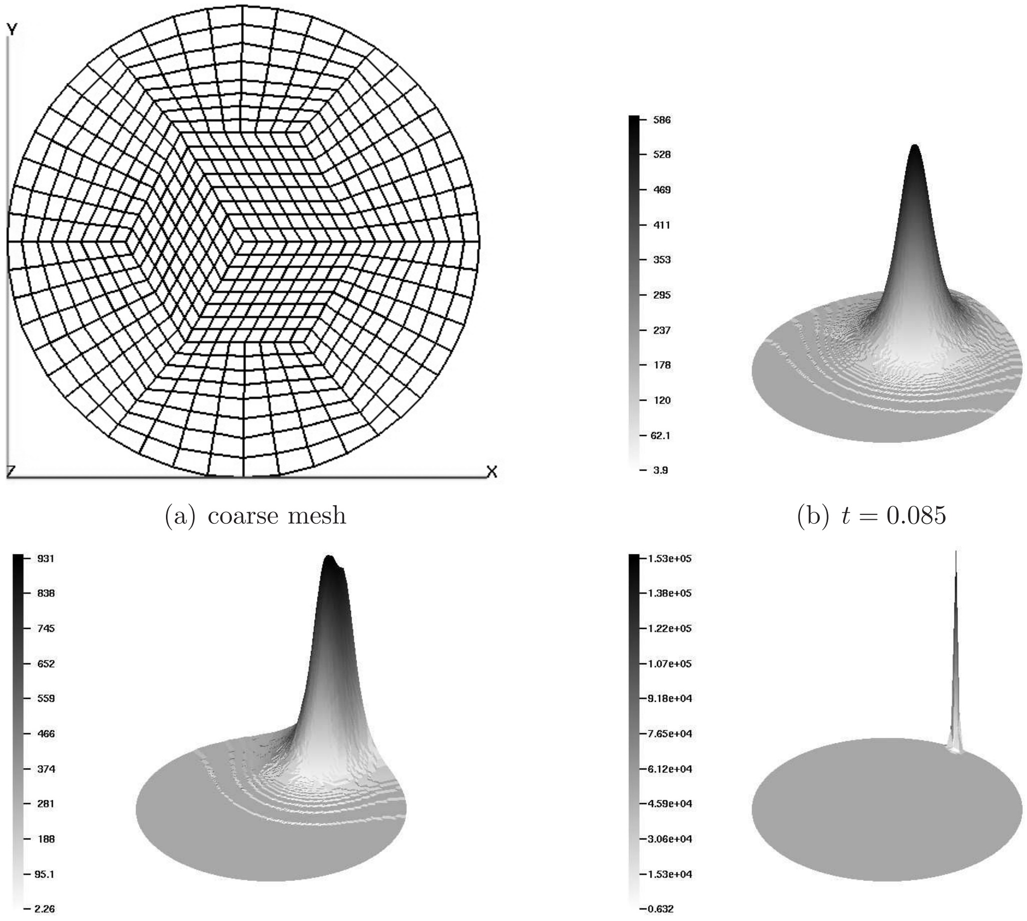
(d) $t = 0.1$

Fig. 3.8. Blow-up in the corner, FEM-FCT scheme, $h = \frac{1}{128}$, $\Delta t = 10^{-3}$.

The point of blow-up may depend on the geometry on the computational domain, as well as on the imposed boundary conditions [11]. For example, let Ω be a circle of radius 0.5 centered at the point $(0.5, 0.5)$. A typical coarse mesh is depicted in Fig. 3.9 (a). The purpose of the numerical experiment to be performed is to find out if the blow-up point tends to any particular location. The peak of the initial profile u_0 is placed at the point $(0.6, 0.6)$

$$\begin{aligned} u_0(x,y) &= 1000\, e^{-100((x-0.6)^2+(y-0.6)^2)}, \\ c_0(x,y) &= 0. \end{aligned} \tag{3.5}$$

All other settings are the same as in the case of the square domain. The FEM-FCT results in Fig. 3.9 (b,c,d) were obtained with 9216 bilinear elements. The distribution of the cell density moves in the radial direction and blows up at the boundary of the circle in finite time.



(a) coarse mesh

(b) $t = 0.085$

(c) $t = 0.14$

(d) $t = 0.2$

Fig. 3.9. Blow-up at a circular boundary, FEM-FCT scheme, $\Delta t = 10^{-3}$.

3.3. Pattern formation

In the last example, we consider a more complicated and realistic chemotaxis model. It describes the complex space-time patterns formed by motile cells of *Escherichia coli.* There are several different approaches to modeling the distribution of these bacteria. One of them leads to the following system of differential equations [5]:

$$u_t = D_1 \Delta u - \alpha \nabla \cdot \left(\frac{u}{(1+c)^2} \nabla c \right), \tag{3.6}$$

$$c_t = D_2 \Delta c + \beta \frac{w\, u^2}{\sigma + u^2}. \tag{3.7}$$

For theoretical analysis, numerical algorithms, and simulation results we refer to [7, 16, 27].

In another model, proposed by Mimura and Tsujikawa [21], only the diffusion, the chemotaxis, and the growth of bacteria are taken into account. The corresponding PDE system reads

$$u_t = D_1 \Delta u - \chi \nabla \cdot (u \nabla c) + u^2(1-u), \tag{3.8}$$

$$c_t = \Delta c - \beta c + u. \tag{3.9}$$

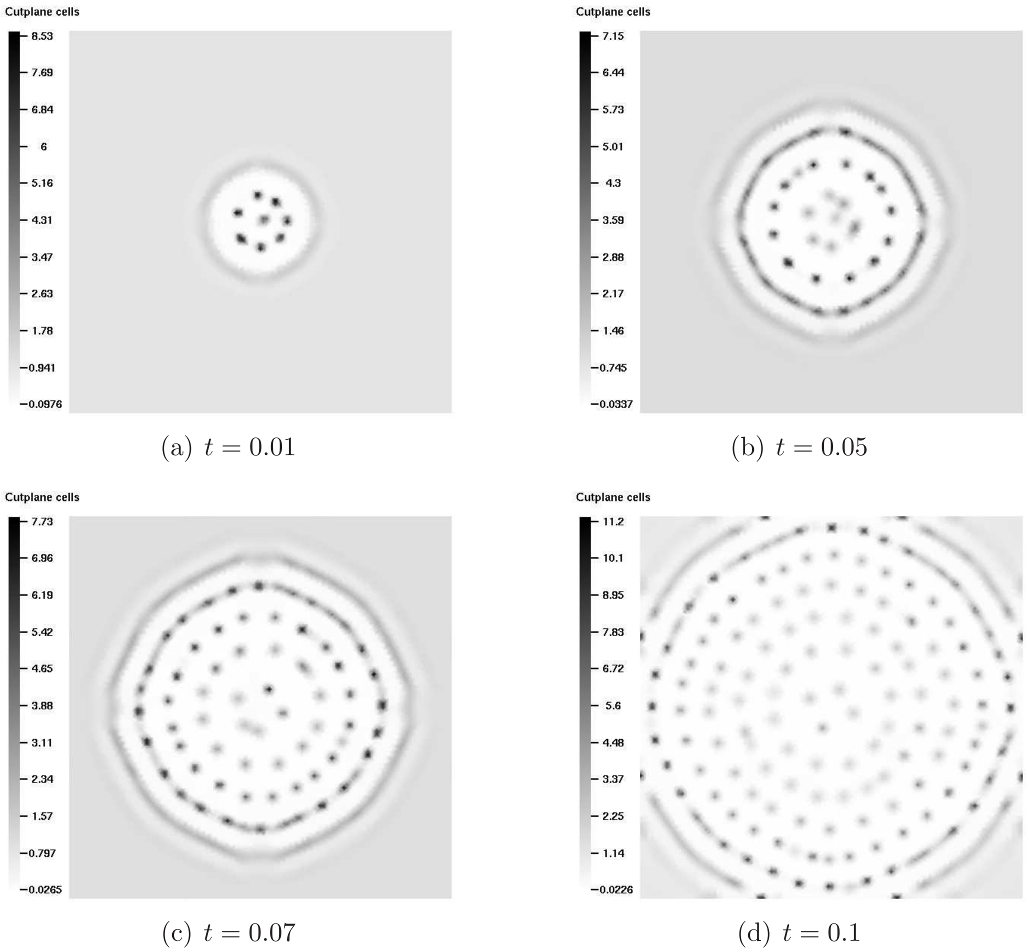
For a detailed presentation of this approach see, e.g., [1, 2]. Although both systems (3.6)–(3.7) and (3.8)–(3.9) model the space-time patterns formed by motile cells of *Escherichia coli* and fit the structure of (1.1)–(1.2), in this article we consider only the Mimura-Tsujikawa model (3.8)–(3.9) with $D_1 = 0.0625$, $\chi = 8.5$, and $\beta = 32$. These parameter settings are taken from [1, 2]. The initial conditions are given by

$$\begin{aligned} u_0(x,y) &= 1 + \sigma(x,y), \\ c_0(x,y) &= 1/32, \end{aligned}$$

where $\sigma(x,y)$ is a small perturbation defined as

$$\sigma(x,y) = \begin{cases} \text{random}, & \text{if} \quad \|\boldsymbol{x} - (8,8)^T\| \leqslant 1.5, \\ 0, & \text{otherwise.} \end{cases}$$

Numerical simulations are performed in the square domain $\Omega = (0,16)^2$ discretized using a uniform mesh of conforming bilinear finite elements. The employed mesh size $h = 1/8$ corresponds to 16384 cells. The time step is taken to be $\Delta t = 0.1$. The solutions are very sensitive to the choice of parameters, especially χ, σ, etc. Figure 3.10 illustrates the temporal evolution of the cell distribution obtained with the implicit FEM-FCT algorithm. The presented results are in quantitative agreement with those reported in [1, 2]. The same formation patterns have been observed experimentally [5].



(a) $t = 0.01$

(b) $t = 0.05$

(c) $t = 0.07$

(d) $t = 0.1$

Fig. 3.10. Pattern formation simulated with the FEM-FCT algorithm, $\Delta t = 0.1, h = \frac{1}{8}$.

4. Conclusions

An implicit flux-corrected transport algorithm has been developed for the unified form (1.1)–(1.2) of chemotaxis models. Positivity constraints were enforced using a nonlinear blend of high- and low-order approximations. The employed limiting strategy is fully multidimensional and applicable to (multi-)linear finite element discretizations on unstructured meshes. The resultant scheme satisfies the discrete maximum principle and resolves steep gradients without excessive smearing. The local order of accuracy varies between first (low-order solution) and second (high-order solution), depending on the amount of artificial diffusion retained at the flux correction step. The robustness and efficiency of the linearized FEM-FCT algorithm make it an attractive alternative to other stabilization techniques for the chemotaxis problems proposed in the literature [7, 9, 10, 16, 23, 28].

A preliminary numerical study of the implicit FEM-FCT scheme has been performed for the minimal Keller-Segel model. The flux-corrected Galerkin approximation has been shown to be sufficiently accurate and positivity-preserving, even in the case of solutions with sharp peaks that blow-up in the center or at the boundary of the domain. An example that illustrates the benefits of local mesh refinement was included. Furthermore, realistic simulation results were obtained for a representative model of chemotactical pattern formation. The proposed methodology is suitable for a 3D implementation and seems to be a promising approach to the numerical treatment of real-life chemotaxis problems in medicine and biology. Further research will concentrate on the design of FCT algorithms for (1.1)–(1.2) with stronger coupling. The implications of the time-stepping method also call for a detailed investigation. Last but not least, a detailed quantitative comparison with existing numerical results [7, 9, 10, 25] is required to illustrate the pros and cons of different discretization/stabilization techniques.

Acknowledgments. This research was supported by the German Research Foundation, grant KU 1530/3.

References

1. M. Aida, T. Tsujikawa, M. Efendiev, A. Yagi, and M. Mimura, *Lower estimate of the attractor dimension for a chemotaxis growth system*, Journal of the London Mathematical Society, **74** (2006), no. 2, pp. 453–474.

2. M. Aida and A. Yagi, *Target pattern solutions for chemotaxis-growth system*, Scientiae Mathematicae Japonicae, **59** (2004), no. 3, pp. 577–590.

3. A. Bonami, D. Hilhorst, E. Logak, and M. Mimura, *Singular limit of chemotaxis-growth model*, Adv. Differential Equations, **6** (2001), pp. 1173–1218.

4. J. P. Boris and D. L. Book, *Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works*, J. Comput. Phys., **11** (1973), pp. 38–69.

5. E. O. Budrene and H. C. Berg, *Dynamics of formation of symmetrical patterns by chemotactic bacteria*, Nature, **376** (1995), no. 6535, pp. 49–53.

6. M. Burger, M. Di Francesco, and Y. Dolak-Stru, *The Keller-Segel model for chemotaxis with prevention of overcrowding: Linear vs. nonlinear diffusion*, SIAM J. Math. Anal., **38** (2006), pp. 1288–1315.

7. A. Chertock and A. Kurganov, *A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models*, Numer. Math., **111** (2008), pp. 169–205.

8. Y. Dolak and C. Schmeiser, *The Keller-Segel model with logistic sensitivity function and small diffusivity*, SIAM J. Appl. Math., **66** (2005), pp. 595–615.

9. Y. Epshteyn, *Discontinuous Galerkin methods for the chemotaxis and haptotaxis models*, J. Comput. Appl. Math., **224** (2009), no. 1, pp. 168–181.

10. Y. Epshteyn and A. Kurganov, *New interior penalty discontinuous Galerkin methods for the Keller-Segel chemotaxis model*, SIAM J. Numer. Anal., **47** (2008), no. 1, pp. 386–408.

11. F. Filbet, *A finite volume scheme for the Patlak-Keller-Segel chemotaxis model*, Numer. Math., **104** (2006), no. 4, pp. 457–488.

12. M. A. Herrero and J. J. L. Velazquez, *A blow-up mechanism for a chemotaxis model*, Ann. Sc. Norm. Super., **24** (1997), pp. 633–683.

13. T. Hillen and K. J. Painter, *A user's guide to PDE models for chemotaxis*, J. Math. Biol., **58** (2009), no. 1, pp. 183–217.

14. E. F. Keller and E. F. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theor. Biol., **26** (1970), pp. 399–415.

15. E. F. Keller and E. F. Segel, *Model for chemotaxis*, J. Theor. Biol., **30** (1971), pp. 225–234.

16. B. S. Kirk and G. F. Carey, *A parallel, adaptive finite element scheme for modeling chemotactic biological systems*, Commun. Numer. Meth. Engrg.,(in press).

17. D. Kuzmin, *Explicit and implicit FEM-FCT algorithms with flux linearization*, J. Comput. Phys., **228** (2009), pp. 2517–2534.

18. D. Kuzmin and M. Möller, *Algebraic flux correction I. Scalar conservation laws*, in: D. Kuzmin, R. Löhner, S. Turek (Eds.), *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Springer, Berlin, (2005), pp. 155–206.

19. D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., **175** (2002), pp. 525–558.

20. I. R. Lapidus and R. Schiller, *Model for the chemotactic response of a bacterial population*, Biophys J., **16** (1976), no. 7, pp. 779–789.

21. M. Mimura and T. Tsujikawa, *Aggregating pattern dynamics in a chemotaxis model including growth*, Physica A, **230** (1996),pp. 499–543.

22. A. B. Potapov and T. Hillen, *Metastability in chemotaxis models*, J. Dyn. Diff. Eq., **17** (2005), pp. 293–330.

23. D. L. Ropp and J. N. Shadid, *Stability of operator splitting methods for systems with indefinite operators: Advection-diffusion-reaction systems*, J. Comput. Phys., (in press).

24. H. R. Schwetlick, *Travelling fronts for multidimensional nonlinear transport equations*, Analyse non linéaire, **17** (2000), no. 4, pp. 523–550.

25. N. Saito, *Conservative upwind finite-element method for a simplified Keller-Segel system modelling chemotaxis*, IMA J. Numer. Anal., **27** (2007), pp. 332–365.

26. T. Suzuki, *Free energy and self-interacting particles*, Boston: Birkhauser, 2005.

27. R. Tyson, S. R. Lubkin, and J. D. Murray, *A minimal mechanism for bacteria pattern formation*, Proc. Biol. Sci., **266** (1999), no. 1416, pp. 299–304.

28. R. Tyson, L. G. Stern, and R. J. LeVeque, *Fractional step methods applied to a chemotaxis model*, J. Math. Biol., **41** (1996), pp. 455–475.

29. S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., **31** (1979), pp. 335–362.