LECTURE NOTES IN COMPUTATIONAL SCIENCE AND ENGINEERING 120

Zhongyi Huang · Martin Stynes
Zhimin Zhang *Editors*

Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016

Springer

Lecture Notes in Computational Science and Engineering 120

Editors:

More information about this series at http://www.springer.com/series/3527

Zhongyi Huang • Martin Stynes • Zhimin Zhang
Editors

Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016

Springer

Editors
Zhongyi Huang
Dept. of Mathematical Sciences
Tsinghua University
Beijing, China

Martin Stynes
Applied and Computational Mathematics
Beijing Computational Science Research Center
Beijing, China

Zhimin Zhang
Applied and Computational Mathematics
Beijing Computational Science Research Center
Beijing, China

ISSN 1439-7358 ISSN 2197-7100 (electronic)
Lecture Notes in Computational Science and Engineering
ISBN 978-3-319-67201-4 ISBN 978-3-319-67202-1 (eBook)
https://doi.org/10.1007/978-3-319-67202-1

Library of Congress Control Number: 2017957058

Mathematics Subject Classification (2010): 65-06, 65L11, 65Mxx, 65Nxx, 76-06

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

These Proceedings contain papers associated with a selection of the lectures given at the conference BAIL 2016: Boundary and Interior Layers—Computational and Asymptotic Methods, which was held during 15–19 August 2016 at the Beijing Computational Science Research Centre (CSRC) and Tsinghua University, Beijing, China. The 60 participants came from Chile, China, the Czech Republic, Germany, India, Ireland, the Netherlands, New Zealand, Russia, Serbia, Spain, and the USA.

The BAIL series of conferences were started by Professor John Miller, who organised the first three in Dublin in 1980, 1982, and 1984. Subsequent conferences were then held in Novosibirsk (1986), Shanghai (1988), Copper Mountain, Colorado (1992), Beijing (1994), Perth (2002), Toulouse (2004), Göttingen (2006), Limerick (2008), Zaragoza (2010), Pohang (2012), and Prague (2014). The next BAIL Conference will be in Glasgow, UK, in 2018.

The BAIL conferences aim to bring together mathematicians and engineers/physicists whose research involves layer phenomena, and these Proceedings reflect this desire. Their papers involve both modelling and numerical methods and their analysis, and will demonstrate to the reader the current state of the art in the computation of boundary and interior layer phenomena.

All papers in the Proceedings were subjected to a standard refereeing process. The editors wish to thank the authors for their contributions and their cooperation in preparing their work for this volume of LNCSE. We are also grateful to the anonymous referees for their valuable work, without which it would have been impossible to produce this publication.

Finally, we thank Beijing Computational Science Research Center and Tsinghua University for their support for the conference. A particular thanks to Dr. Jeanne Stynes, and Ms. Sining Wang and Dr. Xiangyun Meng of CSRC, who were hugely helpful in the organisation and smooth running of the conference.

Beijing, China — Zhongyi Huang
Beijing, China — Martin Stynes
Beijing, China — Zhimin Zhang

Contents

Error Estimates in Balanced Norms of Finite Element Methods on Layer-Adapted Meshes for Second Order Reaction-Diffusion Problems

Hans-G. Roos

Abstract Error estimates of finite element methods for reaction-diffusion problems are often realized in the related energy norm. In the singularly perturbed case, however, this norm is not adequate. A different scaling of the H^1 seminorm leads to a balanced norm which reflects the layer behavior correctly. We discuss anisotropic problems, semilinear equations, supercloseness and a combination technique. Moreover, we consider different classes of layer-adapted meshes and sketch the three-dimensional case. Remarks to systems and problems with different layers close the paper.

AMS subject classification: 65 N

1 Introduction

We shall examine the finite element method for the numerical solution of the singularly perturbed linear elliptic boundary value problem

$$Lu \equiv -\varepsilon \Delta u + cu = f \qquad \text{in } \Omega = (0,1) \times (0,1) \tag{1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{1.1b}$$

where $0 < \varepsilon \ll 1$ is a small positive parameter, $c > 0$ is (for simplicity) a positive constant and f is sufficiently smooth.

It is well-known that the problem has a unique solution $u \in V = H_0^1(\Omega)$ which satisfies the stability estimate in the standard energy norm

$$\|u\|_\varepsilon := \varepsilon^{1/2}|u|_1 + \|u\|_0 \preceq \|f\|_0. \tag{1.2}$$

H.-G. Roos (✉)
Institut Numerical Mathematics, Technische Universität Dresden, 01062 Dresden, Germany
e-mail: hans-goerg.roos@tu-dresden.de

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_1

Here we used the following notation: if $A \preceq B$, there exists a (generic) constant C independent of ε (and later also of the mesh used) such that $A \leq C\,B$. Moreover for $D \subset \Omega$ we denote by $\|\cdot\|_{0,D}$, $\|\cdot\|_{\infty,D}$ and $|\cdot|_{1,D}$ the standard norms in $L_2(D)$, $L_\infty(D)$ and the standard seminorm in $H^1(D)$, respectively. We shall omit the notation of the domain in the case $D = \Omega$. Similarly, we want to use the notation $(\cdot,\cdot)_D$ for the inner product in $L_2(D)$ and abbreviate $(\cdot,\cdot)_\Omega$ to $(\cdot,\cdot)$.

The error of a finite element approximation $u^N \in V^N \subset V$ satisfies

$$\|u - u^N\|_\varepsilon \preceq \min_{v^N \in V^N} \|u - v^N\|_\varepsilon. \tag{1.3}$$

When linear or bilinear elements are used on a Shishkin mesh (see Sect. 2), one can prove under certain additional assumptions concerning f for the interpolation error of the Lagrange interpolant $u^I \in V^N$

$$\|u - u^I\|_\varepsilon \preceq \left(\varepsilon^{1/4} N^{-1} \ln N + N^{-2}\right) \tag{1.4}$$

(see [23] or [31]). It follows that the error $u - u^N$ also satisfies such an estimate.

However, the typical boundary layer function $\exp(-x/\varepsilon^{1/2})$ measured in the norm $\|\cdot\|_\varepsilon$ is of order $\mathcal{O}(\varepsilon^{1/4})$. Consequently, error estimates in this norm are less valuable than for convection- diffusion equations where the layers are of the structure $\exp(-x/\varepsilon)$. Wherefore we ask the fundamental question:
Is it possible to prove error estimates in the balanced norm

$$\|v\|_b := \varepsilon^{1/4}|v|_1 + \|v\|_0 \quad ? \tag{1.5}$$

In Sect. 2 we will repeat a basic idea to prove error estimates in a balanced norm and extend the approach to semilinear problems and anisotropic equations. Most of the manuscript is focused on Shishkin meshes. Different classes of layer-adapted meshes are presented in Sect. 3, moreover we demonstrate the situation in the three-dimensional case. Supercloseness and a combination technique are discussed in Sect. 4. Finally, we present a direct mixed method in Sect. 5 and sketch some open problems in Sect. 6.

We restrict ourselves to second order problems, for fourth-order problems see [12]. For the *hp*-FEM on spectral boundary layer meshes we refer to [23, 24].

2 The Basic Error Estimate in a Balanced Norm and Some Extensions

2.1 Linear Problems

The mesh Ω^N used is the tensor product of two one-dimensional piecewise uniform Shishkin meshes. I.e., $\Omega^N = \Omega_x \times \Omega_y$, where Ω_x (analogously Ω_y) splits $[0, 1]$ into the subintervals $[0, \lambda_x]$, $[\lambda_x, 1 - \lambda_x]$ and $[1 - \lambda_x, 1]$. The mesh distributes $N/4$ points

equidistantly within each of the subintervals $[0, \lambda_x]$, $[1-\lambda_x, 1]$ and the remaining points within the third subinterval. For simplicity, assume

$$\lambda = \lambda_x = \lambda_y = \min\{1/4, \lambda_0\sqrt{\varepsilon/c^*}\ln N\} \quad \text{with } \lambda_0 = 2 \text{ and } c^* < c.$$

We remark that the choice of λ_0 mainly depends on the polynomial degree of the finite element space. We use for the step sizes

$$h := \frac{4\lambda}{N} \quad \text{and} \quad H := \frac{2(1-2\lambda)}{N}.$$

Let $V^N \subset H_0^1(\Omega)$ be the space of bilinear finite elements on Ω^N or the space of linear elements over a triangulation obtained from Ω^N by drawing diagonals.

A standard formulation of problem (1.1) reads: Find $u \in V$, such that

$$\varepsilon(\nabla u, \nabla v) + c(u, v) = (f, v) \quad \forall v \in V. \tag{2.1}$$

By replacing V in (2.1) with V^N one obtains a standard discretization that yields the FEM-solution u^N.

As we mentioned already in the Introduction, certain assumptions on f allow a decomposition of u into smooth components S and layer terms E such that the following estimates for the interpolation error of the Lagrange interpolant hold true (see [9, 25] or [31]):

$$\|u - u^I\|_0 \preceq N^{-2}, \quad \varepsilon^{1/4}|u - u^I|_1 \preceq N^{-1}\ln N \tag{2.2}$$

and

$$\|u - u^I\|_{\infty,\Omega_0} \preceq N^{-2}, \quad \|u - u^I\|_{\infty,\Omega\setminus\Omega_0} \preceq (N^{-1}\ln N)^2, \tag{2.3}$$

here $\Omega_0 = (\lambda_x, 1-\lambda_x) \times (\lambda_y, 1-\lambda_y)$. Let us also introduce $\Omega_f := \Omega \setminus \Omega_0$.

Instead of the Lagrange interpolant we use in our error analysis the L_2 projection $\pi u \in V^N$ from u. Based on

$$u - u^N = u - \pi u + \pi u - u^N$$

we estimate $\xi := \pi u - u^N$:

$$\|\xi\|_\varepsilon^2 \preceq \varepsilon|\nabla\xi|_1^2 + c\,\|\xi\|_0^2 = \varepsilon(\nabla(\pi u - u), \nabla\xi) + c\,(\pi u - u, \xi).$$

Because $(\pi u - u, \xi) = 0$, it follows

$$|\pi u - u^N|_1 \preceq |u - \pi u|_1. \tag{2.4}$$

If we now could prove a similar estimate as (2.2) for the error of the L_2 projection, we obtain an estimate in the balanced norm because we have already the estimate $\|u - u_N\|_0 \preceq N^{-2}$ from the analysis that leads to (1.4).

Lemma 1 *Assuming the validity of* (2.2) *and* (2.3)*, the error of the L_2 projection on the Shishkin mesh satisfies*

$$\|u - \pi u\|_\infty \preceq \|u - u^I\|_\infty, \quad \varepsilon^{1/4}|u - \pi u|_1 \preceq N^{-1}(\ln N)^{3/2}. \tag{2.5}$$

The proof uses the L_∞ stability of the L_2 projection on our mesh [25]. Inverse inequalities are used to move from estimates in W^1_∞ to L_∞, for details see [30].

From (2.4) and Lemma 1 we get

Theorem 1 *Assuming* (2.2) *and* (2.3)*, the error of the Galerkin finite element method with linear or bilinear elements on a Shishkin mesh satisfies*

$$\|u - u^N\|_b \preceq N^{-1}(\ln N)^{3/2} + N^{-2}. \tag{2.6}$$

With a more sophisticated choice of the projection the factor $(\ln N)^{3/2}$ can be replaced by $\ln N$. We remark that for Q_k elements with $k > 1$ one can get an analogous result

$$\|u - u^N\|_b \preceq N^{-k}(\ln N)^{k+1/2} + N^{-(k+1)}$$

because on tensor product meshes the L_2 projection is as well L_∞ stable (see [8] for the one-dimensional result on arbitrary meshes, on tensor product meshes the statement follows immediately).

2.2 Semilinear Problems

It is easy to modify the basic idea to the singularly perturbed semilinear elliptic boundary value problem

$$Lu \equiv -\varepsilon\Delta u + g(\cdot, u) = 0 \qquad \text{in } \Omega = (0,1)\times(0,1) \tag{2.7a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{2.7b}$$

We assume that g is sufficiently smooth and $\partial_2 g \geq \mu > 0$. Then, the so- called reduced problem and our given problem have a unique solution.

If $\partial\Omega$ is smooth, the solution is characterized by the typical boundary layer for linear reaction-diffusion problems, see [14] for the semilinear case. If corners exist, additionally corner layers arise, see [15] for semilinear problems in a polygonal domain. For the analysis of finite element methods on layer-adapted meshes we need a solution decomposition (see Remark 1.27 in Chap. 3 of [31]), in the semilinear

case sufficient conditions for the existence of such a decomposition are not known. Therefore we just assume the existence of a solution decomposition.

A standard weak formulation of our semilinear problem reads: Find $u \in V$, such that

$$\varepsilon(\nabla u, \nabla v) + (g(\cdot, u), v) = 0 \quad \forall v \in V. \tag{2.8}$$

By replacing V in (2.1) with V^N one obtains a standard discretization that yields the FEM solution u^N.

If $\pi u \in V^N$ is some projection of u, we decompose the error into

$$u - u^N = u - \pi u + \pi u - u^N$$

and (assuming we can control the projection error) start the error analysis from the following relation for $\xi := \pi u - u^N$:

$$\begin{aligned} \varepsilon|\nabla\xi|_1^2 + \mu\,\|\xi\|_0^2 &\le \varepsilon(\nabla\xi, \nabla\xi) + (g(\cdot, \pi u) - g(\cdot, u^N), \xi) \\ &= \varepsilon(\nabla(\pi u - u), \nabla\xi) + (g(\cdot, \pi u) - g(\cdot, u), \xi). \end{aligned}$$

If we choose πu to be the standard interpolant of u, the usual error estimate in the energy norm follows:

$$\|u - u^N\|_\varepsilon \preceq \left(\varepsilon^{1/4} N^{-1} \ln N + N^{-2}\right) \tag{2.9}$$

But again we want to prove an error estimate in the balanced norm

$$\|v\|_b := \varepsilon^{1/4}|v|_1 + \|v\|_0 \quad . \tag{2.10}$$

Following the basic idea from [30], we define πu by

$$(g(\cdot, \pi u), v) = (g(\cdot, u), v) \quad \text{for all } v \in V^N. \tag{2.11}$$

Our assumption $\partial_2 g \ge \mu > 0$ immediately tells us that πu is well defined and, moreover,

$$\|u - \pi u\|_0 \preceq \inf_{v^N \in V^N} \|u - v^N\|_0. \tag{2.12}$$

It follows from the definition of our projection that

$$|\pi u - u^N|_1 \preceq |u - \pi u|_1. \tag{2.13}$$

For the standard interpolant u^I of u we have

$$\varepsilon^{1/4}|u - u^I|_1 \preceq N^{-1} \ln N.$$

If we now could prove a similar estimate for our projection error, we would obtain an estimate in the balanced norm because we have already an estimate for $\|u-u_N\|_0$ in (2.9).

Lemma 2 *The projection defined by* (2.11) *is* L_∞ *stable.*

Proof The proof is based on Taylor's formula

$$F(w) - F(v) = (\int_0^1 DF(v + s(w - v))ds)(w - v).$$

Introducing the linear operator

$$\triangle F(v, w) := \int_0^1 DF(v + s(w - v))ds$$

it is obvious that

$$\|w - v\| \leq \|(\triangle F(v, w))^{-1}\| \, \|F(w) - F(v)\|.$$

Therefore, the L_∞ stability of the L_2 projection on our mesh [25] implies the L_∞ stability of our generalized projection as well.

Lemma 3 *The projection error of* (2.11) *on the Shishkin mesh satisfies*

$$\|u - \pi u\|_\infty \preceq \|u - u^I\|_\infty, \quad \varepsilon^{1/4}|u - \pi u|_1 \preceq N^{-1}(\ln N)^{3/2}. \tag{2.14}$$

The proof works analogously as in the linear case. And, consequently, we get the same error estimate as in Theorem 1 also in the semilinear case.

2.3 An Anisotropic Diffusion Problem

Next we consider the anisotropic problem

$$-\varepsilon u_{xx} + u_{yy} + cu = f \qquad \text{in } \Omega = (0, 1) \times (0, 1) \tag{2.15a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{2.15b}$$

Now we have only boundary layers at $x = 0$ and $x = 1$, the layers are of elliptic type. But the layer terms satisfy the same estimates as in the reaction-diffusion regime [16, 17]. Therefore, the estimates (2.2) and (2.3) for the interpolation error on the related Shishkin mesh remain valid, of course, now $\Omega_0 = (\lambda_x, 1 - \lambda_x) \times (0, 1)$. Therefore, defining the energy norm by

$$\|v\|_{\varepsilon,a} := \varepsilon^{1/2}\|u_x\|_0 + \|u_y\|_0 + \|u\|_0$$

it follows for bilinear elements

$$\|u - u^N\|_{\varepsilon,a} \preceq \left(\varepsilon^{1/4} N^{-1} \ln N + N^{-1} + N^{-2}\right).$$

If we want to estimate the error in the balanced norm

$$\|v\|_{b,a} := \varepsilon^{1/4}\|u_x\|_0 + \|u_y\|_0 + \|u\|_0,$$

we start for $\xi := \pi u - u^N$ from

$$\varepsilon\|\xi_x\|_0^2 \le \varepsilon((\pi u - u)_x, \xi_x) + ((\pi u - u)_y, \xi_y) + c\,(\pi u - u, \xi).$$

Now we define in the anisotropic case the projection onto the finite element space by

$$((\pi u - u)_y, \xi_y) + c\,(\pi u - u, \xi) = 0 \quad \forall \xi \in V^N.$$

Consequently it remains to estimate for that projection $\|(\pi u - u)_x\|_0$. But the projection satisfies

$$\pi v = \pi^y(\pi^x v),$$

where π^x is the one-dimensional L_2 projection and π^y the one-dimensional Ritz projection (with respect to a non-singularly perturbed operator on a standard mesh), compare [11]. Consequently, the projection is L_∞ stable and we can repeat our basic idea to prove estimates in the balanced norm.

Remark that in [11] this idea was used to analyse the SDFEM technique for a convection-diffusion problem with two different boundary layers, an exponential layer and a characteristic layer.

3 The 3D Case and Different Classes of Layer-Adapted Meshes

3.1 The 3D Case

In the 3D case with $\Omega = (0,1)^3$ new difficulties arise. Shishkin meshes are anisotropic meshes, therefore, anisotropic interpolation error estimates are needed. In 2D the bilinear interpolant satisfies on a rectangle K with step sizes h_1, h_2

$$\|v - v^I\|_{0,p,K} \le C \sum_{|\alpha|=m} h^\alpha \|D^\alpha v\|_{0,p,K} \qquad \text{for } m = 1,2 \qquad \text{and} \tag{3.1a}$$

$$\|\partial_x(v - v^I)\|_{0,p,K} \le C \sum_{|\alpha|=1} h^\alpha \|D^\alpha \partial_x v\|_{0,p,K} \tag{3.1b}$$

for $1 \leq p \leq \infty$. These estimates are needed mostly for $p = 2$, for instance, to estimate $\varepsilon(\nabla(u - u^I), \nabla v^N)$ using Cauchy-Schwarz.

But in 3D the second estimate of (3.1) does not hold for $p = 2$! (see [1, 3, 4, 10]) In this case the constant is of order

$$C = C(p) \approx \frac{c}{(p-2)^{p/2}}. \tag{3.2}$$

Moreover it is known that alternatively one can assume more smoothness than H^2 [10] or use different interpolants on locally uniform meshes.

The contribution to the error of the smooth part of the solution and of the layer components in the interior domain (where the layer components are small) can be estimated as in the two-dimensional case.

Let E be some layer component. We wish to estimate $\varepsilon(\nabla(E - E^I), \nabla v^N))$ in that subdomain Ω_f, where E is not small and anisotropic elements occur. Instead of Cauchy-Schwarz we use the Hölder inequality ($\frac{1}{p} + \frac{1}{q} = 1$)

$$\varepsilon|(\nabla(E - E^I), \nabla v^N))_{\Omega_f}| \leq \varepsilon|E - E^I|_{1,p,\Omega_f}|v^N|_{1,q,\Omega_f}. \tag{3.3}$$

For $p > 2$ we can now apply (3.1) and obtain

$$\varepsilon|\nabla(E - E^I)|_{0,p,\Omega_f} \preceq \varepsilon C(p)h|E|_{2,p} \preceq C(p)N^{-1}\ln N\varepsilon^{1/2+1/(2p)}. \tag{3.4}$$

Using $\text{meas}(\Omega_f) \preceq \varepsilon^{1/2}\ln N$ we get

$$|v^N|_{1,q,\Omega_f} \preceq \varepsilon^{1/4-1/(2p)}(\ln N)^{1/2-1/p}|v^N|_{1,2,\Omega_f}. \tag{3.5}$$

Summarizing the estimate for the crucial term the situation in 3D is not much worse than in 2D: For arbitrary $p > 2$ we have

$$\varepsilon|(\nabla(E - E^I), \nabla v^N)_{\Omega_f}| \leq \tilde{C}(p)\varepsilon^{1/4}(\ln N)^{1/2-1/p}N^{-1}\ln N\|v^N\|_\varepsilon. \tag{3.6}$$

Consequently one obtains in the energy norm for the reaction-diffusion problem in 3D

$$\|u - u^N\|_\varepsilon \leq \hat{C}(p)\varepsilon^{1/4}(\ln N)^{1/2-1/p}N^{-1}\ln N + N^{-2}.$$

This gives us also an estimate for the L_2 part of the balanced norm.

To estimate $\varepsilon^{1/4}|u - u^N|_1$, we just follow [30]. All ingredients used are also available in 3D: the stability properties of the L_2 projection and the interpolation error estimates in L_2 and L_∞.

3.2 *Different Classes of Layer-Adapted Meshes*

So far error estimates in balanced norms are only known for Shishkin meshes [19, 27, 30] and spectral boundary layer meshes [24]. In the following we discuss again the two-dimensional reaction-diffusion problem

$$Lu \equiv -\varepsilon \Delta u + cu = f \qquad \text{in } \Omega = (0,1) \times (0,1) \tag{3.7a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{3.7b}$$

and its discretization with bilinear or linear elements on a mesh of tensor-product type.

Shishkin-type meshes introduced in [28] use the same transition point(s) from the fine to a coarse mesh as the original Shishkin mesh but the fine mesh is graded. Let ϕ be the mesh generating function and ψ defined by $\phi = -\ln \psi$. Then, under some standard assumptions, especially

$$\max \phi' \preceq N, \tag{3.8}$$

one has in the energy norm

$$\|u - u^N\|_\varepsilon \preceq \varepsilon^{1/4} N^{-1} \max \psi' + (N^{-1} \max \psi')^2.$$

For the Bakhvalov-Shishkin mesh or the Vulanovic-Shishkin mesh $\max \psi'$ is uniformly bounded, thus one gets optimal error estimates. But the energy norm is not balanced.

The approach of [30] leads for S-type meshes to error estimates in the balanced norm too. The only difficulty is the application of an inverse inequality on the fine mesh. Denoting by h_f the minimal step size of the fine mesh, we have

$$|u^I - \pi u|_{1,\Omega_f} \preceq \frac{(\operatorname{meas} \Omega_f)^{1/2}}{h_f} (N^{-1} \max \psi')^2.$$

To get the optimal order with respect to N^{-1}, we need the assumption

$$N^{-1} \preceq \phi(1/N), \tag{3.9}$$

which is satisfied for all meshes mentioned. Then we get in the balanced norm for S-type meshes

$$\|u - u^N\|_b \preceq N^{-1} (\ln N)^{1/2} (\max \psi')^2. \tag{3.10}$$

There exist surprisingly few results concerning finite element methods on *Bakhvalov-type meshes*, see [26, 29].

In [29] Bakhvalov-type meshes are analysed based on their relation to Shishkin-type meshes. It turns out that the analysis in that paper (for convection-diffusion problems) yields for reaction-diffusion problems in the energy norm

$$\|u - u^N\|_\varepsilon \preceq \varepsilon^{1/4} N^{-1} + N^{-2}. \tag{3.11}$$

The ideas from [29] allow also some error estimate in the balanced norm. In the exceptional strip $x \in [x_{N/4-1}, x_{N/4}]$, for instance, the application of an inverse inequality to estimate $|u - \pi u|_1$ generates some additional factor, resulting in

$$\|u - u^N\|_b \preceq Q(N, \varepsilon) N^{-1} (\ln N)^{1/2} \tag{3.12}$$

with

$$Q(N, \varepsilon) := \max(1, N^{-1} (\ln \frac{1}{\varepsilon})^{1/2}).$$

Remark $Q(N, \varepsilon) \leq \sqrt{\ln 10}$ if $N \geq 10$ and $\varepsilon \geq 10^{-100}$.

Recursively generated meshes appear more often in the literature than Bakhvalov-type meshes (combined with finite element methods), let us mention papers by Duran, Franz, Gartland, Liu, Ludwig, Skalicky, Teofanova, Uzelac, Xenophontos and Xu.

In 1D, recursively generated meshes for a problem with a boundary layer characterized by the parameter ε and a layer width of order ε do have the form

$$x_1 = \varepsilon N^{-1}, \tag{3.13a}$$

$$x_i = x_{i-1} + g(\varepsilon, N, x_{i-1}), i = 2, \ldots, M. \tag{3.13b}$$

It makes sense to choose the smallest M such that $M \geq \tau$, here τ is the transition point due to Shishkin to a coarse uniform mesh. For a Gartland-Shishkin mesh we have $g = \varepsilon N^{-1} e^{x_{i-1}/(2\varepsilon)}$, for a Duran-Shishkin mesh the simpler $g = 2N^{-1} x_{i-1}$.

It was shown in [32] as well as in [7], that for a Gartland-Shishkin mesh the grid generation function has the property (3.8) and $\max \psi' \leq C$, moreover $M = O(N)$. Therefore, for a Gartland-Shishkin mesh we get

$$\|u - u^N\|_b \preceq N^{-1} (\ln N)^{1/2}. \tag{3.14}$$

For a Duran-Shishkin mesh the result is a little worse because $\max \psi' \leq C \ln \ln N$ and $M = O(N \ln N)$.

4 Supercloseness and a Combination Technique

We come back to the linear reaction-diffusion problem

$$Lu \equiv -\varepsilon \Delta u + cu = f \qquad \text{in } \Omega = (0,1) \times (0,1) \tag{4.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega \tag{4.1b}$$

for now *bilinear* elements on the corresponding Shishkin mesh. It is well known that we have the supercloseness property (assuming $\lambda_0 \geq 2.5$)

$$\|u^N - u^I\|_\varepsilon \preceq \left(\varepsilon^{1/2}(N^{-1}\ln N)^2 + N^{-2}\right). \tag{4.2}$$

Now we ask: Does there exist some projection onto the finite element space such that a supercloseness property holds with respect to the balanced norm?

With $v_N := u^N - \Pi u$ we start from

$$\varepsilon |v_N|_1^2 + c\,\|v_N\|_0^2 \preceq \varepsilon(\nabla(u - \Pi u), \nabla v_N) + c\,(u - \Pi u, v_N).$$

Next we use the decomposition of solution into a smooth part S and a layer part E with $u = S + E$, decompose also $\Pi u = \Pi S + \Pi E$ (so far ΠS and ΠE are not defined) and use different still to fix projections into our bilinear finite element space for S and E. We choose:

- $\Pi S \in V^N$ satisfies

$$(\Pi S, v) = (S, v) \quad \forall v \in V_0^N$$

  with given values in the grid points on the boundary.
- ΠE is zero in Ω_0 and the standard bilinear interpolation operator in the fine subdomain with exception of one strip of the width of the fine stepsize in the transition region (and, of course, bilinear in that strip and globally continuous)

With this choice we obtain

$$\varepsilon |v_N|_1^2 + c\,\|v_N\|_0^2 \preceq \varepsilon(\nabla(u - \Pi u), \nabla v_N) + c\,(E - \Pi E, v_N)_{\Omega_f}.$$

In the second term we hope to get some extra power of ε, in the first term we want to apply superconvergence techniques for the estimation of the expression $(\nabla(E - \Pi E), \nabla v_N)$. First let us remark that ΠE satisfies the same estimates as the bilinear interpolant E^I on Ω_f:

$$\|E - \Pi E\|_{0,\Omega_f} \preceq \varepsilon^{1/4}(N^{-1}\ln N)^2$$

and (based on Lin identities)

$$\varepsilon|(\nabla(E-\Pi E),\nabla v_N)| \preceq N^{-2}\varepsilon^{3/4}|v_N|_1.$$

It is only a technical question to prove that for our modified interpolant on the exceptional strip the same estimates for the interpolation error hold true as for the bilinear interpolant. Here we use the fact that E is on that strip as small as we want and that the measure of the strip is small as well.

Consequently we get

$$|v_N|_1^2 \preceq |S-\Pi S|_1^2 + \varepsilon^{-1/2}(N^{-1}\ln N)^4.$$

For the L_2 projection of S we have $\|S-\Pi S\|_\infty \preceq N^{-2}$ and $\|S-\Pi S\|_{\infty,\Omega_f} \preceq (\varepsilon^{1/2}N^{-1}\ln N)^2$. It follows

$$|S-\Pi S|_{1,\Omega_0} \preceq N^{-1}, \quad |S-\Pi S|_{1,\Omega_f} \preceq \varepsilon^{1/2}N^{-1}\ln N.$$

Summarizing we get a weak supercloseness result

$$\varepsilon^{1/4}|u^N-\Pi u|_1 \preceq \varepsilon^{1/4}N^{-1} + (N^{-1}\ln N)^2.$$

The result is not satisfactory, but so far we see no possibility to improve it. It is no problem to estimate the L_2 error.

Next we present an application of the supercloseness result to the combination technique. It is clear that the result cannot be optimal because the supercloseness result is nor optimal so far. We analyse the version of the combination technique presented in [13], for a different sparse grid method see [21]. We also remark that in [22] the authors observe numerically that their sparse grid technique appears to converge in a balanced norm.

Writing N for the maximum number of mesh intervals in each coordinate direction, our combination technique simply adds or subtracts solutions that have been computed by the Galerkin FEM on $N\times\sqrt{N}$, $\sqrt{N}\times N$ and $\sqrt{N}\times\sqrt{N}$ meshes. We obtain the same accuracy as on an $N\times N$ mesh with less degrees of freedom. In the following we use the notation of [13].

In the combination technique for bilinear elements we compute a two-scale finite element approximation $u^N_{\hat{N},\hat{N}}$ with $\hat{N}=\sqrt{N}$ by

$$u^N_{\hat{N},\hat{N}} := u_{N,\hat{N}} + u_{\hat{N},N} - u_{\hat{N},\hat{N}}.$$

We proved (in our new notation)

$$\|u-u_{NN}\|_b \preceq N^{-1}(\ln N)^{3/2} + N^{-2}. \tag{4.3}$$

The question is whether or not $u^N_{\hat{N},\hat{N}}$ satisfies a similar estimate.

Analogously to $u^N_{\hat{N},\hat{N}}$ we define $I^N_{\hat{N},\hat{N}}E$ and $\Pi^N_{\hat{N},\hat{N}}S$. Then we can decompose the error to estimate as follows:

$$u^N_{\hat{N},\hat{N}} - u_{NN} = T_{cl,1}(S) + (\Pi^N_{\hat{N},\hat{N}}S - \Pi_{N,N}S) + T_{cl,2}(E) + (I^N_{\hat{N},\hat{N}}E - I_{N,N}E).$$

Thus we have two terms representing the error for two-scale projection operators (related to L_2 projection and interpolation, respectively) and two terms which can be estimated based on our supercloseness result:

$$T_{cl,1}(S) := (S_{N,\hat{N}} - \Pi_{N,\hat{N}}S) + (S_{\hat{N},N} - \Pi_{\hat{N},N}S) - (S_{\hat{N},\hat{N}} - \Pi_{\hat{N},\hat{N}}S) - (S_{N,N} - \Pi_{N,N}S),$$

analogously

$$T_{cl,2}(E) := (E_{N,\hat{N}} - I_{N,\hat{N}}E) + (E_{\hat{N},N} - I_{\hat{N},N}E) - (E_{\hat{N},\hat{N}} - I_{\hat{N},\hat{N}}E) - (E_{N,N} - I_{N,N}E).$$

For the two-scale interpolation error $(I^N_{\hat{N},\hat{N}}E - I_{N,N}E)$ the results of [13] remain valid (Lemma 2.3 and 2.5, modified for the reaction-diffusion problem). For the two-scale projection error an estimate in L_2 and L_∞ is easy. The estimate in the seminorm $|\cdot|_1$ as in Sect. 2 follows from an inverse inequality, applied separately in Ω_0 and Ω_f. Finally we get for $\hat{N} = \sqrt{N}$ the estimate

$$\|u^N_{\hat{N},\hat{N}} - u_{NN}\|_b \preceq \varepsilon^{1/4}N^{-1/2} + N^{-1}\ln N. \tag{4.4}$$

The result is not satisfactory, so far we can only prove the desired estimate for the combination technique if $\varepsilon \preceq N^{-2}$.

5 A Direct Mixed Method

The first balanced error estimate was presented by Lin and Stynes [19] using a first order system least squares (FOSLS-like) mixed method. For the variables $(u, \bar{q})$ with $-\bar{q} = \nabla u$ and its discretizations on a Shishkin mesh they proved

$$\varepsilon^{1/4}\|\bar{q} - \bar{q}^N\|_0 + \|u - u^N\|_0 \preceq N^{-1}\ln N \tag{5.1}$$

(see also [2] for a modified version of the method).

We shall proof that the estimate (5.1) is also valid for a direct mixed method (instead of the more complicated least-squares approach from [19]). We remark that Li and Wheeler [18] analyzed the method in the energy norm on so called A-meshes, which are simpler to analyze than S-meshes.

Introducing $\bar{q} = -\nabla u$, a weak formulation of (1.1) reads:
Find $(u, \bar{q}) \in W \times V$ such that

$$\varepsilon(div\, \bar{q}, w) + c(u, w) = (f, w) \qquad \text{for all } w \in W, \tag{5.2a}$$

$$\varepsilon(\bar{q}, \bar{v}) - \varepsilon(div\, \bar{v}, u) = 0 \qquad \text{for all } \bar{v} \in V, \tag{5.2b}$$

with $V = H(div, \Omega)$, $W = L^2(\Omega)$.

For the discretization on a standard rectangular Shishkin mesh we use $(u^N, \bar{q}^N) \in V^N \times W^N$. Here W^N is the space of piecewise constants on our rectangular mesh and V^N the lowest order Raviart-Thomas space RT_0. That means, on each mesh rectangle elements of RT_0 are vectors of the form

$$(span(1, x), span(1, y))^T.$$

Our discrete problem reads: Find $(u^N, \bar{q}^N) \in W^N \times V^N$ such that

$$\varepsilon(div\, \bar{q}^N, w) + c(u^N, w) = (f, w) \qquad \text{for all } w \in W^N, \tag{5.3a}$$

$$\varepsilon(\bar{q}^N, \bar{v}) - \varepsilon(div\, \bar{v}, u^N) = 0 \qquad \text{for all } \bar{v} \in V^N. \tag{5.3b}$$

Setting $w := u^N$, $\bar{v} := \bar{q}^N$ results in the stability estimate

$$\varepsilon \|\bar{q}^N\|_0^2 + \frac{c}{2}\|u^N\|_0^2 \preceq \|f\|_0^2. \tag{5.4}$$

The unique solvability of the discrete problem follows (if $f \equiv 0$).

For the error estimation we introduce projections $\Pi : V \mapsto V^N$ and $P : W \mapsto W^N$. As usual, instead of $u - u^N$ and $\bar{q} - \bar{q}^N$ we estimate $Pu - u^N$ and $\Pi\bar{q} - \bar{q}^N$, assuming that we can estimate the projection errors. Subtraction of the continuous and the discrete problem results in

$$\varepsilon(\nabla \cdot (\Pi\bar{q} - \bar{q}^N), w) + c(Pu - u^N, w) = \varepsilon(\nabla \cdot (\Pi\bar{q} - \bar{q}), w) + c(Pu - u, w), \tag{5.5a}$$

$$\varepsilon(\Pi\bar{q} - \bar{q}^N, \bar{v}) - \varepsilon(\nabla \cdot \bar{v}, Pu - u^N) = \varepsilon(\Pi\bar{q} - \bar{q}, \bar{v}) - \varepsilon(\nabla \cdot \bar{v}, Pu - u). \tag{5.5b}$$

Setting $\bar{v} := \Pi\bar{q} - \bar{q}^N = \bar{\mu}$ and $w := Pu - u^N = \tau$ we obtain the error equation

$$\varepsilon(\bar{\mu}, \bar{\mu}) + c(\tau, \tau) = \varepsilon(\nabla \cdot (\Pi\bar{q} - \bar{q}), \tau) + c(Pu - u, \tau) + \varepsilon(\Pi\bar{q} - \bar{q}, \bar{\mu}) - \varepsilon(\nabla \cdot \bar{\mu}, Pu - u). \tag{5.6}$$

From the error equation it is easy to derive a first order uniform convergence result in the energy norm (one could also think about supercloseness similar as in [18]). But we want to investigate whether or not an estimate of the type (5.1) is possible.

If P denotes the L_2 projection, we have

$$(Pu - u, \tau) = 0 \quad \text{and} \quad (\nabla \cdot \bar{\mu}, Pu - u) = 0,$$

because $\nabla \cdot \bar{\mu}$ is piecewise constant for $\bar{\mu} \in V^N$. Therefore, from the right hand side of the error equation two terms disappear and it follows

$$\|\bar{\mu}\|_0^2 \preceq \varepsilon \|\nabla \cdot (\Pi(\nabla u) - \nabla u)\|_0^2 + \|\Pi(\nabla u) - \nabla u\|_0^2. \tag{5.7}$$

Now let us denote by Π^* the standard local projection operator into the Raviart-Thomas space V^N. This operator satisfies

$$(\nabla \cdot (\bar{v} - \Pi^* \bar{v}), w) = 0 \qquad \text{for all } w \in W^N. \tag{5.8}$$

Consequently, the choice $\Pi = \Pi^*$ would eliminate one more term in the error equation and thus in (5.7). But do we have for the projection error the desired estimate

$$\varepsilon^{1/4} \|\Pi^*(\nabla u) - \nabla u\|_0 \preceq N^{-1} \ln N \quad ? \tag{5.9}$$

The answer is no (see Lin and Stynes [19], page 2738). The reason lies in the fact that Π^* is applied to ∇u and its behavior near the transition point of the mesh is different from the behavior of u (a factor $\varepsilon^{-1/2}$).

Therefore, Lin and Stynes define a modified interpolant $\Pi \bar{v} \in W^N$, such that

$$\varepsilon^{1/4} \|\Pi(\nabla u) - \nabla u\|_0 \preceq N^{-1} \ln N \tag{5.10}$$

([19], Corollary 4.6). The operator Π is defined differently for every component of the solution decomposition. For the smooth part one takes simply $\Pi = \Pi^*$.

For the layer components, however, Π^* is modified. Consider, for instance, the layer component w_1 related to $\exp(-\sqrt{c^*}y/\sqrt{\varepsilon})$. Then Π and Π^* differ only in the small strip R_1 defined by

$$R_1 := [0,1] \times [\lambda - h^*, \lambda] \qquad \text{with} \quad \lambda = 2\sqrt{\varepsilon} \ln N / \sqrt{c} \quad \text{and} \quad h^* = O(\sqrt{\varepsilon} N^{-1} \ln N).$$

On that strip we loose the property (5.8), therefore we additionally have to estimate

$$M_{1,R_1} := \varepsilon^{1/2} \|\nabla \cdot (\Pi(\nabla w_1) - \nabla w_1)\|_{0,R_1}. \tag{5.11}$$

On R_1 we have $\|\Delta w_1\|_\infty \preceq \varepsilon^{-1} N^{-2}$, consequently

$$\varepsilon^{1/2} \|\Delta w_1\|_{0,R_1} \preceq \varepsilon^{-1/2} N^{-2} \varepsilon^{1/4} N^{-1/2} (\ln N)^{1/2} = \varepsilon^{-1/4} N^{-5/2} (\ln N)^{1/2}. \tag{5.12}$$

By construction the components of $\Pi(\nabla w_1)$ satisfy $(\Pi(\nabla w_1))_1 = 0$ on R_1 and $\|(\Pi(\nabla w_1))_2\|_\infty \preceq \varepsilon^{-1/2}N^{-2}$. It follows

$$\varepsilon^{1/2}\|\nabla\cdot(\Pi\nabla w_1)\|_{0,R_1} \preceq \varepsilon^{1/2}\frac{1}{h^*}\varepsilon^{-1/2}N^{-2}(h^*)^{1/2} = \varepsilon^{-1/4}N^{-3/2}(\ln N)^{-1/2}. \tag{5.13}$$

Therefore

$$M_{1,R_1} \preceq \varepsilon^{-1/4}N^{-3/2}. \tag{5.14}$$

The other layer components of the solution decomposition of u are treated similarly. We obtain finally

$$\varepsilon^{1/4}\|\Pi\bar{q} - \bar{q}^N\|_0 \preceq N^{-1}\ln N \tag{5.15}$$

and

$$\varepsilon^{1/4}\|\nabla u - \bar{q}^N\|_0 \preceq N^{-1}\ln N. \tag{5.16}$$

Remark 1 It is well known [5, 6] that mixed methods can be reformulated as non-mixed formulations, more precisely as projected nonconforming methods. This allows error estimates for certain nonconforming methods to be established. Moreover, certain mixed methods can be implemented as a nonconforming method.

6 Remarks and Further Open Problems

First let us remark that for systems

$$-\varepsilon u'' + Au = f \qquad \text{in } \Omega = (0,1), \tag{6.1a}$$

$$u(0) = u(1) = 0 \qquad \text{on } \partial\Omega, \tag{6.1b}$$

so far there exists only a result of Lin and Stynes [20] in a balanced norm. Following the basic idea from [19], but using C^1 elements instead of mixed finite elements, they introduce the bilinear form

$$\varepsilon(w', v') + (Aw, v) + \varepsilon^{3/2}(w'', v'') + \varepsilon^{1/2}((Aw)', v')$$

and analyse the finite element method for quadratic C^1 elements. The analysis for the Galerkin method with C^0 elements is open. It would be especially very interesting to get estimates in balanced norms for systems with several small parameters.

New difficulties arise as well for problems with different layers in one coordinate direction, even for the simple 1D problem

$$-\varepsilon^3 u'' + \varepsilon b(x)u' + c(x)u = f.$$

Here the layers E_0 at $x = 0$ and E_1 at $x = 1$ can be very different:

$$E_0 \approx \exp(-x/\varepsilon), \quad \text{but} \quad E_1 \approx \exp(-(1-x)/\varepsilon^2).$$

That means, in a balanced norm one should scale the H^1 seminorm near $x = 0$ with $\varepsilon^{1/2}$, but near $x = 1$ with ε (in the energy norm, however, $\varepsilon^{3/2}$ arises as scaling factor).

References

1. Acosta, G.: Lagrange and average interpolation over 3D anisotropic elements. J. Comput. Appl. Math. **135**, 91–109 (2001)
2. Adler, J., MacLachlan, S., Madden, N.: A first-order system Petrov-Galerkin discretisation for a reaction-diffusion problem on a fitted mesh. IMA J. Numer. Anal. **36**, 1281–1309 (2016)
3. Al Shenk, N.: Uniform error estimates for certain narrow Lagrange finite elements. Math. Comput. **63**, 105–119 (1994)
4. Apel, T.: Anisotropic interpolation error estimates for isoparametric quadrilateral finite elements. Computing **60**, 157–174 (1998)
5. Arbogast, T., Chen, Z.: On the implementation of mixed methods as nonconforming methods for second order elliptic problems. Math. Comput. **64**, 943–972 (1995)
6. Arnold, D.N., Brezzi, F.: Mixed and nonconforming finite element methods. RAIRO Modél. Math. Anal. Numer. **19**, 7–32 (1985)
7. Constantinou, P., Xenophontos, C.: Finite element analysis of an exponentially graded mesh for singularly perturbed problems. Comput. Methods Appl. Math. **15**, 135–143 (2015)
8. Crouzeix, M., Thomee, V.: The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces. Math. Comput. **48**, 521–532 (1987)
9. Dobrowolski, M., Roos, H.-G.: A priori estimates for the solution of convection-diffusion problems and interpolation on Shishkin meshes. Z. Anal. Anwend. **16**, 1001–1012 (1997)
10. Duran, R.G.: Error estimates for 3-d narrow finite elements. Math. Comput. **68**, 187–199 (1999)
11. Franz, S., Roos, H.-G.: Error estimates in a balanced norm for a convection-diffusion problem with two different boundary layers. Calcolo **51**, 423–440 (2014)
12. Franz, S., Roos, H.-G.: Robust error estimation in energy and balanced norms for singularly perturbed fourth order problems. Comput. Math. Appl. **72**, 233–247 (2016)
13. Franz, S., Liu, F., Roos, H.-G., Stynes, M., Zhou, A.: The combination technique for a two-dimensional convection-diffusion problem with exponential layers. Appl. Math. **54**, 203–223 (2009)
14. de Jager, E.M., Furu, J.: The Theory of Singular Perturbations. North Holland, Amsterdam (1996)
15. Kellogg, R.B., Kopteva, N.: A singularly perturbed semilinear reaction-diffusion problem in a polygonal domain. J. Differ. Equ. **248**, 184–208 (2010)
16. Li, J.: Quasioptimal uniformly convergent finite element methods for the elliptic boundary layer problem. Comput. Math. Appl. **33**, 11–22 (1997)

17. Li, J.: Uniform error estimates in the finite element method for a singularly perturbed reaction-diffusion problem. Appl. Numer. Math. **36**, 129–154 (2001)
18. Li, J., Wheeler, M.F.: Uniform convergence and superconvergence of mixed finite element methods on anisotropically refined grids. SIAM J. Numer. Anal. **38**, 770–798 (2000)
19. Lin, R., Stynes, M.: A balanced finite element method for singularly perturbed reaction-diffusion problems. SIAM J. Numer. Anal. **50**, 2729–2743 (2012)
20. Lin, R., Stynes, M.: A balanced finite element method for a system of singularly perturbed reaction-diffusion two-point boundary value problems. Numer. Algorithms **70**, 691–707 (2015)
21. Liu, F., Madden, N., Stynes, M., Zhou, A.: A two-scale sparse grid method for singularly perturbed reaction-diffusion problems in two dimensions. IMA J. Numer. Anal. **29**, 986–1007 (2009)
22. Madden, N., Russell, S.: A multiscale sparse grid finite element method for a two-dimensional singularly perturbed reaction-diffusion problem. Adv. Comput. Math. **41**, 987–1014 (2015)
23. Melenk, J.M.: Hp-Finite Element Methods for Singular Perturbations. Springer, Berlin (2002)
24. Melenk, J.M., Xenophontos, C.: Robust exponential convergence of hp-FEM in balanced norms for singularly perturbed reaction-diffusion equations. Calcolo **53**, 105–132 (2016)
25. Oswald, P.: L_∞-bounds for the L_2-projection onto linear spline spaces. In: Recent Advances in Harmonic Analysis and Applications, pp. 303–316. Springer, New York, (2013)
26. Roos, H.-G.: Error estimates for linear finite elements on Bakhvalov-type meshes. Appl. Math. **51**, 63–72 (2006)
27. Roos, H.-G.: Robust numerical methods for singularly perturbed differential equations: a survey covering 2008–2012. ISRN Appl. Math. article ID 379547, 1–30 (2012)
28. Roos, H.-G., Linss, T.: Sufficient conditions for uniform convergence on layer-adapted grids. Computing **63**, 27–45 (1999)
29. Roos, H.-G., Schopf, M.: Analysis of finite element methods on Bakhvalov-type meshes for linear convection-diffusion problems in 2D. Appl. Math. **57**, 97–108 (2012)
30. Roos, H.-G., Schopf, M.: Convergence and stability in balanced norms for finite element methods on Shishkin meshes for reaction-diffusion problems. Z. Angew. Math. Mech **95**, 334–351 (2015)
31. Roos, H.-G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Springer, Berlin (2008)
32. Roos, H.-G., Teofanov, L., Uzelac, Z.: Graded meshes for higher order FEM. J. Comput. Math. **33**, 1–16 (2015)

Numerical Studies of Higher Order Variational Time Stepping Schemes for Evolutionary Navier-Stokes Equations

Naveed Ahmed and Gunar Matthies

Abstract We present in this paper numerical studies of higher order variational time stepping schemes combined with finite element methods for simulations of the evolutionary Navier-Stokes equations. In particular, conforming inf-sup stable pairs of finite element spaces for approximating velocity and pressure are used as spatial discretization while continuous Galerkin–Petrov methods (cGP) and discontinuous Galerkin (dG) methods are applied as higher order variational time discretizations. Numerical results for the well-known problem of incompressible flows around a circle will be presented.

1 Introduction

The flow of incompressible fluids is described by the time-dependent, incompressible Navier–Stokes equations. In order to solve them numerically, one has to discretize in space and time. Often the method of lines is applied where the problem is discretized in space first while the time remains continuous. This technique leads to a large system of ordinary differential equations which can be solved by suitable ODE solvers. Note that the resulting system of ODE is nonlinear due to the nonlinear convection term in the Navier–Stokes equations.

We will consider continuous Galerkin–Petrov and discontinuous Galerkin methods as higher order variational time discretizations. In continuous Galerkin–Petrov (cGP) methods, the ansatz functions are continuous in time while the discontinuous test functions allow a time marching process. In discontinuous Galerkin (dG) schemes, ansatz and test functions are from the same space and allowed to be discontinuous at the discrete time points. Hence, a time marching process is possible

N. Ahmed (✉)
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
e-mail: naveed.ahmed@wias-berlin.de

G. Matthies
Technische Universität Dresden, Institut für Numerische Mathematik, 01062 Dresden, Germany
e-mail: gunar.matthies@tu-dresden.de

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_2

as well. Variational time discretizations by cGP and dG methods allow A-stable or even strongly A-stable methods of any order while time discretizations of BDF-type are A-stable only for orders up to 2.

The cGP method has been studied in [1] for the heat equation. Theoretical and numerical investigations of higher order variational time discretizations applied to different types of incompressible flow problems can be found in [2–6]. Note that cGP methods are A-stable whereas dG methods are even strongly A-stable which might lead to different damping properties with respect to high frequency error components. We refer to [7] for more information on dG methods.

The inf-sup condition plays a fundamental role for solving incompressible flow problems without additional pressure stabilization. Using inf-sup stable pairs of finite element spaces for approximation velocity and pressure is guided by the observation that flow problems are often part of coupled problems of flow and transport where mass conservation depends crucially on the properties of the discrete velocity, see [8]. Since the property of a velocity field being discretely divergence-free is disturbed by pressure stabilization, the use of inf-sup stable discretizations is favorable.

We will describe in this paper the discretization of the evolutionary Navier–Stokes equations in space by inf-sup stable finite element pairs for approximating velocity and pressure together with higher order variational time stepping schemes using continuous Galerkin–Petrov and discontinuous Galerkin methods. In addition, a post-processing technique given in [9] for systems of ordinary differential equations is adapted in order to improve the accuracy of the numerical solution. The proposed solution strategy will be applied to the well-know benchmark problem of an incompressible flow around a circle.

The remainder of this paper is organized as follows. Section 2 introduces the evolutionary, incompressible Navier–Stokes equations and their finite element discretizations. Variational time discretizations by continuous Galerkin–Petrov and discontinuous Galerkin methods are described in Sect. 3 where also the post-processing techniques is given. Numerical results for the benchmark problem "flow around a circle" will be given in Sect. 4.

2 Model Problem and Its Finite Element Discretization

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a Lipschitz domain with polyhedral boundary $\partial\Omega$ and $T > 0$ a finite time. The motion of incompressible fluids is modeled by the time-dependent, incompressible Navier–Stokes equations which in dimensionless form are defined by

$$\begin{aligned} \mathbf{u}' - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f} \qquad \text{in } (0,T]\times\Omega, \\ \nabla\cdot\mathbf{u} &= 0 \qquad \text{in } (0,T]\times\Omega. \end{aligned} \tag{1}$$

Here, $\mathbf{f}$ is a given body force, ν the viscosity, $\mathbf{u}$ and p denote the velocity and pressure fields, respectively. The partial differential equations in (1) have to be closed by appropriate initial and boundary conditions. For simplicity, we consider homogeneous Dirichlet boundary conditions on $[0,T] \times \partial\Omega$ and a given initial velocity field $\mathbf{u}_0$ in Ω.

We introduce the spaces $\mathbf{V} = H_0^1(\Omega)^d$, $Q = L_0^2(\Omega)$, and $W = \{\mathbf{v} \in L^2(0,T;\mathbf{V}) : \mathbf{v}' \in L^2(0,T;\mathbf{V}')\}$ with $\mathbf{V}' = H^{-1}(\Omega)^d$ as dual space of $\mathbf{V}$.

Assuming $\mathbf{f} \in L^2\big(0,T;L^2(\Omega)^d\big)$, a variational formulation of problem (1) reads:

Find $\mathbf{u} \in W$ and $p \in L^2(0,T;Q)$ such that $\mathbf{u}(0) = \mathbf{u}_0$ and for almost all $t \in (0,T)$

$$\begin{aligned} \langle \mathbf{u}'(t), \mathbf{v}\rangle + \nu(\nabla \mathbf{u}(t), \nabla \mathbf{v}) + ((\mathbf{u}(t)\cdot\nabla)\mathbf{u}(t), \mathbf{v}) - (p(t), \nabla\cdot\mathbf{v}) &= (\mathbf{f}(t), \mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{V}, \\ (q, \operatorname{div}\mathbf{u}) &= 0 \qquad \forall q \in Q. \end{aligned} \tag{2}$$

Note that $\langle\cdot,\cdot\rangle$ denote the duality pairing between $\mathbf{V}$ and $\mathbf{V}'$ while $(\cdot,\cdot)$ is the inner product in $L^2(\Omega)$ and its vector-valued and tensor-valued versions. The corresponding L^2-norm is given by $\|\cdot\|_0$ while $|\cdot|_m$ indicates the semi-norm in $H^m(\Omega)$ and its vector-valued version.

For finite element discretizations of (2), we are given a family $\{\mathcal{T}_h\}$ of shape-regular decomposition of Ω into d-simplices, quadrilaterals, or hexahedra. The diameter of a cell K is denoted by h_K and the mesh size h is defined by $h := \max_{K\in\mathcal{T}_h} h_K$.

We consider pairs of conforming finite element spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ for approximation velocity and pressure where we assume that $\mathbf{V}_h = Y_h^d$ with a scalar finite element space Y_h. The unique solvability of the system arising from the discretization and linearization of (2) in space requires to satisfy the inf-sup stability condition

$$\inf_{q_h\in Q_h}\sup_{\mathbf{v}_h\in\mathbf{V}_h} \frac{(q_h, \nabla\cdot\mathbf{v}_h)}{\|q_h\|_0|\mathbf{v}_h|_1} \geq \beta > 0. \tag{3}$$

Then, the finite element discretization of (2) reads:

Find $\mathbf{u}_h \in H^1(0,T;\mathbf{V}_h)$ and $p_h \in L^2(0,T;Q_h)$ such that with $\mathbf{u}_h(0) = \mathbf{u}_{0,h}$ and for almost all $t \in (0,T)$

$$(\mathbf{u}_h'(t), \mathbf{v}_h) + A\left(\mathbf{u}_h(t), (\mathbf{u}_h(t), p_h(t)), (\mathbf{v}_h, q_h)\right) = (\mathbf{f}(t), \mathbf{v}_h) \quad \forall (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h \tag{4}$$

where $\mathbf{u}_{0,h} \in \mathbf{V}_h$ is a suitable approximation of the initial velocity $\mathbf{u}_0$ and A is defined by

$$A\left(\mathbf{w}, (\mathbf{u}, p), (\mathbf{v}, q)\right) = \nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + ((\mathbf{w}\cdot\nabla)\mathbf{u}, \mathbf{v}) - (p, \nabla\cdot\mathbf{v}) + (q, \nabla\cdot\mathbf{u}).$$

Note that A is linear in its second and third argument while the problem (4) is non-linear. For small values of the viscosity parameter ν, spatial stabilization becomes necessary and additional terms will appear in the discrete scheme, see [10, 11] for examples.

3 Variational Time-Stepping Schemes

In this section, we discretize problem (4) in time by continuous Galerkin–Petrov (cGP) and discontinuous Galerkin (dG) methods. To this end, we consider a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of the time interval $I := [0, T]$ and set $I_n := (t_{n-1}, t_n]$, $\tau_n = t_n - t_{n-1}$, $n = 1, \ldots, N$, and $\tau := \max_{1 \le n \le N} \tau_n$. For a given non-negative integer k, we define the time-continuous and time-discontinuous velocity spaces

$$X_k^{\mathrm{c}} := \Big\{ \mathbf{u} \in C(0, T; \mathbf{V}_h) : \mathbf{u}|_{I_n} \in \mathbb{P}_k(I_n, \mathbf{V}_h), n = 1, \ldots, N \Big\},$$
$$X_k^{\mathrm{dc}} := \Big\{ \mathbf{u} \in L^2(0, T; \mathbf{V}_h) : \mathbf{u}|_{I_n} \in \mathbb{P}_k(I_n, \mathbf{V}_h), n = 1, \ldots, N \Big\}$$

and time-continuous and time-discontinuous pressure spaces

$$Y_k^{\mathrm{c}} := \Big\{ q \in C(0, T; Q_h) : q|_{I_n} \in \mathbb{P}_k(I_n, Q_h), n = 1, \ldots, N \Big\},$$
$$Y_k^{\mathrm{dc}} := \Big\{ q \in L^2(0, T; Q_h) : q|_{I_n} \in \mathbb{P}_k(I_n, Q_h), n = 1, \ldots, N \Big\}$$

where

$$\mathbb{P}_k(I_n, W_h) := \Big\{ u : I_n \to W_h : u(t) = \sum_{i=0}^{k} U_i t^i, t \in I_n, U_i \in W_h, i = 0, \ldots, k \Big\}$$

denotes the space of W_h-valued polynomials of degree less than or equal to k in time. The function in the spaces X_k^{dc} and Y_k^{dc} are allowed to be discontinuous at the nodes t_n, $n = 1, \ldots, N-1$. For a piecewise smooth function w, let

$$w_n^- := \lim_{t \to t_n - 0} w(t), \quad w_n^+ := \lim_{t \to t_n + 0} w(t), \quad [w]_n := w_n^+ - w_n^-$$

denote the left-sided value, the right-sided value, and the jump, respectively. □

3.1 The Continuous Galerkin-Petrov Method

In this section, we discretize the semi-discrete problem (4) in time by cGP methods to obtain a fully discrete formulation of (2). Now, the cGP(k) method reads:

Find $\mathbf{u}_{h,\tau} \in X_k^{\mathrm{c}}$ and $p_{h,\tau} \in Y_k^{\mathrm{c}}$ such that $\mathbf{u}_h(0) = \mathbf{u}_{0,h}$ and

$$\int_0^T \left[(\mathbf{u}'_{h,\tau}, \mathbf{v}_{h,\tau}) + A\left(\mathbf{u}_{h,\tau}, (\mathbf{u}_{h,\tau}, p_{h,\tau}), (\mathbf{v}_{h,\tau}, q_{h,\tau})\right) \right]$$
$$= \int_0^T (\mathbf{f}, \mathbf{v}_{h,\tau}) \qquad \forall \mathbf{v}_{h,\tau} \in X_{k-1}^{\mathrm{dc}}, \forall q_{h,\tau} \in Y_{k-1}^{\mathrm{dc}} \tag{5}$$

where the index h, τ refers to the full discretization in space and time, respectively.

Since the test functions are allowed to be discontinuous at the discrete time points $t_n, n = 1, \ldots, N-1$, we can choose the test function $(\mathbf{v}_{h,\tau}, q_{h,\tau}) = (\mathbf{v}_h, q_h)\psi(t)$ with time independent $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ and a scalar function $\psi : I_n \to \mathbb{R}$ which is zero on $I \setminus I_n$ and a polynomial of degree less than or equal to $k-1$ on I_n. Then, the solution of the cGP(k) method can be determined by successively solving a single local problem on each time interval.

The fully discrete time marching scheme associated to (5) reads:

Find $\mathbf{u}_{h,\tau}\big|_{I_n} \in \mathbb{P}_k(I_n, \mathbf{V}_h)$ and $p_{h,\tau}\big|_{I_n} \in \mathbb{P}_k(I_n, Q_h)$ such that for all $\psi \in \mathbb{P}_{k-1}(I_n)$

$$\int_{I_n} \left[(\mathbf{u}'_{h,\tau}, \mathbf{v}_h) + A\left(\mathbf{u}_{h,\tau}, (\mathbf{u}_{h,\tau}, p_{h,\tau}), (\mathbf{v}_h, q_h)\right) \right] \psi(t)$$

$$= \int_0^T (\mathbf{f}, \mathbf{v}_h)\psi(t) \qquad \forall (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$$

with $\mathbf{u}_{h,\tau}\big|_{I_1}(t_0) = \mathbf{u}_{0,h}$ and $\mathbf{u}_{h,\tau}\big|_{I_n}(t_{n-1}) = \mathbf{u}_{h,\tau}\big|_{I_{n-1}}(t_{n-1})$ for $n \geq 2$.

We apply for the numerical integration of the time integrals the Gauß-Lobatto quadrature rule with $(k+1)$ points. This formula is exact for polynomials of degree less than or equal to $2k-1$. Let $\hat{t}_j$ and $\hat{w}_j, j = 0, \ldots, k$, be the Gauß-Lobatto points and the corresponding quadrature weights on $[-1, 1]$, respectively. Furthermore, we denote by $\hat{\phi}_j \in \mathbb{P}_k, j = 0, \ldots, k$, and $\hat{\psi}_j \in \mathbb{P}_{k-1},\ j = 1, \ldots, k$, the Lagrange basis function with respect to $\hat{t}_j, j = 0, \ldots, k$, and $\hat{t}_j, j = 1, \ldots, k$, respectively. The time polynomials $\phi_{n,j} \in \mathbb{P}_k(I_n)$, $j = 0, \ldots, k$, and $\psi_{n,j} \in \mathbb{P}_{k-1}(I_n)$, $j = 1, \ldots, k$, are defined by

$$\phi_{n,j}(t) := \hat{\phi}_j\left(T_n^{-1}(t)\right) \quad \text{and} \quad \psi_{n,j}(t) := \hat{\psi}_j\left(T_n^{-1}(t)\right)$$

with the affine reference transformation

$$T_n : [-1, 1] \to \overline{I_n}, \quad \hat{t} \mapsto t_{n-1} + \frac{\tau_n}{2}(\hat{t} + 1), \tag{6}$$

see [9].

Since the restrictions of $\mathbf{u}_{h,\tau}$ and $p_{h,\tau}$ to the interval I_n are $\mathbf{V}_h$-valued and Q_h-valued polynomials of degree less than or equal to k , they can be represented as

$$\mathbf{u}_{h,\tau}\big|_{I_n} = \sum_{j=0}^{k} U_{n,h}^j \phi_{n,h}^j(t), \quad p_{h,\tau}\big|_{I_n} = \sum_{j=0}^{k} P_{n,h}^j \phi_{n,h}^j(t), \quad t \in I_n,$$

with coefficients $U_{n,h}^j \in \mathbf{V}_h$ and $P_{n,h}^j \in Q_h$, $j = 0, \ldots, k$. The particular ansatz ensures

$$\mathbf{u}_{h,\tau}(t_{n,j}) = U_{n,h}^j, \quad p_{h,\tau}(t_{n,j}) = P_{n,h}^j, \quad j = 0, \ldots, k,$$

where $t_{n,j} := T_n(\hat{t}_j)$, $j = 0, \ldots, k$. Since $t_{n,0} = t_{n-1}$ and $t_{n,k} = t_n$ hold, the initial conditions on the intervals I_n, $n = 1, \ldots, N$, are equivalent to the conditions

$$U^0_{1,h} = \mathbf{u}_{0,h}, \quad \text{and} \quad U^0_{n,h} = \mathbf{u}_{h,\tau}\big|_{I_n}(t_{n-1}) = U^k_{n-1,h} \quad \text{if } n \geq 2.$$

Using the properties of the basis functions in time, we obtain the following coupled system of nonlinear equations:

For $U^0_{1,h} = \mathbf{u}_{0,h}$ and $U^0_{n,h} = U^k_{n-1,h}$ if $n \geq 2$, find the coefficients $U^j_{n,h} \in \mathbf{V}_h$ and $P^j_{n,h}$, $j = 1, \ldots, k$, such that

$$\begin{aligned} &\sum_{j=0}^{k} \alpha^c_{i,j} \left(U^j_{n,h}, \mathbf{v}_h \right) + \frac{\tau_n}{2} A\left(U^i_{n,h}, (U^i_{n,h}, P^i_{n,h}), (\mathbf{v}_h, q_h) \right) \\ &\qquad = \frac{\tau_n}{2} \left\{ \left(\mathbf{f}(t_{n,i}), \mathbf{v}_h \right) + \beta^c_i \left(\mathbf{f}(t_{n-1}), \mathbf{v}_h \right) \right\} \end{aligned} \tag{7}$$

for $i = 1, \ldots, k$, for all $\mathbf{v}_h \in \mathbf{V}_h$, and for all $q_h \in Q_h$, where $\alpha^c_{i,j}$ and β^c_i are defined by

$$\alpha^c_{i,j} := \hat{\phi}'_j(\hat{t}_i) + \beta^c_i \hat{\phi}'_j(\hat{t}_0), \quad \beta^c_i := \hat{w}_0 \hat{\psi}_i(\hat{t}_0), \quad i = 1, \ldots, k, j = 0, \ldots, k,$$

see [9].

In the following, we write (7) as a nonlinear algebraic block system. For simplicity, we restrict ourselves to the two-dimensional case. The three-dimensional case is obtained in a straightforward manner.

Let $\{\phi_i \in Y_h, i = 1, \ldots, m_h\}$ be a finite element basis of Y_h and $\boldsymbol{\xi}^j_{n,1}, \boldsymbol{\xi}^j_{n,2} \in \mathbb{R}^{m_h}$ denote the nodal vectors associated to the components of the finite element function $U^j_{n,h} \in \mathbf{V}_h$ such that

$$U^j_{n,h}(x) = \sum_{l=1}^{2} \left(\sum_{\nu=1}^{m_h} \left(\boldsymbol{\xi}^j_{n,l} \right)_\nu \phi_\nu(x) \right) e^l, \quad x \in \Omega,$$

where $e^1, e^2 \in \mathbb{R}^2$ are the canonical unit vectors. Similarly for the pressure, let $\{\psi_i \in Q_h, i = 1, \ldots, n_h\}$, denote a finite element basis of Q_h and $\boldsymbol{\eta}^j_n$ the nodal vector of $P^j_{n,h} \in Q_h$ such that

$$P^j_{n,h}(x) = \sum_{\nu=1}^{m_h} \left(\boldsymbol{\eta}^j_n \right)_\nu \psi_\nu(x), \quad x \in \Omega.$$

Furthermore, the mass matrix $M \in \mathbb{R}^{m_h \times m_h}$, the matrix $A \in \mathbb{R}^{m_h \times m_h}$, the velocity-pressure coupling matrices $B_i \in \mathbb{R}^{n_h \times m_h}$, and the right-hand side vectors $F^j_{n,i} \in \mathbb{R}^{m_h}$, $i = 1, 2$, are given by

$$\begin{aligned} (M)_{s,k} &:= (\phi_k, \phi_s), & (A)_{s,k} &:= \nu(\nabla \phi_k, \nabla \phi_s), \\ (B_i)_{s,k} &:= -\left(\psi_s, \nabla \cdot (\phi_k e^i) \right), & (F^j_{n,i})_k &:= \left(f(t_{n,j}), \phi_k e^i \right), \quad i = 1, 2. \end{aligned}$$

For a given discrete velocity field $\mathbf{w}_h \in \mathbf{V}_h$ and its nodal vector $\overline{\mathbf{w}} \in \mathbb{R}^{2m_h}$, the matrix representation $N(\overline{\mathbf{w}}) \in \mathbb{R}^{m_h \times m_h}$ of the nonlinear term is defined by

$$\big(N(\overline{\mathbf{w}})\big)_{s,k} := \big((\mathbf{w} \cdot \nabla)\phi_k, \phi_s\big)\,. \tag{8}$$

We define the block matrices

$$\mathscr{M} = \begin{bmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathscr{A} = \begin{bmatrix} A + N(\overline{\mathbf{w}}) & 0 & B_1^T \\ 0 & A + N(\overline{\mathbf{w}}) & B_2^T \\ B_1 & B_2 & 0 \end{bmatrix}, \tag{9}$$

and the block vectors

$$F_n^j = \begin{bmatrix} F_{n,1}^j \\ F_{n,2}^j \\ 0 \end{bmatrix}, \qquad \boldsymbol{\zeta}_n^j = \begin{bmatrix} \boldsymbol{\xi}_{n,1}^j \\ \boldsymbol{\xi}_{n,2}^j \\ \boldsymbol{\eta}_n^j \end{bmatrix}. \tag{10}$$

Then, the fully discrete problem (7) on I_n is equivalent to the nonlinear $k \times k$ block system:

Find $\boldsymbol{\zeta}_n^j \in \mathbb{R}^{2m_h+n_h}$, $j = 1, \ldots, k$, such that

$$\sum_{j=0}^{k} \mathscr{M}\boldsymbol{\zeta}_n^j + \frac{\tau_n}{2}\mathscr{A}\boldsymbol{\zeta}_n^i = \frac{\tau_n}{2}\Big\{F_n^i + \beta_i^c\big(F_n^0 - \mathscr{A}\boldsymbol{\zeta}_n^0\big)\Big\}, \quad i = 1, \ldots, k. \tag{11}$$

3.2 The Discontinuous Galerkin Method

The discontinuous Galerkin (dG) method applied to (4) leads to the following problem in I_n:

Given $\mathbf{u}_n^-$ with $\mathbf{u}_1^- = \mathbf{u}_{0,h}$, find $\mathbf{u}_{h,\tau}\big|_{I_n} \in \mathbb{P}_k(I_n, \mathbf{V}_h)$ and $p_{h,\tau}\big|_{I_n} \in \mathbb{P}_k(I_n, Q_h)$ such that for all $\psi \in \mathbb{P}_k(I_n)$

$$\int_{I_n} \Big[\, \big(\mathbf{u}'_{h,\tau}, \mathbf{v}_{h,\tau}\big) + A\left(\mathbf{u}_{h,\tau}, (\mathbf{u}_{h,\tau}, p_{h,\tau}), (\mathbf{v}_h, q_h)\right) \Big]\psi(t) + \big([\mathbf{u}_{h,\tau}]_n, \mathbf{v}_{n-1}^+\big)\,\psi(t_{n-1})$$
$$= \int_{I_n} (\mathbf{f}, \mathbf{v}_{h,\tau})\,\psi(t)$$

for all $\mathbf{v}_h \in \mathbf{V}_h$ and all $q_h \in Q_h$. Here, the right-sided Gauß-Radau quadrature formula with $(k+1)$ points is applied to evaluate the time integrals numerically. Note that this quadrature rule is exact for polynomials of degree less than or equal to $2k$. Let $\hat{t}_j$ and $\hat{w}_j$, $j = 1, \ldots, k+1$, denote the points and weights for this quadrature formula on $[-1, 1]$, respectively.

Since $\mathbf{u}_{h,\tau}$ and $p_{h,\tau}$ restricted to the interval I_n are $\mathbf{V}_h$-valued and Q_h-valued polynomials of degree less than or equal to k , they can be represented as

$$\mathbf{u}_{h,\tau}\big|_{I_n}(t) = \sum_{j=1}^{k+1} U_{n,h}^j \phi_{n,h}^j(t), \qquad p_{h,\tau}\big|_{I_n}(t) = \sum_{j=1}^{k+1} P_{n,h}^j \phi_{n,h}^j(t)$$

with $U_{n,h}^j \in \mathbf{V}_h$ and $P_{n,h}^j \in Q_h$, $j = 1, \ldots, k+1$. Following [2], one obtains the following coupled system of nonlinear equations:

Given $U_{n,h}^0 = \mathbf{u}_{0,h}$ for $n = 1$ and $U_{n,h}^0 = U_{n-1}^{k+1}$ for $n \geq 2$, find the coefficients $\left(U_{n,h}^j, P_{n,h}^j\right) \in \mathbf{V}_h \times Q_h, j = 1, \ldots, k+1$, such that

$$\sum_{j=1}^{k+1} \alpha_{i,j}^d \left(U_{n,h}^j, \mathbf{v}_h\right) + \frac{\tau_n}{2} A\left(U_{n,h}^i, (U_{n,h}^i, P_{n,h}^i), (\mathbf{v}_h, q_h)\right) = \beta_i \left(U_{n,h}^0, \mathbf{v}_h\right) + \frac{\tau_n}{2} \left(\mathbf{f}(t_{n,i}), \mathbf{v}_h\right) \tag{12}$$

for $i = 1, \ldots, k+1$ and for all $(\mathbf{v}_h, q_h) \in (\mathbf{V}_h, Q_h)$ where

$$\alpha_{i,j}^d := \hat{\phi}_j' + \beta_i^d \hat{\phi}_j(-1), \quad \beta_i^d := \frac{1}{\hat{w}_i} \hat{\phi}_i(-1).$$

Similarly as for cGP, problem (12) on I_n results in the $(k+1) \times (k+1)$ nonlinear algebraic block system:

Find $\boldsymbol{\xi}_n^j \in \mathbb{R}^{2m_h + n_h}$ for $j = 1, \ldots, k+1$ such that

$$\sum_{j=1}^{k+1} \alpha_{i,j}^d \mathscr{M} \boldsymbol{\xi}_n^j + \frac{\tau_n}{2} A \boldsymbol{\xi}_n^i = \beta_i^d \mathscr{M} \boldsymbol{\xi}_n^0 + \frac{\tau_n}{2} F_n^i. \tag{13}$$

After solving this system, we enter the next time interval and set the initial value of the time interval I_{n+1} to $\boldsymbol{\xi}_{n+1}^0 := \boldsymbol{\xi}_n^{k+1}$.

3.3 Post-Processing

In [9], a simple post-processing for systems of ordinary differential equations was presented which was extended to time-dependent convection-diffusion-reaction equations in [12] and to transient Stokes problems in [2]. This simple post-processing allows to construct numerical approximations being in integral-based norms at least one order better than the originally obtained numerical solution provided that the exact solution is sufficiently smooth in time.

We will generalize the idea to the Navier-Stokes equations. Let $\mathbf{u}_{h,\tau}$ and $p_{h,\tau}$ denote the solution of either cGP(k) or dG(k). The post-processed solution $(\Pi \mathbf{u}_{h,\tau}, \Pi p_{h,\tau})$ on the time interval I_n is given by

$$(\Pi \mathbf{u}_{h,\tau})(t) = \mathbf{u}_{h,\tau}(t) + g_n \zeta_n(t), \quad (\Pi p_{h,\tau})(t) = p_{h,\tau}(t) + d_n \zeta_n'(t), \qquad t \in I_n,$$

where $g_n \in \mathbf{V}_h$ and $d_n \in Q_h$ are finite element functions and

$$\zeta_n(t) = \frac{\tau_n}{2} \hat{\zeta}(\hat{t}), \quad \hat{t} := T_n^{-1}(t),$$

with T_n from (6). For cGP(k), the polynomial $\hat{\zeta} \in \mathbb{P}_{k+1}$ vanishes in all Gauß-Lobatto points while the polynomial $\hat{\zeta} \in \mathbb{P}_{k+1}$ for dG(k) vanishes in all Gauß-Radau points. In both cases, it is scaled such that $\hat{\zeta}'(1) = 1$. The nodal vectors $\gamma_{n,1} \in \mathbb{R}^{m_h}$, $\gamma_{n,2} \in \mathbb{R}^{m_h}$ of the finite element function $g_n \in \mathbf{V}_h$ and the nodal vector $\delta_n \in \mathbb{R}^{n_h}$ of the finite element function $d_n \in Q_h$ are the solution of the saddle-point problem

$$\begin{bmatrix} M & 0 & B_1^T \\ 0 & M & B_2^T \\ B_1 & B_2 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_{n,1} \\ \boldsymbol{\gamma}_{n,2} \\ \boldsymbol{\delta}_n \end{bmatrix} = \begin{bmatrix} F_{n,1}^e \\ F_{n,2}^e \\ 0 \end{bmatrix} - \begin{bmatrix} A + N(\xi_n) & 0 & B_1^T \\ 0 & A + N(\xi_n) & B_2^T \\ B_1 & B_2 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{n,1}^e \\ \boldsymbol{\xi}_{n,2}^e \\ \boldsymbol{\eta}_n^e \end{bmatrix} - \begin{bmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\chi}_{n,1}^e \\ \boldsymbol{\chi}_{n,2}^e \\ 0 \end{bmatrix} \tag{14}$$

where $\boldsymbol{\chi}_{n,1}^e$, $\boldsymbol{\chi}_{n,2}^e \in \mathbb{R}^{m_h}$ denote the nodal representation of the components of $\mathbf{u}_{h,\tau}'(t_n) \in \mathbf{V}_h$ while $\xi_n^e = (\xi_{n,1}^e, \xi_{n,2}^e)^T$ and η_n^e are the nodal vectors for $\mathbf{u}_{h,\tau}(t_n)$ and $p_{h,\tau}$, respectively. The matrices are given in (8) and (9).

It has been shown in [9] for systems of ordinary equations that the post-processed solution $\Pi \mathbf{u}_{h,\tau}(t)$ can be interpreted as the solution of a time stepping scheme with ansatz order $k + 1$. The extension to the transient Stokes problems and transient Oseen problems can be found in [2] and [13]. It has been shown numerically that the simple post-processing leads to solutions which show at the discrete time points a super-convergence of order $2k$ (cGP(k)) and $2k + 1$ (dG(k)) for both velocity and pressure. Note that the post-processing requires, even for the Navier–Stokes equations, just the solution of a linear saddle point system. The post-processing for the three-dimensional case is obtained in the obvious way.

4 Numerical Results

This section is devoted to an example which illustrates accuracy and performance of combinations of inf-sup stable spatial discretizations with higher order variational time discretization schemes. All computations used the finite element code MooNMD [14].

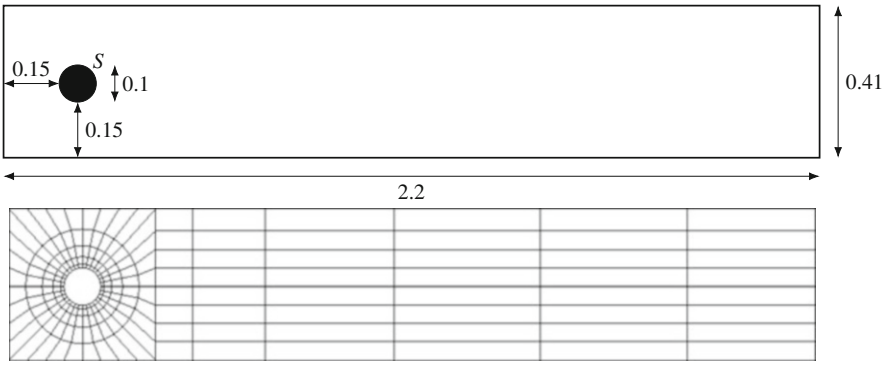

Fig. 1 Domain (*top*) and initial mesh (*bottom*) of the test problem

We consider the well-known benchmark problem of the flow around a circle defined in [15]. The geometry and the initial grid (level 0) are given in Fig. 1.

The Navier-Stokes equations (1) are considered with source term $\mathbf{f} = 0$, viscosity $\nu = 10^{-3}$, and the final time $T = 8$. The inflow and outflow boundary conditions are prescribed by

$$\mathbf{u}(t; 0, y) = \mathbf{u}(t; 2.2, y) = \frac{1}{0.41^2} \sin\left(\frac{\pi t}{8}\right) \begin{pmatrix} 6y(0.41 - y) \\ 0 \end{pmatrix}, \qquad 0 \le y \le 0.41,$$

while no-slip conditions are applied on all other boundaries. The diameter of the cylinder is $L = 0.1$ and the mean inflow velocity is $U(t) = \sin(\pi t/8)$ such that $U_{\max} = 1$. The density of the fluid is $\rho = 1$. Hence, the Reynolds number of this flow is $Re = 100$. Note that the standard Galerkin discretization in space is used since the moderate Reynolds number doesn't require a spatial stabilization.

Important quantities of interest in this example are the drag coefficient c_d at the circle and the lift coefficient c_l at the circle which are defined by

$$c_d(t) := \frac{2}{\rho L U_{\max}^2} \int_S \left(\rho \nu \frac{\partial u_{t_S}(t)}{\partial \mathbf{n}} n_y - p(t) n_x \right) dS,$$

$$c_l(t) := -\frac{2}{\rho L U_{\max}^2} \int_S \left(\rho \nu \frac{\partial u_{t_S}(t)}{\partial \mathbf{n}} n_x + p(t) n_y \right) dS,$$

where $\mathbf{n} = (n_x, n_y)^T$ is the unit normal vector on S directing into Ω, $\mathbf{t}_S = (n_y, -n_x)^T$ the unit tangential vector and $u_{t_S} := \mathbf{u} \cdot \mathbf{t}_S$ the tangential velocity along the circle. Using integration by parts and the weak formulation (4) of the Navier–Stokes equations, we get

$$c_d(t) = -20\big\{(\mathbf{u}_t, \mathbf{v}_d) + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}_d) + \big((\mathbf{u} \cdot \nabla)\mathbf{u}, \mathbf{v}_d\big) - (p, \nabla \cdot \mathbf{v}_d)\big\}$$

for any function $\mathbf{v}_d \in \left(H^1(\Omega)\right)^2$ with $(\mathbf{v}_d)\big|_S = (1,0)^T$ and $\mathbf{v}_d = (0,0)^T$ on all other boundaries. Similarly, the lift coefficient can be obtained by

$$c_l(t) = -20\big\{(\mathbf{u}_t, \mathbf{v}_l) + \nu(\nabla\mathbf{u}, \nabla\mathbf{v}_l) + \big((\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{v}_l\big) - (p, \nabla\cdot\mathbf{v}_l)\big\}$$

with any $\mathbf{v}_l \in \left(H^1(\Omega)\right)^2$ as a test function such that $\mathbf{v}_l\big|_S = (0,1)^T$ and $\mathbf{v}_l = (0,0)^T$ on all other boundaries.

The third benchmark parameter is the pressure difference between the front and the back of the circle, given by

$$\Delta p(t) = p(t; 0.15, 0.2) - p(t; 0.25, 0.2).$$

The Navier-Stokes equations were discretized in space with the inf-sup stable pairs Q_2/P_1^{disc} and Q_3/P_2^{disc} on quadrilateral meshes. They are obtained from the coarsest mesh (level 0) given in Fig. 1 by regular refinement with boundary adaption to take the curved boundary at the circle into consideration. The computations were performed on mesh level 4. This results in $107,712$ degrees of freedom for the velocity and $39,936$ pressure degrees of freedom for Q_2/P_1^{disc} while there are $241,440$ velocity degrees of freedom and $79,872$ pressure degrees of freedom for Q_3/P_2^{disc}. The temporal discretizations cGP$(k+1)$ and dG(k) lead both to a single $(k+1)\times(k+1)$ block system of nonlinear equations in each time step. The computations were performed with the time step lengths $\tau = 0.02\times 2^{-j}, j = 1,\ldots,4$. The nonlinearity is resolved by a Picard iteration (fixed point iteration) and the resulting linear systems were solved by a flexible GMRES method [16] where coupled multigrid methods with Vanka-type smoothers were used as preconditioner.

The accuracy is measured with respect to the reference values

$$(t_{d,\max}^{\text{ref}}, c_{d,\max}^{\text{ref}}) = (3.93625, 2.950921575),\ (t_{l,\max}^{\text{ref}}, c_{l,\max}^{\text{ref}})$$
$$(t_{l,\max}^{\text{ref}}, c_{l,\max}^{\text{ref}}) = (5.693125, 0.47795),\ \Delta p^{\text{ref}}(8) = -0.1116$$

given in [17] where $t_{d,\max}^{\text{ref}}$ and $t_{l,\max}^{\text{ref}}$ denote the times at which drag and lift coefficients achieve their maximal values $c_{d,\max}^{\text{ref}}$ and $c_{l,\max^{\text{ref}}}$, respectively. We compute the error to the reference values with respect to the drag and lift coefficients by the distance formula

$$\text{err}_d = \sqrt{(t_{d,\max}^{\text{ref}} - t_{d,\max})^2 + (c_{d,\max}^{\text{ref}} - c_{d,\max})^2},$$
$$\text{err}_l = \sqrt{(t_{l,\max}^{\text{ref}} - t_{l,\max})^2 + (c_{l,\max}^{\text{ref}} - c_{l,\max})^2},$$

see [18]. The error for the pressure difference will be computed by the simple distance to the reference value.

All numbers which will be presented in the following graphs are based on post-processed velocity and post-processed pressure.

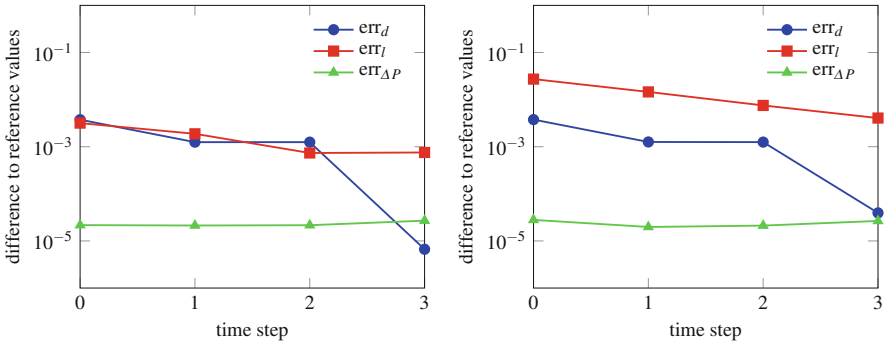

Fig. 2 Difference to the reference values vs the time step lengths for the cGP(2) (*left*) and dG(1) (*right*) methods combined with the finite element pair Q_2/P_1^{disc}

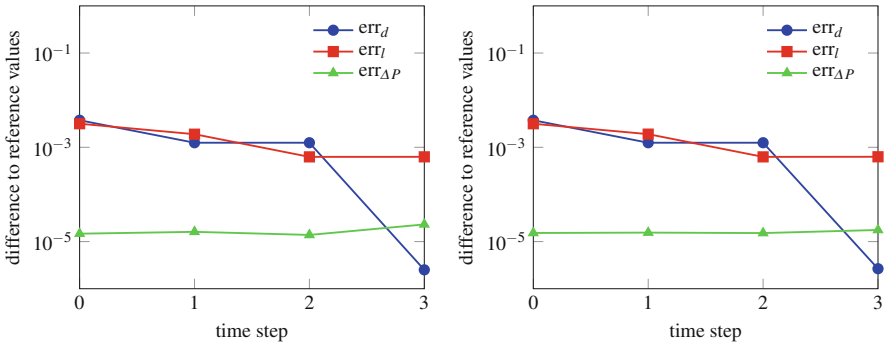

Fig. 3 Difference to the reference values vs the time step lengths for the cGP(3) (*left*) and dG(2) (*right*) methods combined with the finite element pair Q_3/P_2^{disc}

Results for the time stepping schemes cGP(2) and dG(1) in combination with the Q_2/P_1^{disc} finite element pair are plotted in Fig. 2. We observe from the simulations that the behavior concerning the accuracy and efficiency is different for different quantities of interest. For the drag coefficient, the best result for both time discretization methods can be obtained by using the time step length $\tau = 0.00125$. For the lift coefficient, dG(1) needs a smaller time step length than cGP(2), see the left plot in Fig. 2. However, the results for the pressure difference are almost independent of the time step length. Comparing the results for both methods, cGP(2) shows the best combination of efficiency and accuracy.

In Fig. 3, the differences to reference values for the combination of cGP(3) and dG(2) with the pair Q_3/P_2^{disc} are plotted. Similar conclusions can be made as for the combination of cGP(2) and dG(1) with the pair Q_2/P_1^{disc} if the drag coefficient is of main interest. However, both time discretization methods perform similar for the lift coefficient and pressure difference. Moreover, it is observed that the time error

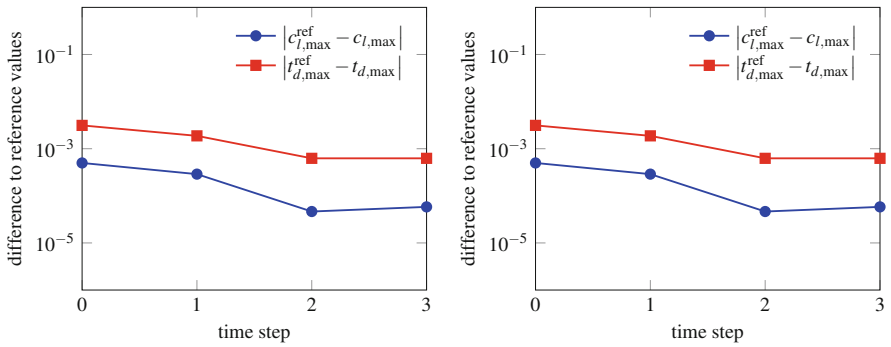

Fig. 4 Difference to the reference maximum time and lift values vs the time step lengths for the cGP(3) (*left*) and dG(2) (*right*) methods combined with the finite element pair Q_3/P_2^{disc}

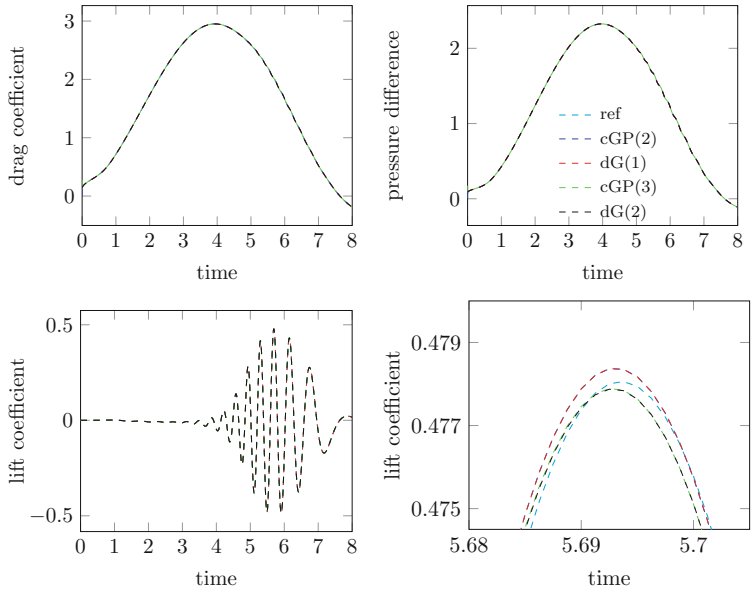

Fig. 5 Evaluation of the drag coefficient (*top left*), pressure difference (*top right*), lift coefficient (*bottom left*), and zoom of the lift coefficient around $(t_{l,\max}^{\text{ref}}, c_{l,\max}^{\text{ref}})$ (*bottom right*)

is dominant in the computations of the error of the lift coefficient. This can be seen in Fig. 4 where the difference to the reference time $t_{l,\max}^{\text{ref}}$ and value $c_{l,\max}^{\text{ref}}$ are plotted.

Figure 5 shows for the four considered combinations of spatial and temporal discretizations the drag and lift coefficients as well as the pressure difference as a function of time. The corresponding reference curves from [17] are also given in all plots. If the drag coefficient and pressure difference are of concern, all methods produce similarly accurate results. Considering the accuracy of the lift coefficient, the situation is considerably more delicate. The higher order methods

cGP(3) and dG(2), both in combination with the higher order pair Q_3/P_2^{disc} as spatial discretization, generate values which are closer to the reference data than the results obtained for cGP(2) and dG(1), both together with Q_2/P_1^{disc} as discretization in space. This can be seen in a zoom of the lift coefficient around $(l_{l,\max}^{\text{ref}}, c_{l,\max}^{\text{ref}})$, shown in the right bottom picture of Fig. 5. The results of the four discretizations suggest that the behavior of the lift coefficient is much more influenced by the spatial discretization than the variational time discretization.

References

1. Aziz, A.K., Monk, P.: Continuous finite elements in space and time for the heat equation. Math. Comput. **52**(186), 255–274 (1989)
2. Ahmed, N., Becher, S., Matthies, G.: Higher-order discontinuous Galerkin time stepping and local projection stabilization techniques for the transient Stokes problem. Comput. Methods Appl. Mech. Eng. **313**(1), 28–52 (2017)
3. Hussain, S., Schieweck, F., Turek, S.: Higher order Galerkin time discretization for nonstationary incompressible flow. In: Numerical Mathematics and Advanced Applications 2011, pp. 509–517. Springer, Heidelberg (2013)
4. Hussain, S., Schieweck, F., Turek, S.: A note on accurate and efficient higher order Galerkin time stepping schemes for the nonstationary Stokes equations. Open Numer. Methods J. **4**, 35–45 (2012)
5. Hussain, S., Schieweck, F., Turek, S.: An efficient and stable finite element solver of higher order in space and time for nonstationary incompressible flow. Int. J. Numer. Methods Fluids **73**(11), 927–952 (2013)
6. Hussain, S., Schieweck, F., Turek, S.: Efficient Newton-multigrid solution techniques for higher order space-time Galerkin discretizations of incompressible flow. Appl. Numer. Math. **83**, 51–71 (2014)
7. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, 2nd edn. Springer Series in Computational Mathematics, vol. 25. Springer-Verlag, Berlin (2006)
8. Matthies, G., Tobiska, L.: Mass conservation of finite element methods for coupled flow-transport problems. Int. J. Comput. Sci. Math. **1**(2–4), 293–307 (2007)
9. Matthies, G., Schieweck, F.: Higher order variational time discretizations for nonlinear systems of ordinary differential equations, Preprint 23/2011, Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg (2011)
10. Ahmed, N., Chacón Rebollo, T., John, V., Rubino, S.: A review of variational multiscale methods for the simulation of turbulent incompressible flows. Arch. Comput. Methods Eng. **24**(1), 115–164 (2017)
11. Ahmed, N., Chacón Rebollo, T., John, V., Rubino, S.: Analysis of a full space-time discretization of the navier-stokes equations by a local projection stabilization method. IMA J. Numer. Anal. **37**, 1437–1467 (2017)
12. Ahmed, N., Matthies, G.: Higher order continuous Galerkin-Petrov time stepping schemes for transient convection-diffusion-reaction equations. ESAIM Math. Model. Numer. Anal. **49**(5), 1429–1450 (2015)
13. Ahmed, N., Matthies, G.: Numerical studies of variational-type time-discretization techniques for transient Oseen problem. In: Algoritmy 2012. 19th Conference on Scientific Computing, Vysoké Tatry, Podbanské, Slovakia, September 9–14, 2012. Proceedings of Contributed Papers and Posters. Slovak University of Technology, Faculty of Civil Engineering, Department of Mathematics and Descriptive Geometry, Bratislava, pp. 404–415. (2012)

14. John, V., Matthies, G.: MooNMD—a program package based on mapped finite element methods. Comput. Vis. Sci. **6**(2–3), 163–169 (2004)
15. Turek, S., Schäfer, M.: Benchmark computations of laminar flow around cylinder. In: Hirschel, E. (ed.) Flow Simulation with High-Performance Computers II, vol. 52, pp. 547–566. Vieweg, Braunschweig (1996)
16. Saad, Y.: A flexible inner-outer preconditioned GMRES algorithm. SIAM J. Sci. Comput. **14**(2), 461–469 (1993)
17. John, V.: Reference values for drag and lift of a two-dimensional time dependent flow around a cylinder. Int. J. Numer. Methods Fluids **44**, 777–788 (2004)
18. John, V., Rang, J.: Adaptive time step control for the incompressible navier-stokes equations. Comput. Methods Appl. Mech. Engrg. **199**, 514–524 (2010)

Uniform Convergent Monotone Iterates for Nonlinear Parabolic Reaction-Diffusion Systems

Igor Boglaev

Abstract This paper deals with a uniform convergent monotone method for solving nonlinear singularly perturbed parabolic reaction-diffusion systems. The uniform convergence on a piecewise uniform mesh is established. Numerical experiments are presented.

1 Introduction

In this paper we give a numerical treatment for the following semi-linear singularly perturbed parabolic system:

$$\frac{\partial u_i}{\partial t} - \varepsilon_i \frac{\partial^2 u_i}{\partial x^2} + f_i(x,t,u) = 0, \quad (x,t) \in \omega \times (0,T], \tag{1}$$

$$u_i(0,t) = 0, \quad u_i(1,t) = 0, \quad t \in [0,T],$$

$$u_i(x,0) = \psi_i(x), \quad x \in \overline{\omega}, \quad \omega = (0,1), \quad i = 1,2,$$

where $0 < \varepsilon_1 \le \varepsilon_2 \le 1$, $u \equiv (u_1, u_2)$, the functions f_i and ψ_i, $i = 1,2$, are smooth in their respective domains.

In the study of numerical methods for nonlinear singularly perturbed problems, the two major points to be developed are: (1) constructing robust difference schemes (this means that unlike classical schemes, the error does not increase to infinity, but rather remains bounded, as the small parameters approach zero); (2) obtaining reliable and efficient computing algorithms for solving nonlinear discrete problems. For solving these nonlinear discrete systems, the iterative approach presented in this paper is based on the method of upper and lower solutions and associated monotone iterates. The basic idea of the method of upper and lower solutions is the construction of two monotone sequences which converge monotonically from above

I. Boglaev (✉)
Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
e-mail: I.Boglaev@massey.ac.nz

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_3

and below to a solution of the problem. The monotone property of the iterations gives improved upper and lower bounds of the solution in each iteration. An initial iteration in the monotone iterative method is either an upper or lower solution, which can be constructed directly from the difference equation, this method simplifies the search for the initial iteration as is often required in Newton's method.

In [5], uniformly convergent numerical methods for solving linear singularly perturbed systems of type (1) were constructed. These uniform numerical methods are based on the piecewise uniform meshes of Shishkin-type [6].

In [2], we investigated uniform convergence properties of the monotone iterative method for solving scalar nonlinear singularly perturbed problems of type (1). In this paper, we extend our investigation to the case of the nonlinear singularly perturbed system (1).

The structure of the paper as follows. In Sect. 2, we introduce a nonlinear difference scheme for solving (1). The monotone iterative method is presented in Sect. 3. An analysis of the uniform convergence of the monotone iterates to the solution of the nonlinear difference scheme and to the solution of (1) is given in Sect. 4. The final Sect. 5 presents the results of numerical experiments with a gas-liquid interaction model.

2 The Nonlinear Difference Scheme

On $\overline{\omega} = [0, 1]$ and $[0, T]$, we introduce meshes $\overline{\omega}^h$ and $\overline{\omega}^\tau$:

$$\overline{\omega}^h = \{x_m, 0 \le m \le M_x; x_0 = 0, x_{M_x} = 1; h_m = x_{m+1} - x_m\},$$

$$\overline{\omega}^\tau = \{t_k, 0 \le k \le N_\tau; t_0 = 0, t_{N_\tau} = T; \tau_k = t_k - t_{k-1}\},$$

and consider the nonlinear implicit difference scheme

$$\mathcal{L}_i U_i(x_m, t_k) + f_i(x_m, t_k, U) - \tau_k^{-1} U_i(x_m, t_{k-1}) = 0, \quad (x_m, t_k) \in \omega^h \times \omega^\tau, \tag{2}$$

$$\mathcal{L}_i U_i(x_m, t_k) \equiv -\varepsilon_i \mathcal{L}_i^h U_i(x_m, t_k) + \tau_k^{-1} U_i(x_m, t_k).$$

$$U_i(x_0, t_k) = U_i(x_{M_x}, t_k) = 0, \quad U_i(x_m, 0) = \psi_i(x_m), \quad x_m \in \overline{\omega}^h, \quad i = 1, 2,$$

where $U \equiv (U_1, U_2)$, and the difference operators $\mathcal{L}_i^h$, $i = 1, 2$, are defined by

$$\mathcal{L}_i^h U_i(x_m, t_k) = \left[\frac{U_i(x_{m+1}, t_k) - U_i(x_m, t_k)}{\hbar_m h_m} - \frac{U_i(x_m, t_k) - U_i(x_{m-1}, t_k)}{\hbar_m h_{m-1}} \right],$$

$$\hbar_m = (h_m + h_{m-1})/2, \quad i = 1, 2.$$

On each time level t_k, $k \geq 1$, we introduce the linear problems

$$(\mathcal{L}_i + c_i)W_i(x_m, t_k) = \Phi_i(x_m, t_k), \quad W_i(x_0, t_k) = W_i(x_{M_x}, t_k) = 0, \tag{3}$$

$$c_i(x_m, t_k) \geq 0, \quad x_m \in \omega^h, \quad i = 1, 2.$$

In the following lemma, we state the maximum principle and we give estimates on solutions of (3) from [8].

Lemma 1

(i) If mesh functions $W_i(x_m, t_k)$, $i = 1, 2$, satisfy the conditions

$$(\mathcal{L}_i + c_i)W_i(x_m, t_k) \geq 0 \ (\leq 0), \quad x_m \in \omega^h,$$

$$W_i(x_0, t_k) \geq 0 \ (\leq 0), \quad W_i(x_{M_x}, t_k) \geq 0 \ (\leq 0),$$

then $W_i(x_m, t_k) \geq 0 \ (\leq 0)$ in $\overline{\omega}^h$, $i = 1, 2$.

(ii) The following estimates on the solutions of (3) hold true

$$\|W_i(\cdot, t_k)\|_{\overline{\omega}^h} \leq \max_{x_m \in \omega^h} \left\{ \frac{|\Phi_i(x_m, t_k)|}{c_i(x_m, t_k) + \tau_k^{-1}} \right\}, \quad i = 1, 2, \tag{4}$$

where $\|W_i(\cdot, t_k)\|_{\overline{\omega}^h} = \max_{x_m \in \overline{\omega}^h} |W_i(x_m, t_k)|$.

3 The Monotone Iterative Method

We say that the mesh functions

$$\widetilde{U}(x_m, t_k) = (\widetilde{U}_1(x_m, t_k), \widetilde{U}_2(x_m, t_k)), \quad \widehat{U}(x_m, t_k) = (\widehat{U}_1(x_m, t_k), \widehat{U}_2(x_m, t_k))$$

are ordered upper and lower solutions if they satisfy the following inequalities:

$$\widetilde{U}(x_m, t_k) \geq \widehat{U}(x_m, t_k), \quad (x_m, t_k) \in \overline{\omega}^h \times \omega^\tau,$$

$$\mathcal{L}_i \widetilde{U}_i(x_m, t_k) + f_i(x_m, t_k, \widetilde{U}) - \tau_k^{-1} \widetilde{U}_i(x_m, t_{k-1}) \geq 0, \quad (x_m, t_k) \in \omega^h \times \omega^\tau,$$

$$\mathcal{L}_i \widehat{U}_i(x_m, t_k) + f_i(x_m, t_k, \widehat{U}) - \tau_k^{-1} \widehat{U}_i(x_m, t_{k-1}) \leq 0, \quad (x_m, t_k) \in \omega^h \times \omega^\tau,$$

$$\widehat{U}_i(x_*, t_k) \leq 0 \leq \widetilde{U}_i(x_*, t_k), \quad x_* = x_0, x_{M_x},$$

$$\widehat{U}_i(x_m, 0) \leq \psi_i(x_m) \leq \widetilde{U}_i(x_m, 0), \quad x_m \in \overline{\omega}^h, \quad i = 1, 2.$$

We introduce the notation

$$\langle \widehat{U}(t_k), \widetilde{U}(t_k)\rangle = \{U(x_m,t_k) : \widehat{U}(x_m,t_k) \le U(x_m,t_k) \le \widetilde{U}(x_m,t_k),\ x_m \in \overline{\omega}^h\},$$

and we assume that on each time level t_k, $k \ge 1$, the reaction functions satisfy the assumptions

$$0 \le \frac{\partial f_i}{\partial u_i}(x_m,t_k,U) \le c_i(x_m,t_k), \quad \text{on } \langle \widehat{U}(t_k), \widetilde{U}(t_k)\rangle, \tag{5}$$

$$0 \le -\frac{\partial f_i}{\partial u_{i'}}(x_m,t_k,U) \le q_i(x_m,t_k), \quad \text{on } \langle \widehat{U}(t_k), \widetilde{U}(t_k)\rangle, \quad i' \ne i,$$

where $c_i(x_m,t_k)$ and $q_i(x_m,t_k)$, $i = 1,2$, are nonnegative bounded functions in $\overline{\omega}^h$.

On each time level t_k, $k \ge 1$, the iterative method is given in the form

$$(\mathcal{L}_i + c_i) Z_i^{(n)}(x_m,t_k) = -\mathcal{R}_i(x_m,t_k,U^{(n-1)}), \quad x_m \in \omega^h, \tag{6}$$

$$\mathcal{R}_i(x_m,t_k,U^{(n-1)}) \equiv \mathcal{L}_i U_i^{(n-1)}(x_m,t_k) + f_i(x_m,t_k,U^{(n-1)}) - \tau_k^{-1} U_i(x_m,t_{k-1}),$$

$$Z_i^{(n)}(x_*,t_k) = 0, \quad n \ge 1, \quad x_* = x_0, x_{M_x},$$

$$Z_i^{(n)}(x_m,t_k) \equiv U_i^{(n)}(x_m,t_k) - U_i^{(n-1)}(x_m,t_k),$$

$$U_i(x_m,0) = \psi_i(x_m), \quad x_m \in \overline{\omega}^h, \quad i = 1,2,$$

where c_i, $i = 1,2$, are defined in (5). For upper sequence, we have $\overline{U}_i(x_m,0) = \psi_i(x_m)$, $\overline{U}^{(0)}(x_m,t_k) = \widetilde{U}_i(x_m,t_k)$ and $\overline{U}_i(x_m,t_k) = \overline{U}_i^{(n_k)}(x_m,t_k)$, $i = 1,2$, $x_m \in \overline{\omega}^h$, where $\overline{U}_i(x_m,t_k)$, $i = 1,2$, are approximations of the exact solutions on time level t_k and n_k is a number of iterative steps on time level t_k. For lower sequence, we have $\underline{U}_i(x_m,0) = \psi_i(x_m)$, $\underline{U}^{(0)}(x_m,t_k) = \widehat{U}_i(x_m,t_k)$ and $\underline{U}(x_m,t_k) = \underline{U}^{(n_k)}(x_m,t_k)$, $i = 1,2$, $x_m \in \overline{\omega}^h$.

The following theorem gives the monotone property of the iterative method (6).

Theorem 1 *Let $\widetilde{U}$ and $\widehat{U}$ be ordered upper and lower solutions, and assumption (5) be satisfied. On each time level t_k, $k \ge 1$, the sequences $\{\overline{U}^{(n)}\}$, $\{\underline{U}^{(n)}\}$ with $\overline{U}^{(0)} = \widetilde{U}$ and $\underline{U}^{(0)} = \widehat{U}$, generated by the iterative method (6), converge monotonically*

$$\underline{U}^{(n-1)}(x_m,t_k) \le \underline{U}^{(n)}(x_m,t_k) \le \overline{U}^{(n)}(x_m,t_k) \le \overline{U}^{(n-1)}(x_m,t_k), \quad x_m \in \overline{\omega}^h, \tag{7}$$

Proof Since $\overline{U}^{(0)} = \widetilde{U}$ and $\underline{U}^{(0)} = \widehat{U}$, then from (6) we conclude that

$$(\mathcal{L}_i + c_i)\overline{Z}_i^{(1)}(x_m,t_1) \le 0, \quad (\mathcal{L}_i + c_i)\underline{Z}_i^{(1)}(x_m,t_1) \ge 0, \quad x_m \in \omega^h,$$

$$\overline{Z}_i^{(1)}(x_*,t_1) \le 0, \quad \underline{Z}_i^{(1)}(x_*,t_1) \ge 0, \quad x_* = x_0, x_{M_x}, \quad i = 1,2.$$

From Lemma 1, it follows that

$$\overline{Z}_i^{(1)}(x_m,t_1)\le 0,\quad \underline{Z}_i^{(1)}(x_m,t_1)\ge 0\quad x_m\in\overline{\omega}^h,\quad i=1,2. \tag{8}$$

We now prove (7) for $n=1$ and $k=1$. From (6), in the notation $W_i^{(n)}=\overline{U}_i^{(n)}-\underline{U}_i^{(n)}$, $n\ge 0$, $i=1,2$, we conclude that

$$(\mathcal{L}_i+c_i)W_i^{(1)}(x_m,t_1)=F_i(x_m,t_1,\overline{U}^{(0)})-F_i(x_m,t_1,\underline{U}^{(0)}),\quad x_m\in\omega^h,$$

$$W_i^{(1)}(x_*,t_1)=0,\quad x_*=x_0,x_{M_x},\quad i=1,2,$$

where $F_i(x_k,t_k,U)=c_i(x_m,t_k)U_i(x_m,t_k)-f_i(x_m,t_k,U)$. Since $\overline{U}^{(0)}(x_m,t_1)\ge\underline{U}^{(0)}(x_m,t_1)$, by Lemma 2 from [1], we conclude that the right hand sides in the difference equations are nonnegative. From Lemma 1, it follows $W_i^{(1)}(p,t_1)\ge 0$, $i=1,2$, and this leads to (7) for $n=1$, $k=1$.

Using the mean-value theorem, from (6) we obtain

$$\mathcal{R}_i(x_m,t_1,\overline{U}^{(1)})=-\left(c_i-\frac{\partial f_i}{\partial u_i}\right)\overline{Z}_i^{(1)}(x_m,t_1)+\frac{\partial f_i}{\partial u_{i'}}\overline{Z}_{i'}^{(1)}(x_m,t_1),\quad i'\ne i, \tag{9}$$

where the partial derivatives are calculated at intermediate points which lie in the sector $\langle\overline{U}^{(1)}(t_1),\overline{U}^{(0)}(t_1)\rangle$. From (5) and (8), we conclude that

$$\mathcal{R}_i(x_m,t_1,\overline{U}^{(1)})\ge 0,\quad x_m\in\omega^h,\quad \overline{U}_i^{(1)}(x_*,t_1)=0,\quad x_*=x_0,x_{M_x},\ i=1,2.$$

Thus, $\overline{U}^{(1)}(x_m,t_1)$ is an upper solution. Similarly, we prove that $\underline{U}^{(1)}(x_m,t_1)$ is a lower solution. By induction on n, we can prove that $\{\overline{U}^{(n)}(x_m,t_1)\}$ and $\{\underline{U}^{(n)}(p,t_1)\}$ are, respectively monotonically decreasing and monotonically increasing sequences.

From (7) with t_1, it follows that for $i=1,2$,

$$\widehat{U}_i(x_m,t_1)\le\underline{U}_i^{(n_1)}(x_m,t_1)\le\overline{U}_i^{(n_1)}(x_m,t_1)\le\widetilde{U}_i(x_m,t_1),\quad x_m\in\overline{\omega}^h. \tag{10}$$

From here and by the assumption of the theorem that $\widetilde{U}(p,t_2)$ and $\widehat{U}(p,t_2)$ are, respectively, upper and lower solutions, we conclude that $\widetilde{U}(x_m,t_2)$ and $\widehat{U}(x_m,t_2)$ are upper and lower solutions with respect to $\overline{U}^{(n_1)}(x_m,t_1)$ and $\underline{U}^{(n_1)}(x_m,t_1)$.

From (6), we conclude that $W^{(1)}(x_m,t_2)$ satisfies

$$(\mathcal{L}_i+c_i)W_i^{(1)}(x_m,t_2)=F_i(x_m,t_2,\overline{U}^{(0)})-F_i(x_m,t_2,\underline{U}^{(0)})+\tau_2^{-1}[\overline{U}_i^{(n_1)}(x_m,t_1)-\underline{U}_i^{(n_1)}(x_m,t_1)],$$

$$x_m\in\omega^h,\quad W_i^{(1)}(x_*,t_2)=0,\quad x_*=x_0,x_{M_x},\quad i=1,2.$$

Since $\overline{U}^{(0)}(x_m,t_2) \geq \underline{U}^{(0)}(x_m,t_2)$ and taking into account (10), by Lemma 2 from [1], we conclude that the right hand sides in the difference equations are nonnegative. From Lemma 1, we have $W_i^{(1)}(p,t_2) \geq 0$, $i=1,2$, that is,

$$\underline{U}_i^{(1)}(p,t_2) \leq \overline{U}_i^{(1)}(p,t_2), \quad p \in \overline{\omega}^h, \quad i=1,2.$$

The proof that $\overline{U}_i^{(1)}(x_m,t_2)$ and $\underline{U}_i^{(1)}(x_m,t_2)$, $i=1,2$, are, respectively, upper and lower solutions is similar to the proof on the time level t_1. By induction on n, we can prove that $\{\overline{U}^{(n)}(x_m,t_2)\}$ and $\{\underline{U}^{(n)}(x_m,t_2)\}$ are, respectively, monotonically decreasing and monotonically increasing sequences.

By induction on k, $k \geq 1$, we prove that $\{\overline{U}^{(n)}(x_m,t_k)\}$ and $\{\underline{U}^{(n)}(p,t_k)\}$ are, respectively, monotonically decreasing and monotonically increasing sequences, which satisfy (7).

3.1 Convergence on $[0,T]$

We now choose the stopping criterion of the iterative method (6) in the form

$$\max_i \|\mathcal{R}_i(\cdot,t_k,U^{(n)})\|_{\omega^h} \leq \delta, \tag{11}$$

where δ is a prescribed accuracy, and $U(x_m,t_k) = U^{(n_k)}(x_m,t_k)$, $x_m \in \overline{\omega}^h$, where n_k is minimal subject to the stopping test.

Instead of (5), we now impose the two-sided constraints on f_i, $i=1,2$, in the form

$$\rho_k \leq \frac{\partial f_i}{\partial u_i}(x_m,t_k,U) \leq c_i(x_m,t_k), \quad \text{on } \langle \widehat{U}(t_k), \widetilde{U}(t_k)\rangle, \tag{12}$$

$$0 \leq -\frac{\partial f_i}{\partial u_{i'}}(x_m,t,U) \leq q_i(x_m,t_k), \quad \text{on } \langle \widehat{U}(t_k), \widetilde{U}(t_k)\rangle, \quad i \neq i',$$

where ρ_k, $k \geq 1$, are defined in (13).

Remark 1 We mention that the assumption $\partial f_i/\partial u_i \geq \rho_k$, $i=1,2$, in (12) can always be obtained via a change of variables. Indeed, introduce the following functions $u_i(x,t) = \exp(\lambda t) z_i(x,t)$, $i=1,2$, where λ is a constant. Now, $z_i(x,t)$, $i=1,2$, satisfy (1) with

$$\varphi_i = \lambda z_i + \exp(-\lambda t) f_i(x,t,\exp(\lambda t)z_1, \exp(\lambda t)z_2),$$

instead of f_i, $i = 1, 2$, and we have

$$\frac{\partial \varphi_i}{\partial z_i} = \lambda + \frac{\partial f_i}{\partial u_i}, \quad \frac{\partial \varphi_i}{\partial z_{i'}} = \frac{\partial f_i}{\partial u_{i'}}, \quad i' \neq i, \quad i = 1, 2.$$

Thus, if $\lambda \geq \max_{k \geq 1} \rho_k$, from here, we conclude that $\partial \varphi_i / \partial z_i$ and $\partial \varphi_i / \partial z_{i'}$ satisfy (12)

We impose the constraint on τ_k

$$\tau_k < \frac{1}{\rho_k}, \quad \rho_k = \max_i \{ \max_{x_m \in \overline{\omega}^h} [q_i(x_m, t_k)] \}. \tag{13}$$

If assumptions (12) and (13) hold, then the nonlinear difference scheme (2) has a unique solution (see Lemmas 3 and 4 in [1] for details).

We prove the following convergence result for the iterative method (6), (11).

Theorem 2 *Assume that the mesh $\overline{\omega}^\tau$ satisfies (13), and $f_i(p, t, U)$, $i = 1, 2$, satisfy (12), where $\widetilde{U}$ and $\widehat{U}$ are ordered upper and lower solutions of (2). Then for the sequences $\{\overline{U}^{(n)}\}$, $\{\underline{U}^{(n)}\}$, generated by (6), (11) with, respectively, $\overline{U}^{(0)} = \widetilde{U}$ and $\underline{U}^{(0)} = \widehat{U}$, the following uniform in ε estimate holds*

$$\max_i \left[\max_{t_k \in \overline{\omega}^\tau} \| U_i(\cdot, t_k) - U_i^*(\cdot, t_k) \|_{\overline{\omega}^h} \right] \leq T\delta, \tag{14}$$

where $U_i^*(p, t_k)$, $i = 1, 2$, is the unique solution to (2).

Proof The difference problem for $U(x_m, t_k) = U^{(n_k)}(x_m, t_k)$, $k \geq 1$, can be represented in the form

$$\mathcal{L}_i U_i(x_m, t_k) + f_i(x_m, t_k, U) - \tau_k^{-1} U_i(x_m, t_{k-1}) = \mathcal{R}_i(x_m, t_k, U^{(n_k)}), \quad x_m \in \omega^h,$$

$$U_i(x_*, t_k) = 0, \quad x_* = x_0, x_{M_x}, \quad i = 1, 2.$$

From here, (2) and using the mean-value theorem, we get the difference problem for $W_i(x_m, t_k) = U_i(x_m, t_k) - U_i^*(x_m, t_k)$

$$\left(\mathcal{L}_i + \frac{\partial f_i}{\partial u_i} \right) W_i(x_m, t_k) = \mathcal{R}_i(x_m, t_k, U) + \frac{1}{\tau_k} W_i(x_m, t_{k-1}) - \frac{\partial f_i}{\partial u_{i'}} W_{i'}(x_m, t_k),$$
$$x_m \in \omega^h, \quad W_i(x_*, t_k) = 0, \quad x_* = x_0, x_{M_x} \quad i' \neq i, \quad i = 1, 2, \tag{15}$$

where the partial derivatives are calculated at intermediate points E_i, $i = 1, 2$, such that $U_i^* \leq E_i \leq \overline{U}_i^{(0)}$, $i = 1, 2$, in the case of upper solutions and $\underline{U}_i^{(0)} \leq E_i \leq U_i^*$, $i = 1, 2$, in the case of lower solutions. Thus, the partial derivatives satisfy (12). From here, (12), using (4) and taking into account that according

to Theorem 1 the stopping criterion (11) can always be satisfied, in the notation $w_k = \max_i \|W_i(\cdot, t_k)\|_{\overline{\omega}^h}$ we have

$$w_k \le \frac{1}{\rho_k + \tau_k^{-1}} \left[\delta + \tau_k^{-1} w_{k-1} + \rho_k w_k\right].$$

Solving the last inequality for w_k and taking into account that $\tau_k^{-1}/(\rho_k + \tau_k^{-1}) > 0$, we have

$$w_k \le \delta \tau_k + w_{k-1}.$$

Since $w_0 = 0$, by induction on k, we conclude (14)

$$w_k \le \delta \sum_{l=1}^{k} \tau_l \le T\delta, \quad k \ge 1.$$

3.2 Construction of Initial Upper and Lower Solutions

Here, we give some conditions on functions f_i and ψ_i, $i = 1, 2$, to guarantee the existence of upper $\widetilde{U}$ and lower $\widehat{U}$ solutions, which are used as the initial iterations in the monotone iterative method (6).

Bounded Reactions Functions Assume that f_i, ψ_i, $i = 1, 2$, from (1) satisfy the conditions

$$-\sigma_i \le f_i(x, t, 0) \le 0, \quad \psi_i(x) \ge 0, \quad u_i(x, t) \ge 0, \quad x \in \overline{\omega},$$

where σ_i, $i = 1, 2$, are positive constants. Then

$$\widehat{U}_i(x_m, t_k) = \begin{cases} \psi_i(x_m), & k = 0, \\ 0, & k \ge 1, \end{cases} \quad x_m \in \overline{\omega}^h, \quad i = 1, 2,$$

are lower solutions to (2). The solutions of the following linear problems:

$$\mathcal{L}_i(x_m, t_k)\widetilde{U}_i(x_m, t_k) = \tau_k^{-1}\widetilde{U}_i(x_m, t_{k-1}) + \sigma_i, \quad x_m \in \omega^h, \quad k \ge 1,$$

$$\widetilde{U}_i(x_*, t_k) = 0, \quad x_* = x_0, x_{M_x}, \quad k \ge 1, \quad \widetilde{U}_i(x_m, 0) = \psi_i(x_m), \quad x_m \in \overline{\omega}^h,$$

are upper solutions to (2).

Constant Upper and Lower Solutions Assume that functions f_i, ψ_i, $i = 1, 2$, from (1) satisfy the conditions

$$f_i(x,t,0) \le 0, \quad f_i(x,t,L) \ge 0, \quad \psi_i(x) \ge 0, \quad u_i(x,t) \ge 0, \quad x \in \overline{\omega}, \tag{16}$$

where $L = \text{const} > 0$. The functions

$$\widehat{U}_i(x_m,t_k) = \begin{cases} \psi_i(x_m), & k = 0, \\ 0, & k \ge 1, \end{cases} \qquad \widetilde{U}_i(x_m,t_k) = L, \quad x_m \in \overline{\omega}^h, \tag{17}$$

are, respectively, lower and upper solutions.

4 Uniform Convergence of the Monotone Iterates

We assume that $0 < \varepsilon_1 \le \varepsilon_2 \le 1$.

In the notation $u = (u_1, u_2)$, $\varepsilon = (\varepsilon_1, \varepsilon_2)$ and $f = (f_1, f_2)$, the following linear system is considered in [5]:

$$\frac{\partial u}{\partial t} - \varepsilon \frac{\partial^2 u}{\partial x^2} + A(x,t)u = f(x,t), \quad A(x,t) = \begin{bmatrix} a_{11}(x,t) & a_{12}(x,t) \\ a_{21}(x,t) & a_{22}(x,t) \end{bmatrix},$$

where the matrix $A(x,t)$ satisfies the assumptions

$$a_{ii}(x,t) > 0, \qquad a_{ii'}(x,t) \le 0, \quad a_{ii}(x,t) + a_{ii'}(x,t) \ge \alpha = \text{const} > 0,$$

$$i \ne i', \quad i = 1, 2, \quad (x,t) \in \overline{\omega} \times [0,T].$$

From [5], we write down the bounds on $\partial u_i / \partial x$, $i = 1, 2$ in the form

$$\left| \frac{\partial u_1}{\partial x}(x,t) \right| \le C \left[1 + \mu_1^{-1} \pi_{\mu_1}(x) + \mu_2^{-1} \pi_{\mu_2}(x) \right], \tag{18}$$

$$\left| \frac{\partial u_2}{\partial x}(x,t) \right| \le C \left[1 + \mu_2^{-1} \pi_{\mu_2}(x) \right], \quad \pi_\gamma(x) \equiv \exp(-\gamma^{-1} x) + \exp(-\gamma^{-1}(1-x)),$$

where $\mu_i = \sqrt{\varepsilon_i}$, $i = 1, 2$, and γ is a positive constant. These bounds show that there are two overlapping boundary layers at $x = 0$ and $x = 1$.

By using the mean-value theorem, we write f_i, $i = 1, 2$, from (1) in the form

$$f_i(x,t,u) = f_i(x,t,0) + \frac{\partial f_i}{\partial u_i}(x,t,v) u_i + \frac{\partial f_i}{\partial u_{i'}}(x,t,v) u_{i'}, \quad i' \ne i, \quad i, i' = 1, 2,$$

where v lies between 0 and u. We suppose that $\partial f_i/\partial u_i$ and $\partial f_i/\partial u_{i'}$, $i' \neq i$, $i, i' = 1, 2$, for $(x, t, v) \in \overline{\omega} \times [0, T] \times (-\infty, \infty)$ satisfy the following assumptions:

$$\frac{\partial f_i}{\partial u_i}(x, t, v) > 0, \quad \frac{\partial f_i}{\partial u_{i'}}(x, t, v) \leq 0, \qquad i' \neq i, \quad i, i' = 1, 2, \tag{19}$$

$$\min_{-\infty \leq v \leq \infty} \left[\frac{\partial f_i}{\partial u_i}(x, t, v) + \frac{\partial f_i}{\partial u_{i'}}(x, t, v) \right] > \alpha = \text{const} > 0.$$

Remark 2 If assumptions (19) hold, then Theorem 3.1, Chap. 8 in [7] guarantees existence and uniqueness of the solution to problem (1).

We may now consider (1) as a linear problem and use bounds (18) on the exact solutions. We introduce the piecewise uniform mesh $\overline{\omega}^h$ of Shishkin-type from [5], where the boundary layer thicknesses $\varsigma_{\varepsilon_i}$, $i = 1, 2$, and mesh spacings h_{ε_i}, $i = 1, 2$, h are defined by

$$\varsigma_{\varepsilon_2} = \min \left\{ 1/4, 2\sqrt{\varepsilon_2} \ln M_x \right\}, \quad \varsigma_{\varepsilon_1} = \min \left\{ \varsigma_{\varepsilon_2}/2, 2\sqrt{\varepsilon_1} \ln M_x \right\}, \tag{20}$$

$$h_{\varepsilon_1} = 8\varsigma_{\varepsilon_1}/M_x, \quad h_{\varepsilon_2} = 8(\varsigma_{\varepsilon_2} - \varsigma_{\varepsilon_1})/M_x \quad h = 2(1 - 2\varsigma_{\varepsilon_2})/M_x.$$

The mesh $\overline{\omega}^h$ is constructed thus: in each of the subintervals $[0, \varsigma_{\varepsilon_1}]$, $[\varsigma_{\varepsilon_1}, \varsigma_{\varepsilon_2}]$, $[\varsigma_{\varepsilon_2}, 1-\varsigma_{\varepsilon_2}]$, $[1-\varsigma_{\varepsilon_2}, 1-\varsigma_{\varepsilon_1}]$ and $[1-\varsigma_{\varepsilon_1}, 1]$, mesh points are distributed uniformly with $M_x/8+1, M_x/8+1, M_x/2+1, M_x/8+1$ and $M_x/8+1$ mesh points, respectively. The mesh spacings h_{ε_1}, h_{ε_2} and h are in use, respectively, in the first and last, in the second and fourth, in the third domains.

Theorem 3 *Assume that meshes $\overline{\omega}^\tau$ and $\overline{\omega}^h$ satisfy, respectively, (13) and (20), and $f_i(x, t, u)$, $i = 1, 2$, satisfy (19). Then the nonlinear difference scheme (2) converges ε-uniformly to the solution of (1)*

$$\max_i \left[\max_{t_k \in \overline{\omega}^\tau} \| U_i^*(\cdot, t_k) - u_i^*(\cdot, t_k) \|_{\overline{\omega}^h} \right] \leq C(M_x^{-1} \ln M_x + \tau), \quad \tau = \max_k \tau_k, \tag{21}$$

where U_i^* and u_i^*, $i = 1, 2$, are, respectively, the exact solutions to (2) and (1), C is a generic constant which is independent of ε, M_x and τ.

Proof Since the proof of the theorem follows the proof of Theorem 1 from [3], then we only present the sketch of it.

The exact solutions $u_i^*(x, t)$, $i = 1, 2$, can be presented on $[x_{m-1}, x_{m+1}]$ in the integral-difference form (compare with (5) from [3])

$$\varepsilon_i \mathcal{L}_i^h u_i^*(x_m, t_k) = \frac{\partial u_i^*}{\partial t} + f_i(x_m, t_k, u^*) + I_i(x_m, t_k, u^*), \quad x_m, t_k \in \omega^h \times \omega^\tau,$$

where $u^* = (u_1^*, u_2^*)$, $\mathcal{L}_i^h$, $i = 1, 2$, are defined in (2) and I_i, $i = 1, 2$, are given in the form

$$I_i(x_m, t_k, u^*) = \frac{1}{\hbar_m} \int_{x_{m-1}}^{x_m} \phi_{2,m-1}(s) \left(\int_{x_m}^{s} \frac{d\psi_i(\xi, t_k)}{d\xi} d\xi \right) ds + \frac{1}{\hbar_m} \int_{x_m}^{x_{m+1}} \phi_{1,m}(s) \left(\int_{x_m}^{s} \frac{d\psi_i(\xi, t_k)}{d\xi} d\xi \right) ds,$$

$$\psi_i(x, t_k) = f_i(x, t_k, u^*) + \frac{\partial u_i^*(x, t_k)}{\partial t}, \quad x \in [x_{m-1}, x_{m+1}],$$

$$\phi_{1,m}(x) = \frac{x_{m+1} - x}{\hbar_m}, \quad \phi_{2,m}(x) = \frac{x - x_m}{\hbar_m},$$

The truncation errors $T_i(x_m, t_k)$, $i = 1, 2$, can be represented in the form

$$T_i(x_m, t_k) = T_{i,1}(x_m, t_k) - I_i(x_m, t_k, u^*),$$

$$T_{i,1}(x_m, t_k) \equiv \frac{u_i^*(x_m, t_k) - u_i^*(x_m, t_{k-1})}{\tau_k} - \frac{\partial u_i^*(x_m, t_k)}{\partial t}.$$

Using the Taylor expansion about (x_m, t_k), we obtain

$$\|T_i(\cdot, t_k)\|_{\omega^h} \le \frac{1}{2} \max_{(x,t) \in Q} |u_{i,tt}^*| \tau_k + \|I_i(\cdot, t_k)\|_{\omega^h}. \tag{22}$$

Thus, similar to [3], using bounds (18), the following estimates on $d\psi_i/dx$, $i = 1, 2$, hold true

$$\left| \frac{d\psi_i(x, t)}{dx} \right| \le C \left[1 + \mu_1^{-1} \pi_{\mu_1}(x) + \mu_2^{-1} \pi_{\mu_2}(x) \right], \quad i = 1, 2.$$

From here, using the properties of the piecewise uniform mesh of Shishkin-type and repeating the proof of Theorem 1 from [3], we prove the estimates

$$\|I_i(\cdot, t_k)\|_{\omega^h} \le C \left(M_x^{-1} \ln M_x \right), \quad i = 1, 2.$$

From here and (22), we obtain

$$\|T_i(\cdot, t_k)\|_{\omega^h} \le C \left(M_x^{-1} \ln M_x + \tau \right), \quad i = 1, 2.$$

The difference problems for u_i^*, $i = 1, 2$, can be represented in the form

$$\mathcal{L}_i u_i^*(x_m, t_k) + f_i(x_m, t_k, u^*) - \tau_k^{-1} u_i^*(x_m, t_{k-1}) = T_i(x_m, t_k), \quad x_m \in \omega^h,$$

$$u_i^*(x_*, t_k) = 0, \quad x_* = x_0, x_{M_x}, \quad i = 1, 2.$$

From here, (2) and using the mean-value theorem, we get the difference problem for $W_i(x_m,t_k)=U_i(x_m,t_k)-u_i^*(x_m,t_k)$ in the form

$$\left(\mathcal{L}_i+\frac{\partial f_i}{\partial u_i}\right)W_i(x_m,t_k)=-T_i(x_m,t_k)+\frac{1}{\tau_k}W_i(x_m,t_{k-1})-\frac{\partial f_i}{\partial u_{i'}}W_{i'}(x_m,t_k),$$

$$x_m\in\omega^h,\quad W_i(x_*,t_k)=0,\quad x_*=x_0,x_{M_x}\quad i'\neq i,\quad i=1,2.$$

Now the proof of the theorem repeats the proof of Theorem 2 starting from (15), where $-T_i$, $i=1,2$, are in use instead of $\mathcal{R}_i$, $i=1,2$, in (15).

Theorem 4 *Assume that all the assumptions in Theorem 3 are satisfied. Then for the sequences* $\{\overline{U}^{(n)}\}$ *and* $\{\underline{U}^{(n)}\}$*, generated by (6), (11) with, respectively,* $\overline{U}^{(0)}=\widetilde{U}$ *and* $\underline{U}^{(0)}=\widehat{U}$*, the uniform in* ε *estimate holds*

$$\max_i\left[\max_{t_k\in\overline{\omega}^\tau}\|U_i(\cdot,t_k)-u_i^*(\cdot,t_k)\|_{\overline{\omega}^h}\right]\le C(\delta+M_x^{-1}\ln M_x+\tau),$$

where $U_i(p,t_k)=\overline{U}^{(n_k)}(p,t_k)$ *or* $U_i(p,t_k)=\underline{U}^{(n_k)}(p,t_k)$ *and* u_i^**,* $i=1,2$*, are the exact solutions to (1).*

Proof The proof of the theorem follows from Theorems 2 and 3.

5 Gas-Liquid Interaction Model

The gas-liquid interaction model in the non-dimensional variables can be presented in the form (see [4] for details)

$$\frac{\partial u_1}{\partial t}-\frac{\partial u_1}{\partial x^2}-\kappa_1(1-u_1)u_2=0,\quad (x,t)\in\omega\times(0,T],$$

$$\frac{\partial u_2}{\partial t}-\varepsilon\frac{\partial u_2}{\partial x^2}+\kappa_2(1-u_1)u_2=0,\quad (x,t)\in\omega\times(0,T],$$

$$u_1(0,t)=u_1(1,t)=0,\quad u_2(0,t)=u_2(1,t)=1,$$

$$u_1(x,0)=0,\quad u_2(x,0)=\sin(\pi x),\quad x\in\overline{\omega},$$

where u_1 and u_2 are, respectively, concentrations of a dissolved gas and a dissolved reactant and κ_i, $i=1,2$, are positive constants. The test problem, which corresponds to the case $\varepsilon_1=1$, $\varepsilon_2=\varepsilon$, for small values of ε is singularly perturbed and u_2 has boundary layers of width $\mathcal{O}(\sqrt{\varepsilon})$ near $x=0$ and $x=1$.

It is easy to verify that assumptions (16) with $L_i = 1$, $i = 1, 2$, hold true. Thus, $\widehat{U}_i$ and $\widetilde{U}_i$, $i = 1, 2$, from (17) are, respectively, lower and upper solutions to the test problem. From here, it follows that the inequalities in (12) hold, and one can choose $c_i(x_m, t_k) = \kappa_i$, $i = 1, 2$, in (5) The exact solution is not available, so we estimate the error of the numerical solutions $U_i^{M_x}$, $i = 1, 2$, with respect to the reference solutions $U_i^{2M_x}$, $i = 1, 2$,

$$E_{M_x} = \max_{i=1,2} \|U_i^{M_x}(\cdot, t_{N_\tau}) - U_i^{2M_x}(\cdot, t_{N_\tau})\|_{\overline{\omega}^h},$$

and assume that $E_{M_x} = C(1/M_x)^{p_{M_x}}$, where constant C is independent of M_x, and p_{M_x} is the order of maximum numerical error. For each M_x, we compute p_{M_x} from

$$p_{M_x} = \log_2 \frac{E_{M_x}}{E_{2M_x}}.$$

We choose $\delta = 10^{-8}$ in the stopping test (11). In Table 1, for parameters $\kappa_i = 1$, $i = 1, 2$, $t_{N_\tau} = 0.5$, $\tau = 5 \times 10^{-4}$ and different values of ε and M_x, we present the maximum numerical error E_{M_x}, the order of maximum numerical error p_{M_x} and the number of monotone iterations n_{M_x} on each time level. The data in the table show that for $\varepsilon \le 10^{-4}$, the numerical solution converges uniformly in ε, has the first-order accuracy in the space variable, and the monotone sequences converge in few iterations.

Table 1 Numerical results

M_x		32	64	128	256	512
$\varepsilon = 1$	E_{M_x}	5.949e − 5	2.046e − 5	8.296e − 6	3.712e − 6	1.753e − 6
	p_{M_x}	1.539	1.302	1.160	1.081	
	n_{M_x}	2	2	1	1	1
$\varepsilon = 10^{-1}$	E_{M_x}	4.265e − 4	1.684e − 4	7.054e − 5	3.280e − 5	1.583e − 5
	p_{M_x}	1.341	1.255	1.105	1.051	
	n_{M_x}	2	2	1	1	1
$\varepsilon = 10^{-2}$	E_{M_x}	2.001e − 3	9.127e − 4	4.293e − 4	2.078e − 4	1.021e − 4
	p_{M_x}	1.133	1.088	1.047	1.025	
	n_{M_x}	3	3	2	2	2
$\varepsilon = 10^{-3}$	E_{M_x}	2.058e − 3	9.371e − 4	4.411e − 4	2.135e − 4	1.049e − 4
	p_{M_x}	1.135	1.087	1.047	1.025	
	n_{M_x}	3	3	2	2	2
$\varepsilon \le 10^{-4}$	E_{M_x}	2.103e − 3	9.557e − 4	4.498e − 4	2.177e − 4	1.070e − 4
	p_{M_x}	1.138	1.087	1.047	1.024	
	n_{M_x}	3	3	2	2	2

References

1. Boglaev, I.: Monotone iterates for solving coupled systems of nonlinear parabolic equations. Computing **92**, 65–95 (2011)
2. Boglaev, I.: Uniform quadratic convergence of monotone iterates for nonlinear singularly perturbed parabolic problems. Numer. Algor. **64**, 617–631 (2013)
3. Boglaev, I., Hardy, M.: Uniform convergence of a weighted average scheme for a nonlinear reaction-diffusion problem. Comput. Appl. Math. **200**, 705–721 (2007)
4. Danckwerts, P.V.: Gas-Liquid Reactions. McGraw-Hill, New York (1970)
5. Gracia, J.L., Lisbona, F.: A uniformly convergent scheme for a system of reaction-diffusion equations. J. Comput. Appl. Math. **206**, 1–16 (2007)
6. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems, Revised edn. World Scientific, Singapore (2012)
7. Pao, C.V.: Nonlinear Parabolic and Elliptic Equations. Plenum Press, New York (1992)
8. Samarskii, A.: The Theory of Difference Schemes. Marcel Dekker, New York/Basel (2001)

Order Reduction and Uniform Convergence of an Alternating Direction Method for Solving 2D Time Dependent Convection-Diffusion Problems

C. Clavero and J.C. Jorge

Abstract In this work we solve efficiently 2D time dependent singularly perturbed problems. The fully discrete numerical scheme is constructed by using a two step discretization process, firstly in space, by using the classical upwind finite difference scheme on a special mesh of Shishkin type, and later on in time by using the fractional implicit Euler method. The method is uniformly convergent with respect to the diffusion parameter having first order in time and almost first order in space. We focus our interest on the analysis of the influence of general Dirichlet boundary conditions in the convergence of the algorithm. We propose a simple modification of the natural evaluations, which avoid the order reduction associated to those natural evaluations. Some numerical tests are shown in order to exhibit, from a practical of point of view, the robustness of the numerical method as well as the influence of the improved boundary conditions.

1 Introduction

Let us consider 2D time dependent convection-diffusion singularly perturbed problems defined by

$$\begin{aligned} &\mathscr{L}u \equiv \frac{\partial u}{\partial t} + (\mathscr{L}_{1,\varepsilon}(t) + \mathscr{L}_{2,\varepsilon}(t))\, u = f, \ \text{in } \Omega \times (0,T], \\ &u(x,y,0) = \varphi(x,y), \ \text{in } \Omega, \\ &u(x,y,t) = g(x,y,t), \ \text{in } \partial\Omega \times [0,T], \end{aligned} \tag{1}$$

C. Clavero (✉)
Department of Applied Mathematics and IUMA, University of Zaragoza, Zaragoza, Spain
e-mail: clavero@unizar.es

J.C. Jorge
Department of Computational and Mathematical Engineering and ISC, Public University of Navarra, Pamplona, Spain
e-mail: jcjorge@upna.es

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_4

where $\Omega \equiv (0,1)^2$, and the spatial differential operators $\mathscr{L}_{i,\varepsilon}$, $i = 1,2$ are given by

$$\mathscr{L}_{1,\varepsilon}(t) \equiv -\varepsilon \frac{\partial^2}{\partial x^2} + v_1(x,y,t)\frac{\partial}{\partial x} + k_1(x,y,t), \tag{2}$$
$$\mathscr{L}_{2,\varepsilon}(t) \equiv -\varepsilon \frac{\partial^2}{\partial y^2} + v_2(x,y,t)\frac{\partial}{\partial y} + k_2(x,y,t),$$

respectively. We assume that the diffusion parameter ε, $0 < \varepsilon \leq 1$, can be very small with respect to the convective coefficients which will be considered strictly positive here, i.e., $v_i(x,y,t) \geq \overline{v} > 0$; also, the reaction terms satisfy $k_i(x,y,t) \geq 0$, $i = 1,2$. We assume that sufficient smoothness and compatibility conditions between data hold so that the solution is four times derivable in space and twice in time (see [1, 3] for instance).

It is well known that, in general, when $\varepsilon \ll \overline{v}$, the solution of these problems presents a multiscale character even for smooth data, and the exact solution has regular boundary layers of size $\mathscr{O}(\varepsilon)$ at the sides $x = 1$ and $y = 1$ of the boundary of Ω (see [6–9]). In such case, the use of standard finite difference or finite element methods, defined on uniform meshes, is inappropriate because a large number (ε-dependent) of mesh points will be necessary to obtain accurate approximations. Then, the use of uniformly convergent methods is a much better choice, due to the rates of convergence and the associated error constants being independent of ε and, consequently, they are able to obtain reliable solutions using meshes with a reasonable number of mesh points independently of the value of ε. Here, we use a fitted mesh method (see [7, 9]), which concentrates appropriately the grid points in the boundary layer regions, to obtain a uniformly convergent scheme.

Similar 2D parabolic singularly perturbed problems are analyzed in many works. In [4, 5] the numerical algorithm was defined by using a two step process, discretizing firstly in time and secondly in space. In [1, 2] the technique discretizes first in space and later on integrates in time, via the implicit Euler method, the derived stiff initial value problems. The resulting numerical algorithm in [1, 2] must solve pentadiagonal linear systems at each time level; therefore, the computational cost of the algorithm is high. To reduce the computational cost, here we follow the same technique as in [1, 2], but now we use the fractional implicit Euler method to discretize in time; in this way, only tridiagonal systems have to be solved. We prove that the fully discrete scheme, which combines the fractional implicit Euler method, on a uniform mesh, and the classical upwind scheme, defined on a piecewise uniform Shishkin mesh, is uniformly convergent of first order in time and of almost first order in space.

We focus special attention to the influence of considering general time dependent Dirichlet boundary conditions. It is well known that, when using one step methods, a classical evaluation of the boundary conditions causes, in general, a reduction, both theoretically and numerically, in the order of convergence. This is the rationale for as to consider a different and very simple modification of these evaluations. We prove that the new evaluations of the boundary conditions retain the first

order of consistency of the fractional implicit Euler method, without increasing the computational cost of the algorithm.

The paper is structured as follows: in Sect. 2, we introduce the spatial discretization of the continuous problem on a special nonuniform mesh of Shishkin type and we prove its almost first order uniform convergence. In Sect. 3 we introduce the time discretization and we prove the uniform convergence of the fully discrete method. Finally, in Sect. 4 some numerical results corroborating in practice the theoretical results are shown.

Henceforth, C denotes a generic positive constant independent of the diffusion parameter ε and also of the discretization parameters N and M.

2 Spatial Discretization

In this section we describe the spatial discretization chosen for (1). First we construct the mesh $\Omega_{\overline{N}} \equiv I_{x,\varepsilon,N} \times I_{y,\varepsilon,N}$, as a tensor product of one dimensional piecewise uniform Shishkin meshes, $I_{x,\varepsilon,N} = \{0 = x_0 < \ldots < x_N = 1\}$, $I_{y,\varepsilon,N} = \{0 = y_0 < \ldots < y_N = 1\}$. We give the details of the construction of $I_{x,\varepsilon,N}$. Let us choose N as an even number. We define the transition parameter

$$\sigma_x = \min(1/2, m_x \varepsilon \ln N), \tag{3}$$

where $m_x \geq 1/\overline{v}$; then, the piecewise uniform mesh has $N/2+1$ points in $[0, 1-\sigma_x]$ and $[1-\sigma_x, 1]$, and the mesh points are given by

$$x_i = \begin{cases} 2i(1-\sigma_x)/N, \ i = 0, \ldots, N/2, \\ 1 - \sigma_x + 2(i - N/2)\sigma_x/N, \ i = N/2+1, \ldots, N. \end{cases} \tag{4}$$

In a similar way, defining the transition parameter

$$\sigma_y = \min(1/2, m_y \varepsilon \ln N), \tag{5}$$

where $m_y \geq 1/\overline{v}$, we can construct the mesh $I_{y,\varepsilon,N}$.

Let us denote Ω_N the subgrid composed by all of the points of $\Omega_{\overline{N}}$ which are in the interior of Ω. Let us denote $u_N(t)$ the semidiscrete approximations which we are going to define in Ω_N and let us denote $u_{\overline{N}}(t)$ the natural extension of $u_N(t)$ to $\Omega_{\overline{N}}$, by adding the corresponding evaluations of the boundary data. On these meshes, $\mathscr{L}_{i,\varepsilon,N}$, $i = 1, 2$, are the discretization differential operators of $\mathscr{L}_{i,\varepsilon}$, $i = 1, 2$, using the simple upwind finite difference scheme, which is given by

$$\begin{aligned} &\mathscr{L}_{1,\varepsilon,N}(t)u_N(t)(x_i, y_j) \equiv l_{i-,j}u_N(t)(x_{i-1}, y_j) + l_{i+,j}u_N(t)(x_{i+1}, y_j) + \\ &l^1_{i,j}u_N(t)(x_i, y_j), \quad i = 1, \ldots, N-1, \ j = 0, \ldots, N, \end{aligned} \tag{6}$$

where

$$l_{i-,j} = \frac{-\varepsilon}{h_{x,i}\tilde{h}_{x,i}} - \frac{v_1(x_i, y_j, t)}{h_{x,i}}, \quad l_{i+,j} = \frac{-\varepsilon}{h_{x,i+1}\tilde{h}_{x,i}}, \tag{7}$$
$$l^1_{i,j} = -l_{i-,j} - l_{i+,j} + k_1(x_i, y_j, t),$$

and analogously

$$\mathscr{L}_{2,\varepsilon,N}(t)u_N(t)(x_i, y_j) \equiv l_{i,j-}u_N(t)(x_i, y_{j-1}) + l_{i,j+}u_N(t)(x_i, y_{j+1}) + l^2_{i,j}u_N(t)(x_i, y_j), \ j = 1, \ldots, N-1, \ i = 0, \ldots, N, \tag{8}$$

where

$$l_{i,j-} = \frac{-\varepsilon}{h_{y,j}\tilde{h}_{y,j}} - \frac{v_2(x_i, y_j, t)}{h_{y,j}}, \quad l_{i,j+} = \frac{-\varepsilon}{h_{y,j+1}\tilde{h}_{y,j}}, \tag{9}$$
$$l^2_{i,j} = -l_{i,j-} - l_{i,j+} + k_2(x_i, y_j, t),$$

with $h_{x,i} = x_i - x_{i-1}, \ i = 1, \ldots, N, \ h_{y,j} = y_j - y_{j-1}, \ j = 1, \ldots, N, \ \tilde{h}_{x,i} = (h_{x,i} + h_{x,i+1})/2, \ i = 1, \ldots, N-1, \ \tilde{h}_{y,j} = (h_{y,j} + h_{y,j+1})/2, \ j = 1, \ldots, N-1.$

Let us denote $[.]_N$, the restriction to Ω_N of any function defined in Ω. In [1], it was proven that it holds

$$\|[u(x, y, t)]_N - u_N(t)\|_{\Omega_N} \leq CN^{-1} \ln N, \quad \forall \ t \in (0, T], \tag{10}$$

showing the almost first order of uniform convergence of the spatial discretization.

3 Time Discretization: Uniform Convergence

In this section we discretize in time, by means of the fractional implicit Euler method (see [4]), the stiff initial value problem

$$\begin{aligned} &u'_N(t) + (\mathscr{L}_{1,\varepsilon,N}(t) + \mathscr{L}_{2,\varepsilon,N}(t))\, u_{\overline{N}}(t) = [f]_N, \ \text{ in } \Omega_N, \\ &u_{\overline{N}}(t) = [g]_{\overline{N}}, \ \text{ in } \Omega_{\overline{N}} \backslash \Omega_N, \\ &u_N(0) = [\varphi]_N, \ \text{ in } \Omega_N, \end{aligned} \tag{11}$$

Let $\tau \equiv T/M$ be the time step, and let us consider the mesh $\bar{I}_M = \{t_m = m\tau, \ m = 0, 1, \ldots, M\}$. Let $u^m_N \approx u_N(x, y, t_m), \ m = 0, 1, \ldots, M$. Then, the fully discrete

method is given by

$$
\begin{aligned}
&(i)\ \text{(initialize)}\\
&u_N^0 = [\varphi(x,y)]_N,\ \text{in } \Omega_N.\\
&u_{\overline{N}}^0 = [g(x,y,0)]_{\overline{N}},\ \text{in } \Omega_{\overline{N}}\backslash\Omega_N.\\
&(ii)\ \text{(first half step)}\\
&(I+\tau\mathscr{L}_{1,\varepsilon,N}(t_{m+1}))u_{\overline{N}}^{m+1/2} = u_{\overline{N}}^m + \tau f_{1,\overline{N}}^{m+1},\ \text{in } \Omega_{\overline{N}}\backslash\{0,1\}\times[0,1],\\
&u_{\overline{N}}^{m+1/2} = g_{\overline{N}}^{m+1/2},\ \text{in } \Omega_{\overline{N}}\cap\{0,1\}\times[0,1].\\
&(iii)\ \text{(second half step)}\\
&(I+\tau\mathscr{L}_{2,\varepsilon,N}(t_{m+1}))u_{\overline{N}}^{m+1} = u_{\overline{N}}^{m+1/2} + \tau f_{2,\overline{N}}^{m+1},\ \text{in } \Omega_{\overline{N}}\backslash[0,1]\times\{0,1\},\\
&u_{\overline{N}}^{m+1}(x,y) = g_{\overline{N}}^{m+1},\ \text{in } \Omega_{\overline{N}}\cap[0,1]\times\{0,1\},\\
&\qquad\qquad m=0,\ldots,M-1,
\end{aligned}
\tag{12}
$$

being $f = f_1 + f_2, f_{1,\overline{N}}^{m+1} = [f_1(x,y,t_{m+1})]_{\overline{N}}$, $f_{2,\overline{N}}^{m+1} = [f_2(x,y,t_{m+1})]_{\overline{N}}$.

An important question in the numerical approximation of initial value problems is related with the evaluations of the boundary data. The most classical option for that is given by

$$
\begin{aligned}
&g_{\overline{N}}^{m+1/2} = [g(x,y,t_{m+1})]_{\overline{N}},\ \text{in } \Omega_{\overline{N}}\cap\{0,1\}\times[0,1],\\
&g_{\overline{N}}^{m+1} = [g(x,y,t_{m+1})]_{\overline{N}},\ \text{in } \Omega_{\overline{N}}\cap[0,1]\times\{0,1\}.
\end{aligned}
\tag{13}
$$

Nevertheless, in general, this choice reduces the order of unconditional (independent of N) consistency to zero, and causes a sharp increase in the global error of the method. Then, we propose a different choice for the boundary data, given by

$$
\begin{aligned}
&g_{\overline{N}}^{m+1/2} = (I+\tau\mathscr{L}_{2,\varepsilon,N}(t_{m+1})[g(x,y,t_{m+1})]_{\overline{N}} - \tau f_{2,\overline{N}}^{m+1},\ \text{in } \Omega_{\overline{N}}\cap\{0,1\}\times[0,1],\\
&g_{\overline{N}}^{m+1} = [g(x,y,t_{m+1})]_{\overline{N}},\ \text{in } \Omega_{\overline{N}}\cap[0,1]\times\{0,1\}.
\end{aligned}
\tag{14}
$$

Theorem 1 *Under sufficient smoothness and compatibility conditions on data (see [3]), if we choose the boundary data given in (14), then the error in time satisfies*

$$
\|u_N(t_m) - u_N^M\|_{\Omega_N} \le C\tau,\quad \forall\, m=1,\ldots,M, \tag{15}
$$

therefore, the time integration process (12) is uniformly and unconditionally convergent of first order; in other words, (15) is obtained independently of the size of ε and without restrictions between N and M.

Then, combining the uniform convergence of the spatial and time discretization, the main result follows.

Theorem 2 *Under sufficient smoothness and compatibility conditions on data (see [3]), if we use the improved boundary data (14), then the global error given by*

$$E_{N,M} \equiv \max_{1\le m\le M} \|[u(x,y,t_m)]_N - u_N^m\|_{\Omega_N},$$

satisfies

$$E_{N,M} \le C(N^{-1}\ln N + M^{-1}),$$

and therefore the fully discrete method is uniformly convergent of first order in time and almost first order in space.

Remark 1 In [3], there are the full details of the proofs of the last two results.

4 Numerical Experiments

In this section we solve some test problems using our numerical algorithm. The first example is given by

$$\begin{array}{l} u_t - \varepsilon\Delta u + u_x + u_y + (30t+xy)u = f(x,y,t),\ (x,y,t)\in \Omega\times[0,1], \\ u(x,y,t) = g(x,y,t),\ \text{in } \partial\Omega\times[0,1] \\ u(x,y,0) = \varphi(x,y), \quad x,y\in[0,1], \end{array} \tag{16}$$

where $f(x,y,t)$, $g(x,y,t)$ and $\varphi(x,y)$ are chosen in such way that the exact solution is

$$u(x,y,t) = (e^{-20t}-t)\left(\Psi(x)\Psi(y) - x^2y^2\right),\ \text{ with } \Psi(z) \equiv 1 - z - \frac{1-e^{-\frac{1-z}{\varepsilon}}}{1-e^{-\frac{1}{\varepsilon}}}.$$

Figure 1 shows the solution at the final time $t = 1$; from it, we clearly see the boundary layers at $x = 1$ and $y = 1$.

In all tables corresponding to example (16), we take $m_x = m_y = 1$ to define the transition parameters of the meshes $I_{1,\varepsilon,N}$ and $I_{2,\varepsilon,N}$ respectively. In this example we decompose the right-hand side in the form $f(x,y,t) = f_1(x,y,t) + f_2(x,y,t)$, where $f_2(x,y,t) = f(x,0,t) + y(f(x,1,t) - f(x,0,t))$ and $f_1(x,y,t) = f(x,y,t) - f_2(x,y,t)$.

As the exact solution is known, the maximum global errors at the mesh points can be computed exactly by

$$e_{N,M} = \max_{0\le n\le M}\max_{0\le i\le N}\max_{0\le j\le N} |U_N^n - u(x_i,y_j,t_n)|,$$

and therefore the numerical orders of convergence are calculated by

$$p = \log\left(e_{N,M}/e_{2N,2M}\right)/\log 2.$$

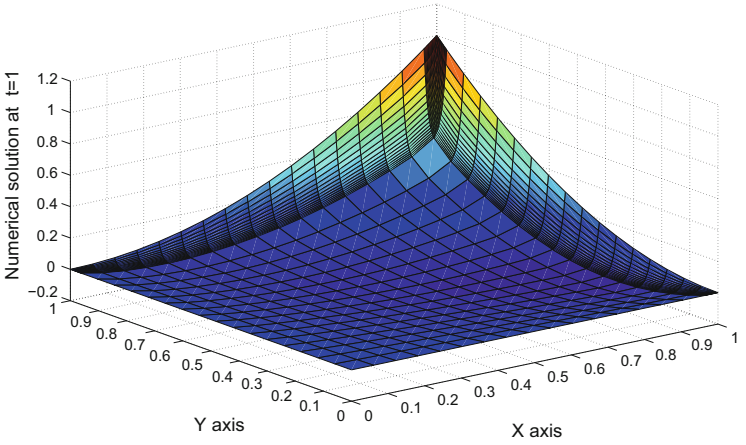

Fig. 1 Numerical solution of example (16) for $\varepsilon = 10^{-2}, N = M = 32$, at the final time $t = 1$

From these values we calculate the uniform maximum errors by $emax^{N,M} = \max_{\varepsilon} e_{N,M}$, and from them, in a usual way, the corresponding numerical uniform orders of convergence are given by

$$p^{uni} = \log\left(emax^{N,M}/emax^{2N,2M}\right)/\log 2.$$

Tables 1 and 2 display the errors and the orders of convergence when natural and improved boundary conditions are used, respectively. From them, we observe the typical almost first order of uniform convergence (up to a logarithmic factor, in both cases; so, we can conclude that in this example the errors associated to the spatial discretization dominate in the global error.

To clarify the influence, in the numerical behavior of the method, of the two options for the boundary data considered here as well as the improvements provided by the non natural evaluations of the boundary conditions, we estimate the local errors in time. As the exact solution is known, such estimates are calculated as

$$\tilde{e}_{N,M} = \max_{0\le m\le M} \max_{0\le i\le N} \max_{0\le j\le N} |\tilde{U}_N^m - u(x_i, y_j, t_m)|,$$

where N must be chosen large enough in order to the contribution of the spatial discretization can be neglected and $\tilde{U}_N^m$ are the result of performing one step of our algorithm, but substituting U_N^{m-1} by $[u(x_i, y_j, t_{m-1}]_N$. From them, the quantities

$$\tilde{p} = \log\left(\tilde{e}_{N,M}/\tilde{e}_{N,2M}\right)/\log 2,$$

permit to estimate the numerical orders of consistency in time, given by $\tilde{p} - 1$.

Next tables show such estimated local errors and the values of $\tilde{p}$ corresponding to the two choices of the boundary data, taking $N = 512$ fixed. Table 3 displays

Table 1 Maximum errors and orders of convergence for (16) with natural boundary conditions

	N=16	N=32	N=64	N=128	N=256
ε	M=8	M=16	M=32	M=64	M=128
1	3.6250E−1	3.1930E−1	2.4345E−1	1.6290E−1	9.8094E−2
	0.183	0.391	0.580	0.732	
2^{-2}	5.1934E−1	4.2391E−1	3.0583E−1	1.9585E−1	1.1409E−1
	0.293	0.471	0.643	0.780	
2^{-4}	7.2837E−1	5.2644E−1	3.5019E−1	2.1350E−1	1.2095E−1
	0.468	0.588	0.714	0.820	
2^{-6}	9.1870E−1	6.2821E−1	3.8314E−1	2.2717E−1	1.2632E−1
	0.548	0.713	0.754	0.847	
2^{-8}	9.8648E−1	6.8344E−1	4.2006E−1	2.3851E−1	1.2930E−1
	0.529	0.702	0.817	0.883	
2^{-10}	1.0042E+0	6.9951E−1	4.3330E−1	2.4729E−1	1.3379E−1
	0.522	0.691	0.809	0.886	
2^{-12}	1.0086E+0	7.0369E−1	4.3702E−1	2.5012E−1	1.3561E−1
	0.519	0.687	0.805	0.883	
2^{-14}	1.0098E+0	7.0474E−1	4.3798E−1	2.5087E−1	1.3614E−1
	0.519	0.686	0.804	0.882	
2^{-16}	1.0100E+0	7.0501E−1	4.3822E−1	2.5107E−1	1.3628E−1
	0.519	0.686	0.804	0.881	
...	...	...	...	...	...
...	...	...	...	...	...
2^{-26}	1.0101E+0	7.0509E−1	4.3830E−1	2.5113E−1	1.3633E−1
	0.519	0.686	0.803	0.881	
$emax^{N,M}$	1.0101E+0	7.0509E−1	4.3830E−1	2.5113E−1	1.3633E−1
p^{uni}	0.519	0.686	0.803	0.881	

the result when natural boundary conditions are used; from it the zero order of consistency of the algorithm can be observed. Table 4 displays the result when improved boundary conditions are used; here, we can appreciate the first order of consistency of the algorithm according to the theoretical results.

The second example that we consider is given by

$$\begin{aligned} &u_t - \varepsilon \Delta u + (1+t+x+y)u_x + (1+xyt^2)u_y + (30t+10xye^{-t})u = \\ &\qquad\qquad e^t\left(x+y+x^2+y^2\right),\ (x,y,t) \in \Omega \times [0,1], \\ &u(x,y,t) = t\left(x+y+x^2+y^2\right),\ \text{in } \partial\Omega \times [0,1] \\ &u(x,y,0) = 0, \quad x,y \in [0,1]. \end{aligned} \tag{17}$$

In this case the exact solution is unknown. We take again $m_x = m_y = 1$ to define the piecewise uniform Shishkin mesh, and we decompose the source term in a different way; now we take $f_1(x,y,t) = f_2(x,y,t) = f(x,y,t)/2$.

Table 2 Maximum errors and orders of convergence for (16) with improved boundary conditions

	N=16	N=32	N=64	N=128	N=256
ε	M=8	M=16	M=32	M=64	M=128
1	8.1032E−2	6.1745E−2	4.1156E−2	2.4708E−2	1.3735E−2
	0.392	0.585	0.736	0.847	
2^{-2}	3.0553E−1	2.1515E−1	1.3340E−1	7.5618E−2	4.0493E−2
	0.506	0.690	0.819	0.901	
2^{-4}	6.6342E−1	4.4616E−1	2.6930E−1	1.5028E−1	7.9834E−2
	0.572	0.728	0.842	0.913	
2^{-6}	8.9060E−1	6.0712E−1	3.6390E−1	2.0135E−1	1.0637E−1
	0.553	0.738	0.854	0.921	
2^{-8}	9.5140E−1	6.6489E−1	4.0674E−1	2.2638E−1	1.1920E−1
	0.517	0.709	0.845	0.925	
2^{-10}	9.6591E−1	6.7917E−1	4.1938E−1	2.3610E−1	1.2533E−1
	0.508	0.696	0.829	0.914	
2^{-12}	9.6947E−1	6.8270E−1	4.2265E−1	2.3889E−1	1.2751E−1
	0.506	0.692	0.823	0.906	
2^{-14}	9.7035E−1	6.8358E−1	4.2347E−1	2.3962E−1	1.2811E−1
	0.505	0.691	0.822	0.903	
2^{-16}	9.7058E−1	6.8380E−1	4.2368E−1	2.3981E−1	1.2827E−1
	0.505	0.691	0.821	0.903	
...	...	...	...	...	...
...	...	...	...	...	...
2^{-26}	9.7065E−1	6.8387E−1	4.2375E−1	2.3987E−1	1.2832E−1
	0.505	0.691	0.821	0.902	
$emax^{N,M}$	9.7065E−1	6.8387E−1	4.2375E−1	2.3987E−1	1.2832E−1
p^{uni}	0.505	0.691	0.821	0.902	

To approximate the maximum pointwise errors, we use a variant of the two-mesh principle. We calculate $\{\hat{u}^N\}$, the numerical solution on the mesh $\{(\hat{x}_i, \hat{y}_j, \hat{t}_n)\}$ containing the original mesh points and its midpoints, i.e.,

$$\begin{aligned} \hat{x}_{2i} &= x_i, \ \ i = 0, \ldots, N, \quad \hat{x}_{2i+1} = (x_i + x_{i+1})/2, \ \ i = 0, \ldots, N-1, \\ \hat{y}_{2j} &= y_j, \ \ j = 0, \ldots, N, \quad \hat{y}_{2j+1} = (y_j + y_{j+1})/2, \ \ j = 0, \ldots, N-1, \\ \hat{t}_{2m} &= t_m, \ \ m = 0, \ldots, M, \quad \hat{t}_{2m+1} = (t_m + t_{m+1})/2, \ \ m = 0, \ldots, M-1. \end{aligned}$$

Then, we estimate the maximum errors at the mesh points of the coarse mesh as

$$d_{i,j,N,M} = \max_{0 \le m \le M} \max_{0 \le i,j \le N} |u^N(x_i, y_j, t_m) - \hat{u}^N(x_i, y_j, t_m)|,$$

the corresponding numerical orders of convergence are given by

$$q = \log \left(d_{i,j,N,M} / d_{i,j,2N,2M} \right) / \log 2.$$

Table 3 Local errors and values of $\tilde{p}$ for (16) with natural boundary conditions, $N = 512$

ε	M=8	M=16	M=32	M=64	M=128
1	4.3707E−1	3.5384E−1	2.5674E−1	1.6629E−1	9.7885E−2
	0.305	0.463	0.627	0.765	
2^{-2}	6.2590E−1	4.8022E−1	3.2776E−1	2.0041E−1	1.1249E−1
	0.382	0.551	0.710	0.833	
2^{-4}	7.2042E−1	5.3813E−1	3.5638E−1	2.1165E−1	1.1575E−1
	0.421	0.595	0.752	0.871	
2^{-6}	8.3783E−1	5.6795E−1	3.7471E−1	2.2236E−1	1.2209E−1
	0.561	0.600	0.753	0.865	
2^{-8}	9.2029E−1	5.7860E−1	3.8173E−1	2.2684E−1	1.2497E−1
	0.670	0.600	0.751	0.860	
2^{-10}	9.4871E−1	5.8178E−1	3.8395E−1	2.2840E−1	1.2611E−1
	0.705	0.600	0.749	0.857	
2^{-12}	9.5726E−1	5.8272E−1	3.8462E−1	2.2889E−1	1.2650E−1
	0.716	0.599	0.749	0.856	
2^{-14}	9.5967E−1	5.8298E−1	3.8481E−1	2.2903E−1	1.2661E−1
	0.719	0.599	0.749	0.855	

Table 4 Local errors and values of $\tilde{p}$ for (16) with improved boundary conditions, $N = 512$

ε	M=8	M=16	M=32	M=64	M=128
1	8.1463E−2	5.6560E−2	3.2374E−2	1.5063E−2	5.8364E−3
	0.526	0.805	1.104	1.368	
2^{-2}	2.9521E−1	1.7902E−1	8.6850E−2	3.4115E−2	1.1376E−2
	0.722	1.044	1.348	1.584	
2^{-4}	6.0748E−1	3.5087E−1	1.6056E−1	5.9411E−2	1.8797E−2
	0.792	1.128	1.434	1.660	
2^{-6}	8.2049E−1	4.6890E−1	2.1207E−1	7.7674E−2	2.4431E−2
	0.807	1.145	1.449	1.669	
2^{-8}	9.1301E−1	5.2039E−1	2.3469E−1	8.5809E−2	2.7069E−2
	0.811	1.149	1.452	1.664	
2^{-10}	9.4559E−1	5.3919E−1	2.4365E−1	8.9700E−2	2.8880E−2
	0.810	1.146	1.442	1.635	
2^{-12}	9.5452E−1	5.4465E−1	2.4662E−1	9.1388E−2	3.0196E−2
	0.809	1.143	1.432	1.598	
2^{-14}	9.5675E−1	5.4605E−1	2.4744E−1	9.1910E−2	3.0710E−2
	0.809	1.142	1.429	1.582	

The uniform maximum errors are estimated by $d^{N,M} = \max_{\varepsilon} d_{i,j,N,M}$; from them, as usual, we define the numerical uniform orders of convergence as

$$q^{uni} = \log\left(d^{N,M}/d^{2N,2M}\right)/\log 2.$$

Tables 5 and 6 display the errors and the orders of convergence when natural and improved boundary conditions are used, respectively. Again, it can be observed that, if the improved boundary conditions are used, the maximum errors present a much better behavior, according to the theoretical results.

Table 5 Maximum errors and orders of convergence for (17) with natural boundary conditions

	N=16	N=32	N=64	N=128	N=256
ε	M=8	M=16	M=32	M=64	M=128
1	1.3913E−1	2.4139E−1	2.6596E−1	2.1717E−1	1.4631E−1
	−.795	−.140	0.292	0.570	
2^{-2}	1.2604E−1	2.1200E−1	2.6799E−1	2.3349E−1	1.6160E−1
	−.750	−.338	0.199	0.531	
2^{-4}	1.5672E−1	1.7202E−1	2.1551E−1	1.9894E−1	1.4647E−1
	−.134	−.325	0.115	0.442	
2^{-6}	2.0382E−1	2.0250E−1	2.4525E−1	2.1963E−1	1.5769E−1
	0.009	−.276	0.159	0.478	
2^{-8}	2.2383E−1	2.1158E−1	2.5590E−1	2.2750E−1	1.6231E−1
	0.081	−.274	0.170	0.487	
2^{-10}	2.2922E−1	2.1399E−1	2.5896E−1	2.3008E−1	1.6393E−1
	0.099	−.275	0.171	0.489	
2^{-12}	2.3064E−1	2.1460E−1	2.5974E−1	2.3083E−1	1.6443E−1
	0.104	−.275	0.170	0.489	
2^{-14}	2.3101E−1	2.1476E−1	2.5993E−1	2.3101E−1	1.6458E−1
	0.105	−.275	0.170	0.489	
2^{-16}	2.3110E−1	2.1480E−1	2.5998E−1	2.3106E−1	1.6461E−1
	0.106	−.275	0.170	0.489	
...	...	...	...	...	...
...	...	...	...	...	...
2^{-26}	2.3113E−1	2.1481E−1	2.6000E−1	2.3108E−1	1.6463E−1
	0.106	−.275	0.170	0.489	
$d^{N,M}$	2.3113E−1	2.4139E−1	2.6799E−1	2.3349E−1	1.6463E−1
q^{uni}	−.063	−.151	0.199	0.504	

Table 6 Maximum errors and orders of convergence for (17) with improved boundary conditions

	N=16	N=32	N=64	N=128	N=256
ε	M=8	M=16	M=32	M=64	M=128
1	1.4057E−1	1.2136E−1	9.2976E−2	6.3759E−2	4.0322E−2
	0.212	0.384	0.544	0.661	
2^{-2}	2.3371E−1	1.6670E−1	1.1193E−1	7.0281E−2	4.2088E−2
	.487	0.575	0.671	0.740	
2^{-4}	2.9010E−1	2.2153E−1	1.5507E−1	1.0209E−1	6.2416E−2
	0.389	0.515	0.603	0.710	
2^{-6}	2.9941E−1	2.2677E−1	1.5615E−1	1.0104E−1	6.1765E−2
	0.401	0.538	0.628	0.710	
2^{-8}	2.9985E−1	2.2861E−1	1.5698E−1	1.0114E−1	6.1679E−2
	0.391	0.542	0.634	0.714	
2^{-10}	2.9954E−1	2.2933E−1	1.5746E−1	1.0321E−1	6.1757E−2
	0.385	0.542	0.609	0.741	
2^{-12}	2.9961E−1	2.2950E−1	1.5845E−1	1.0568E−1	6.3412E−2
	0.385	0.534	0.584	0.737	
2^{-14}	2.9965E−1	2.2954E−1	1.5903E−1	1.0634E−1	6.3966E−2
	0.385	0.529	0.581	0.733	
2^{-16}	2.9966E−1	2.2955E−1	1.5918E−1	1.0650E−1	6.4106E−2
	0.385	0.528	0.580	0.732	
...	...	...	...	...	...
...	...	...	...	...	...
2^{-26}	2.9966E−1	2.2955E−1	1.5923E−1	1.0656E−1	6.4153E−2
	0.385	0.528	0.579	0.732	
$d^{N,M}$	2.9985E−1	2.2955E−1	1.5923E−1	1.0656E−1	6.4153E−2
q^{uni}	0.385	0.528	0.579	0.732	

Acknowledgements This research was partially supported by the project MTM2014-52859 and the Diputación General de Aragón.

References

1. Clavero, C., Jorge, J.C.: Another uniform convergence analysis technique of some numerical methods for parabolic singularly perturbed problems. Comput. Math. Appl. **70**, 222–235 (2015)
2. Clavero, C., Jorge, J.C.: Spatial semidiscretization and time integration of 2D parabolic singularly perturbed problems. In: Lecture Notes in Computational Science and Engineering, vol. 108, pp. 75–85. Springer, Cham (2016)
3. Clavero, C., Jorge, J.C.: A fractional step method for 2D parabolic convection-diffusion singularly perturbed problems: uniform convergence and order reduction. Numer. Algorithms **75**, 809–826 (2017). https://doi.org/10.1007/s11075-016-0221-9

4. Clavero, C., Jorge, J.C., Lisbona, F., Shishkin, G.I.: A fractional step method on a special mesh for the resolution of multidimensional evolutionary convection-diffusion problems. Appl. Numer. Math. **27**, 211–231 (1998)
5. Clavero, C., Gracia, J.L., Jorge, J.C.: A uniformly convergent alternating direction HODIE finite difference scheme for 2D time dependent convection-diffusion problems. IMA J. Numer. Anal. **26**, 155–172 (2006)
6. Linss, T., Stynes, M.: A hybrid difference scheme on a Shishkin mesh for linear convection-diffusion problems. Appl. Numer. Math. **31**, 255–270 (1999)
7. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems, revised edn. World Scientific, Singapore (2012)
8. O'Riordan, E., Stynes, M.: A globally convergent finite element method for a singularly perturbed elliptic problem in two dimensions. Math. Comput. **57**, 47–62 (1991)
9. Roos, H.G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations, 2nd edn. Springer, Berlin (2008)

Laminar Boundary Layer Flow with DBD Plasma Actuation: A Similarity Equation

Gael de Oliveira, Marios Kotsonis, and Bas van Oudheusden

Abstract The framework of self-similar laminar boundary layer flow solutions is extended to include the effect of actuation with body force fields resembling those generated by DBD plasma actuators. The deduction line is similar to previous work investigating the effect of porous wall suction on laminar boundary layers. The starting point of the analysis is a generalised form of the Boundary Layer Partial Differential Equations (BL-PDEs) that includes volume force terms. Actuation force distributions are defined such that the volume force term of the BL-PDE equations conforms to the requirements of similarity. New similarity parameters for the plasma strength and thickness are identified. The procedure yields a general similarity equation which includes the effect of pressure gradients, wall transpiration and DBD plasma actuation. Select numerical solutions of the new similarity equation are presented to develop instinctive understanding and prompt a discussion on the construction of new closure relations for integral boundary layer models.

1 Introduction

Prandtl formulated the Boundary Layer equations for viscous stationary flow over a century ago [1, 2]. His asymptotic analysis confirmed Saint-Venant's [3] justification of drag and explained flow separation [4]. He concluded by pointing that flow separation could be reduced by channelling the boundary layer into a slot.

Active flow control was born at the 1904 mathematical congress [1, 4, 5], but early flow control studies consisted of practical experiments with slot [6–8] and continuous suction [9]. A major breakthrough occurred when Thwaites [10] and Watson [11, 12] extended the framework of similarity solutions [13, 14] to handle continuous wall suction. Their similarity equations prompted further research [15–18] and enabled the design of industrial applications [19–22].

G. de Oliveira (✉) • M. Kotsonis • B. van Oudheusden
Faculty of Aerospace Engineering, Department of Aerodynamics, Wind Energy and Propulsion (AWEP), Delft University of Technology, Kluyverweg 1, 2629HS Delft, Netherlands
e-mail: g.l.deoliveiraandrade@tudelft.nl

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_5

In practice, the implementation of boundary layer suction is often plagued by the complexity and weight of supporting systems [19, 23]. The flow control community is addressing these concerns by developing low footprint concepts for passive [24, 25] and active [23, 26] boundary layer manipulation.

Owing to their low footprint and large bandwidth, Dielectric Barrier Discharge (DBD) plasma actuators have been the object of growing interest as flow control devices [23, 27–29]. In the simplest idealization, DBD actuators impart a controllable force on the flow [28, 30]. Envisioned applications include boundary layer transition [31, 32] and separation [33, 34] control.

The theory of laminar boundary layers under plasma actuation is still incomplete. Asghar [35] and Oliveira [34] extended the Von Karman integral equations [36, 37] to include the effect of plasma forces, but ordinary differential equations for the velocity profile remain unavailable.

The present work extends the Falkner-Skan [14] equation to handle flows with externally imposed body force fields. Section 2 explains the working principle of DBD plasma actuators and describes the flow modelling strategy together with its governing equations. Section 3 describes the procedure for identifying similar flow solutions of the extended Prandtl system introduced in Sect. 2. The main result is a similarity equation which is solved numerically with off-the-shelf solvers in Sect. 4. A final note discusses applications and future research needs.

2 Flow with Idealized DBD Plasma Actuation

DBD actuators consist of an exposed electrode and an encapsulated electrode separated by a dielectric barrier and asymmetrically positioned. When the electric potential between electrodes is varied with appropriate amplitude *O(kV)* and frequency *O(kHz)*, a small region of fluid is ionized near the exposed electrode. The ionized fluid exhibits uneven electric charge distributions and the electric field of the electrodes imparts a force to the flow.

Numerous effects come into play: the dynamic viscosity of air may be affected by chemical changes in the ionized region, temperature increases of a few degrees may lead to buoyancy effects and the force pulsation may excite unstable flow modes. Still, a consensus has emerged amongst physicists [23, 27, 28, 30], suggesting that the main effect of a cold plasma essentially corresponds to that of an externally imposed body force field. Orlov [30] therefore proposed to treat the problem of plasma actuated flow by adding a plasma force term to the steady state incompressible Navier-Stokes equations:

$$\begin{cases} (\mathbf{U}\cdot\nabla)\,\mathbf{U} = -\frac{1}{\rho}\,(\nabla p + \mathbf{F}) \\ \nabla\cdot\mathbf{U} = 0 \end{cases} \tag{1}$$

The force is essentially independent from the flow field whenever the flow velocity is significantly smaller than the ion drift velocity [38], and unsteady effects can be neglected when the excitation period is significantly smaller than the timescale of dominant flow phenomena [39]. Most industrial applications take place in this regime.

2.1 Boundary Layer Equations with Force Terms

The Prandtl [1] system for two-dimensional incompressible stationary flow over plates was extended by Asghar [35] to include the effect of external body force fields like those generated by DBD-Plasma actuators.

$$\begin{cases} U\frac{\partial U}{\partial X} + V\frac{\partial U}{\partial Y} = U_e\frac{\partial U_e}{\partial X} + \nu\frac{\partial^2 U}{\partial Y^2} + \frac{1}{\rho}F_x & \text{momentum equation} \\ \nabla\cdot\mathbf{U} = \frac{\partial U}{\partial X} + \frac{\partial V}{\partial Y} = 0 & \text{continuity equation} \\ \text{Subject to:} & \\ \quad U_{(X,0)} = 0 & \text{No-slip at wall} \\ \quad \lim_{Y\to\infty} U_{(X,Y)} = U^e_{(X)} & \text{Edge velocity} \\ \quad V_{(X,0)} = V^0_{(X)} & \text{Wall transpiration} \end{cases} \tag{2}$$

Asymptotic analysis [34] shows that system (2) approximates physical flows when plasma forces $\left(F_x, F_y\right)$ are small and act along the body. Actuation forces must be significantly smaller than the ratio between stagnation pressure $\left(\frac{1}{2}\rho U_e^2\right)$ and length of the plasma force field $\left(L_p\right)$:

$$O\left(F_x\right) \ll O\left(\frac{\frac{1}{2}\rho U_e^2}{L_p}\right) \qquad , \qquad F_y = 0$$

Following Kotsonis [28], the DBD-plasma force field is approximated through the product of two spatial weighting functions with a constant:

$$F_x = \phi_x^p w^y_{\left(Y,T_p\right)} w^x_{(X)}$$

The $\phi_x^p \in \mathbb{R}$ constant represents the average density of the plasma force field over the actuation region:

$$\phi_x^p = \frac{\int_0^{T_p}\int_0^{L_p} F_x dX dY}{\int_0^{T_p}\int_0^{L_p} dX dY} \tag{3}$$

The thickness of the actuation region is denoted as $T_p \perp X$ and used to define the weighting function $w^y : \mathbb{R}^2 \to \mathbb{R}$ for the normal coordinate, as in references

[28, 34].

$$w^{x}_{(X)} \quad \text{to be determined}$$
$$w^{y}_{(Y,T_p)} = \begin{cases} \frac{\pi}{2} sin\left(\pi\left(\frac{Y}{2T_p} + \frac{1}{2}\right)\right) & , \quad \frac{Y}{T_p} \in [0,1] \\ 0 & , \quad otherwise \end{cases} \tag{4}$$

However, and this differs from references [28, 34], the present work assumes that DBD-plasma actuators can be designed to produce arbitrary force distributions along the longitudinal direction. In fact, the function $w^{x}_{(X)} : \mathbb{R} \rightarrow \mathbb{R}$ will be determined to satisfy the requirements of flow similarity.

3 Similarity Form of the Boundary Layer Equations

Several paths towards similarity forms of the Prandtl system have been proposed since the seminal works of Blasius [13] and Falkner-Skan [14]. Textbook explanations [36, 37, 40, 41] generally follow the exposition given by Schlichting [42]: first the system is rewritten in terms of a streamfunction, then it is postulated that a solution in the form $U = U_e f_{(n)}$ exists and finally it is shown that the momentum equation ceases to depend on x when the similarity postulate is combined with a suitable forcing of the outer flow. Extension of this approach to flow control scenarios is delicate [10–12].

Oleinik and Samokhin [43] proposed a lengthier but more rigorous deduction: they start by integrating the continuity equation to relate velocity components, then rewrite the flow variables across a generic affine transformation and finally determine the conditions under which the transformation leads to self similar forms of the momentum equation. This approach is better suited for flow control applications, so the present work uses the deduction of Oleinik and Samokhin [43] as a template.

3.1 Relation Between Flow Components

Consider a straight path $\boldsymbol{\sigma} \subset \mathbb{R}^2$ running from the wall $(X, 0)$ to some point above it (X, Y) . Integrate the gradient ∇V of the normal velocity field $V : \mathbb{R}^2 \rightarrow \mathbb{R}$ along $\boldsymbol{\sigma}$ with the fundamental theorem of multivariate calculus:

$$V_{(X,Y)} - V_{(X,0)} = \int \underbrace{\left[\begin{array}{cc} \frac{\partial V}{\partial X} & \frac{\partial V}{\partial Y} \end{array}\right]}_{\nabla V} \underbrace{\left[\begin{array}{c} 0 \\ 1 \end{array}\right]}_{\mathbf{n}} d\boldsymbol{\sigma} = \int_0^Y \frac{\partial V}{\partial Y}_{(X,h)} dh \tag{5}$$

Feed the continuity equation into the integral to rewrite the normal speed in terms of the longitudinal velocity and the wall transpiration boundary condition $V_{(X,0)} = V^0_{(X)}$:

$$\begin{aligned} \frac{\partial V}{\partial Y} = -\frac{\partial U}{\partial X} \qquad \Rightarrow V_{(X,Y)} &= V_{(X,0)} - \int_0^Y \frac{\partial U}{\partial X}_{(X,h)} dh \\ &= V^0_{(X)} - \int_0^Y \frac{\partial U}{\partial X} dY \end{aligned}$$

Now feed into the momentum equation to rewrite system (2) into a simpler form with a single partial differential equation, and a single unknown field $U : \mathbb{R}^2 \to \mathbb{R}$:

$$\begin{cases} U\frac{\partial U}{\partial X} + \left(V^0_{(X)} - \int_0^Y \frac{\partial U}{\partial X} dY\right) \frac{\partial U}{\partial Y} = U_e \frac{\partial U_e}{\partial X} + \nu \frac{\partial^2 U}{\partial Y^2} + \frac{\phi^p_x}{\rho} w^y_{(Y,T_p)} w^x_{(X)} \\ \text{Subject to:} \\ \quad U_{(X,0)} = 0 \qquad\qquad \text{No-slip at wall} \\ \quad \lim_{Y\to\infty} U_{(X,Y)} = U^e_{(X)} \quad \text{Edge velocity} \\ \quad V_{(X,0)} = V^0_{(X)} \qquad \text{Wall transpiration} \end{cases} \tag{6}$$

Parameters ϕ^p_x, ρ are constants and the function $w^y_{(Y,T_p)}$ was defined in expression (4). Functions $U^e_{(x)}$, $V^0_{(X)}$ and $w^x_{(X)}$ will be determined to satisfy the requirements of flow similarity in Sect. 3.3.

3.2 Transformation of Flow Variables

Following Oleinik and Samokhin [43], we proceed to search for solutions of system (6) in a transformed space, as per Fig. 1. Without loss of generality, the longitudinal velocity is represented in a form that is suitable for identifying similarity conditions:

$$U_{(X,Y)} = U^e_{(X)} \frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)},\lambda_{(X)})} \qquad \text{with} \qquad \begin{cases} \frac{\partial f}{\partial \eta}_{(0,\lambda)} = 0 \\ \lim_{\eta\to\infty} \frac{\partial f}{\partial \eta}_{(\eta,\lambda)} = 1 \end{cases} \tag{7}$$

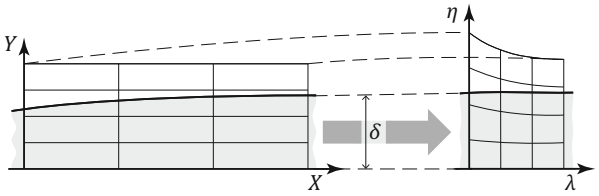

Fig. 1 Map between the (X, Y) and the (λ, η) coordinate systems. The horizontal axis corresponds to the wall, the *shaded region* represents the boundary layer and δ is an estimate for its thickness

Function $f : \mathbb{R}^2 \to \mathbb{R}$ is the main unknown, it will be called similarity function and assumed to be trice differentiable. Map $\eta_{(Y,X)} : \mathbb{R}^2 \to \mathbb{R}$ scales the normal coordinate:

$$\eta_{(Y,X)} = \frac{Y}{\delta_{(X)}} \quad , \quad \frac{\partial \eta}{\partial X} = -\frac{Y}{\delta_{(X)}^2}\frac{\partial \delta}{\partial X}_{(X)} \quad , \quad \frac{\partial \eta}{\partial Y} = \frac{1}{\delta_{(X)}} \tag{8}$$

Map $\lambda_{(X)} : \mathbb{R} \to \mathbb{R}$ fulfills a similar purpose for the longitudinal coordinate and function $\delta_{(X)} : \mathbb{R} \to \mathbb{R}^+$ is usually associated with boundary layer thickness. It is assumed that the $(X, Y) \to (\lambda, \eta)$ map is invertible.

Definitions will be refined at a later stage, but the derivatives of the velocity field can already be rewritten using the generic properties of the transformation (8) and similarity function (7):

$$\begin{aligned}
&\frac{\partial}{\partial Y}\left(\frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)},\lambda_{(X)})}\right) = \frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})}\frac{\partial \eta}{\partial Y} = \frac{1}{\delta_{(X)}}\frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})} && (a)\\
&\frac{\partial^2}{\partial Y^2}\left(\frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)},\lambda_{(X)})}\right) = \frac{\partial}{\partial Y}\left(\frac{1}{\delta_{(X)}}\frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})}\right) = \frac{1}{\delta_{(X)}^2}\frac{\partial^3 f}{\partial \eta^3}_{(\eta_{(Y,X)},\lambda_{(X)})} && (b)\\
&\frac{\partial}{\partial X}\left(\frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)},\lambda_{(X)})}\right) = -\frac{\eta_{(Y,X)}}{\delta_{(X)}}\frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})}\frac{\partial \delta}{\partial X}_{(X)} + \frac{\partial^2 f}{\partial \eta \partial \lambda}_{(\eta_{(Y,X)},\lambda_{(X)})}\frac{\partial \lambda}{\partial X} && (c)
\end{aligned} \tag{9}$$

These derivatives (9) are combined with expression (7) to write key terms from the momentum equation of system (6):

$$\begin{aligned}
\frac{\partial U}{\partial Y}_{(X,Y)} &= \frac{U^e_{(X)}}{\delta_{(X)}}\frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})} && (a)\\
\frac{\partial^2 U}{\partial Y^2}_{(X,Y)} &= \frac{U^e_{(X)}}{\delta^2_{(X)}}\frac{\partial^3 f}{\partial \eta^3}_{(\eta_{(Y,X)},\lambda_{(X)})} && (b)\\
\frac{\partial U}{\partial X}_{(X,Y)} &= \frac{\partial U^e}{\partial X}_{(X)}\frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)},\lambda_{(X)})} - \frac{U^e_{(X)}}{\delta_{(X)}}\eta_{(Y,X)}\frac{\partial^2 f}{\partial \eta^2}_{(\eta_{(Y,X)},\lambda_{(X)})}\frac{\partial \delta_{(X)}}{\partial X}\\
&\qquad + U^e_{(X)}\frac{\partial^2 f}{\partial \eta \partial \lambda}_{(\eta_{(Y,X)},\lambda_{(X)})}\frac{\partial \lambda}{\partial X} && (c)
\end{aligned} \tag{10}$$

The x-derivative of the longitudinal velocity U is integrated along the normal direction Y with the variable change theorem:

$$\begin{aligned}
\int_0^Y \frac{\partial U}{\partial X}_{(X,Y)} dY &= \int_{\eta_{(0,X)}}^{\eta_{(Y,X)}} \frac{\partial U}{\partial X}_{\left(X_{(\lambda)},Y_{(\eta,\lambda)}\right)}\frac{dY}{d\eta}d\eta = \delta\int_0^\eta \frac{\partial U}{\partial X}d\eta\\
&= \delta\left(\frac{\partial U^e}{\partial X}\int_0^\eta\left(\frac{\partial f}{\partial \eta}\right)d\eta - \frac{U^e}{\delta}\frac{\partial \delta}{\partial X}\int_0^\eta\left(\eta\frac{\partial^2 f}{\partial \eta^2}\right)d\eta + U^e\frac{\partial \lambda}{\partial X}\int_0^\eta\left(\frac{\partial^2 f}{\partial \eta \partial \lambda}\right)d\eta\right)
\end{aligned} \tag{11}$$

Notation shorthands are adopted for the sake of readability: η means $\eta_{(X.Y)}$, U^e represents $U^e_{(X)}$ and δ denotes $\delta_{(X)}$. The integrals are solved with the chain rule and

we take $f_{(0,\lambda)} = 0 \Rightarrow \frac{\partial f}{\partial \lambda}_{(0,\lambda)} = 0$ without loss of generality:

$$\int_0^Y \frac{\partial U}{\partial X} dY = \left(f_{(\eta,\lambda)}\right) \frac{\partial}{\partial X}\left(U^e \delta\right) - \eta \frac{\partial f}{\partial \eta} U^e \frac{\partial \delta}{\partial X} + \left(\frac{\partial f}{\partial \lambda}_{(\eta,\lambda)}\right) U^e \delta \frac{\partial \lambda}{\partial X} \tag{12}$$

The momentum equation is rewritten in the transformed space by reworking the velocity terms with expressions (10)*(a–c)* and (12). Extensive algebraic manipulations lead to an interesting form of system (6):

$$\begin{cases} \overbrace{\left(\left(\frac{\partial f}{\partial \eta}\right)^2 - 1 - f\frac{\partial^2 f}{\partial \eta^2}\right) U^e \frac{\partial U^e}{\partial X}}^{Pressure\ Gradient} - \overbrace{\left(f\frac{\partial^2 f}{\partial \eta^2}\right) \frac{U_e^2}{\delta}\frac{\partial \delta}{\partial X}}^{Convection} + \\ \qquad + \nu \frac{U_e}{\delta^2}\left(\overbrace{\frac{V^0 \delta}{\nu}\frac{\partial^2 f}{\partial \eta^2}}^{Suction} - \overbrace{\frac{\partial^3 f}{\partial \eta^3}}^{Shear} - \overbrace{\frac{\delta^2}{\nu U_e}\frac{1}{\rho}F_x}^{Plasma}\right) \\ \qquad = -\overbrace{\left(\left(\frac{\partial f}{\partial \eta}\frac{\partial^2 f}{\partial \eta \partial \lambda}\right) - \frac{\partial^2 f}{\partial \eta^2}\frac{\partial f}{\partial \lambda}_{(\eta,\lambda)}\right) U_e^2 \frac{\partial \lambda}{\partial X}}^{Trasnformation\ Stretching} \\ \text{Subject to:} \\ \quad \frac{\partial f}{\partial \eta}_{(0,\lambda)} = 0 \text{ No-slip at wall} \qquad \lim_{\eta \to \infty} \frac{\partial f}{\partial \eta}_{(\eta,\lambda)} = 1 \text{ Edge velocity} \\ \quad f_{(0,\lambda)} = 0 \quad \text{Integration Constant} \end{cases} \tag{13}$$

When moving from Eqs. (11) to (13), Oleinik and Samokhin [43] eliminated the transpiration term by setting the $f_{(0,\lambda)}$ integration constant as a function of $V^0_{(x)}$. The current presentation adopts a different approach by keeping the suction term explicitly visible, which is consistent with previous flow control work by Thwaites [10] and Watson [11, 12].

3.3 Similarity Conditions

Flat plate flows are said to be self-similar when the velocity profile maintains a constant shape along the plate. Similarity occurs when there exists a map $(X, Y) \to (\eta, \lambda)$ such that the quantity $\frac{U}{U_e} = \frac{\partial f}{\partial \eta}_{(\eta,\lambda)}$ depends on a single scaled coordinate (η):

$$\exists \begin{pmatrix} \eta : \mathbb{R}^2 \to \mathbb{R} \\ \lambda : \mathbb{R} \to \mathbb{R} \end{pmatrix} \quad : \quad \frac{U}{U_e} = \frac{\partial f}{\partial \eta}_{(\eta_{(Y,X)}, \lambda_{(X)})} \perp (X, \lambda)$$

Our endeavour will now consist in identifying a map (η, λ) and a set of boundary conditions $(U^e_{(X)}, V^0_{(X)}, w^x_{(X)})$ such that the solutions $f_{(\eta,\lambda)}$ of the scaled Prandtl system (13) are independent from the longitudinal coordinate.

3.3.1 Blasius Flow

Different maps and boundary condition choices may lead to different types of similar flow. Blasius [13] identified the first similarity solution of the Prandtl system for unactuated flow ($V^0 = 0\,,\ F_x = 0$) with no pressure gradient $\frac{dU_e}{dX} = 0$. Blasius [13, 43] chose a transformation such that:

$$\lambda_{(X)} \equiv \xi_{(X)} \qquad with \qquad U_e^2 \left(\left(\frac{\partial f}{\partial \eta} \frac{\partial^2 f}{\partial \eta \partial \xi} \right) - \frac{\partial^2 f}{\partial \eta^2} \frac{\partial f}{\partial \xi}_{(\eta,\lambda)} \right) \frac{\partial \xi}{\partial X} \equiv 0$$

A rigorous discussion of the reasoning behind this choice can be found in Oleinik [43]. The momentum equation (13) then takes a very simple form:

$$\left.\begin{array}{l} \frac{dU_e}{dX} = 0 \\ \frac{d\lambda}{dX} = 0 \\ V^0 = 0 \\ F_x = 0 \end{array}\right\} \quad \Rightarrow \quad -\left(f \frac{\partial^2 f}{\partial \eta^2}\right) \frac{U_e^2}{\delta} \frac{\partial \delta}{\partial X} - \nu \frac{U_e}{\delta^2} \left(\frac{\partial^3 f}{\partial \eta^3} \right) = 0 \tag{14}$$

The solutions f of Eq. (14) will be independent of $(\lambda\,, X)$ when $\delta_{(X)}$ is chosen such that:

$$\exists\, \kappa_1 \in \mathbb{R} \quad : \quad \left(\frac{U_e^2}{\delta_{(X)}} \frac{\partial \delta}{\partial X} \right) = \kappa_1 \left(\nu \frac{U_e}{\delta^2_{(X)}} \right) \quad \forall X \in \mathbb{R}^+$$

These conditions are satisfied when the normal coordinate (Y) is scaled with the boundary layer thickness estimate proposed by Prandtl [1]:

$$\left.\begin{array}{l} \lambda_{(X)} \equiv \xi_{(X)} \\ \delta_{(X)} \equiv \frac{X}{\sqrt{Re_X}} \\ Re_X \equiv \frac{U_e X}{\nu} \\ U^e_{(X)} \equiv const. \end{array}\right\} \quad \Rightarrow \quad \begin{array}{l} f \frac{\partial^2 f}{\partial \eta^2} + \frac{\partial^3 f}{\partial \eta^3} = 0 \\ \Leftrightarrow \quad f f'' + f''' = 0 \end{array} \tag{15}$$

Prime notation (f') was adopted to ease comparison with previous works.

3.3.2 Falkner-Skan Flow

Falkner and Skan [14] extended the similarity solutions of Blasius [13] to handle flow with non-zero external pressure gradients $\left(\frac{\partial U_e}{\partial X} \neq 0 \right)$. The momentum equation

of system (13) then reads:

$$\left.\begin{array}{l}\frac{d\lambda}{dX}=0\\ V^0=0\\ F_x=0\end{array}\right\} \quad \Rightarrow \begin{array}{r}\left(\left(\frac{\partial f}{\partial \eta}\right)^2-1-f\frac{\partial^2 f}{\partial \eta^2}\right)U^e\frac{\partial U^e}{\partial X}-\ \ldots\\ \ldots-\left(f\frac{\partial^2 f}{\partial \eta^2}\right)\frac{U_e^2}{\delta}\frac{\partial \delta}{\partial X}-\nu\frac{U_e}{\delta^2}\left(\frac{\partial^3 f}{\partial \eta^3}\right)=0\end{array} \tag{16}$$

Equation (16) will not depend on (λ, X) similar if $\delta_{(X)}$ and $U^e_{(X)}$ are chosen such that:

$$\exists\, \kappa_1, \kappa_2 \in \mathbb{R} \quad : \quad \left(\frac{U_e^2}{\delta_{(X)}}\frac{\partial \delta}{\partial X}\right)=\kappa_1\left(\nu\frac{U_e}{\delta^2_{(X)}}\right)=\kappa_2\left(U^e\frac{\partial U^e}{\partial X}\right) \quad \forall X \in \mathbb{R}^+$$

These conditions occur on wedges [36, 37], where the outer flow velocity varies with a power law [14, 43]:

$$\left.\begin{array}{l}\lambda_{(X)}\equiv \xi_{(X)}\\ \delta_{(X)}\equiv \frac{X}{\sqrt{Re_X}}\\ U^e_{(X)}\equiv cX^m\end{array}\right\} \quad \Rightarrow \quad \begin{array}{l} m\left(\frac{df}{d\eta}\right)^2-m-\frac{1}{2}(m+1)f\frac{\partial f^2}{\partial \eta}-\frac{\partial^3 f}{\partial \eta^3}=0\\ \Leftrightarrow \quad m(f')^2-\frac{1}{2}(m+1)ff''=m+f'''\end{array} \tag{17}$$

Constant m is called the pressure gradient parameter, and it is usually varied to derive correlations between boundary layer parameters. Closure relations [16, 44] based on the Falkner-Skan similarity conditions provide excellent approximations to many real (non-similar) flows. This is of immense practical importance for the calculation of subsonic airfoil flows [24] with viscous-inviscid solvers [34, 45].

3.3.3 Actuated Flow with Pressure Gradient

The procedure for extending the Falkner-Skan family of similarity solutions to actuated flows is simple. Having chosen a transformation of flow variables, it suffices to determine a set of boundary conditions $\left(V^0_{(X)}, F^x_{(X,Y)}\right)$ that turns the actuation terms independent of λ in the transformed space. Let us then rewrite the momentum equation of system (13) across the Falkner-Skan transformation:

$$\left.\begin{array}{l}\lambda_{(X)}\equiv \xi_{(X)}\\ \delta_{(X)}\equiv \frac{X}{\sqrt{Re_X}}\\ U^e_{(X)}\equiv cX^m\end{array}\right\} \quad \Rightarrow$$

$$m\left(\frac{df}{d\eta}\right)^2-m-\frac{1}{2}(m+1)f\frac{\partial f^2}{\partial \eta}+\left(\overbrace{\frac{V^0\delta}{\nu}\frac{\partial^2 f}{\partial \eta^2}}^{Suction}-\overbrace{\frac{\partial^3 f}{\partial \eta^3}}^{Shear}-\overbrace{\frac{\delta^2}{\nu U_e}\frac{1}{\rho}F_x}^{Plasma}\right)=0 \tag{18}$$

Observation of the third parcel of Eq. (18) indicates that actuated flows are self-similar when $V^0_{(X)}$ and $F^x_{(X,Y)}$ are chosen such that:

$$\exists\ (\kappa_3, \kappa_4) \in \mathbb{R}\ :\qquad 1 = \kappa_3 \frac{V^0 \delta}{\nu} = \kappa_4 \frac{\delta^2}{\nu U_e} \frac{1}{\rho} F_x \quad \forall\ X \in \mathbb{R}^+$$

Boundary Layer Suction can then be quantified in terms of a suction strength similarity parameter β, consistent with the definitions of Thwaites [10] and Watson [11, 12]:

$$\beta \equiv \frac{V^0 \delta}{\nu} = \left(\frac{V^0}{U_e}\right) Re_\delta = const. \qquad with \qquad Re_\delta = \frac{U_e \delta}{\nu} \tag{19}$$

A similar procedure leads to the identification of plasma actuation similarity conditions. The first step consists in observing that the normal weighting function from expression (4) can be rewritten in scaled variables quite easily:

$$w^y_{(Y,T_p)} = w^y_{\left(\frac{Y}{\delta}, \frac{T_p}{\delta}\right)} \quad \perp \quad \lambda$$

The plasma force term can then decomposed into the product of the normal weighting function with a plasma strength parameter α:

$$\begin{aligned} &\frac{1}{\rho} F_x \equiv \frac{\delta^2}{\nu U_e} \frac{\phi^p_x}{\rho} w^y_{(Y,T_p)} w^x_{(X)} = \alpha w^y_{(\eta,\bar{t}_p)} \\ &with \qquad\qquad \alpha \equiv \frac{\delta^2}{\nu U_e} \frac{\phi^p_x}{\rho} w^x_{(X)} = const. \\ &\qquad\qquad\quad \bar{t}_p \equiv \frac{T_p}{\delta} = const. \end{aligned} \tag{20}$$

Using the definitions from expressions (19) and (20), Eq. (18) is finally rewritten in similarity form:

$$m \left(\frac{df}{d\eta}\right)^2 - m - \frac{1}{2}(m+1) f \frac{\partial f^2}{\partial \eta} + \beta \frac{\partial^2 f}{\partial \eta^2} - \frac{\partial^3 f}{\partial \eta^3} - \alpha w^y_{(\eta,\bar{t}_p)} = 0 \tag{21}$$

It has then been shown that there exist similarity solutions of the extended Prandtl system (2) under the following conditions:

$$\begin{cases} m (f')^2 - m - \frac{1}{2}(m+1) f f'' + \beta f'' - f''' - \alpha w^y_{(\eta,\bar{t}_p)} = 0 \\ \text{Subject to:} \\ \quad \frac{\partial f}{\partial \eta}_{(0,\lambda)} = 0 \qquad\qquad m = const. \\ \quad \lim_{\eta \to \infty} \frac{\partial f}{\partial \eta}_{(\eta,\lambda)} = 1 \quad \alpha = const. \\ \quad f_{(0,\lambda)} = 0 \qquad\qquad \beta = const. \end{cases} \tag{22}$$

4 Numerical Solutions of the Similarity Equation

Equation (21) is the main result of the present contribution. The authors do not have the pretension to identify an optimal numerical method for solving the proposed similarity equation.

There is extensive literature about the numerical solution of boundary value problems for ODEs resembling system (22). Blasius [13], Thwaites [10] and Watson [11, 12] approximated the solutions of their equations with series expansions. Hartree [46, 47] used his *analog computer* to solve the Falkner-Skan [14] problem. In the digital era, early results were presented by Cebeci and Keller [48] and followed by numerous contributions from Asaithambi [49], Farrell et al. [50] and Elgazery [51].

Figures 2 and 3 present results obtained with a popular off-the-shelf solver [52]. The problem was rescaled to handle far-field boundary conditions effectively, as suggested by Farrell et al. [50] (Chap. 11). The two algorithms described in reference [52] yield consistent results as long as the pressure gradient (m), suction (β) and plasma strength (α) remain sufficiently favourable.

The effect of co-flow plasma forces is similar to that of favorable pressure gradients or continuous wall suction, whereas counter-flow plasma actuators are comparable with adverse pressure gradients and wall blowing. Even so, differences between the two phenomena exist: plasma forces act closer to the wall, and therefore lead to greater changes in skin friction than pressure gradients do. Furthermore,

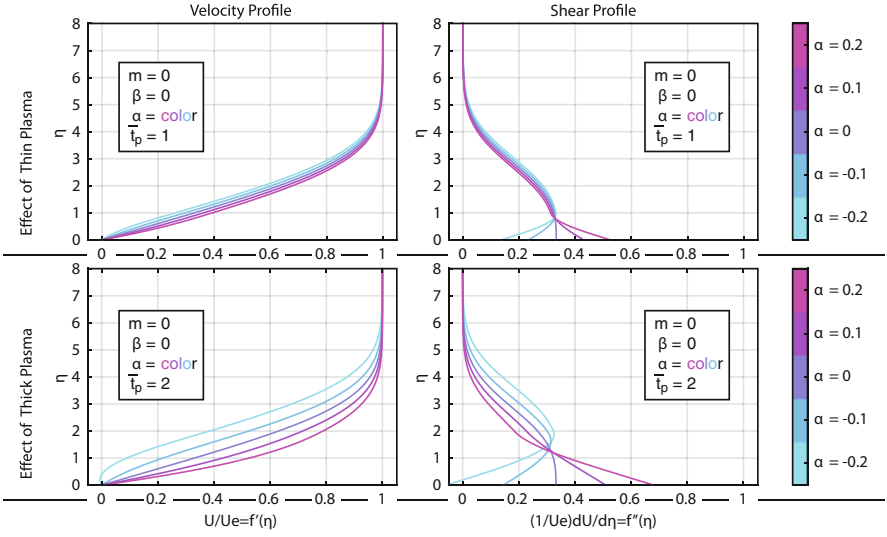

Fig. 2 Numerical solutions of Eq. (21) for different plasma force (α) and thickness parameters ($\bar{t}_p$)

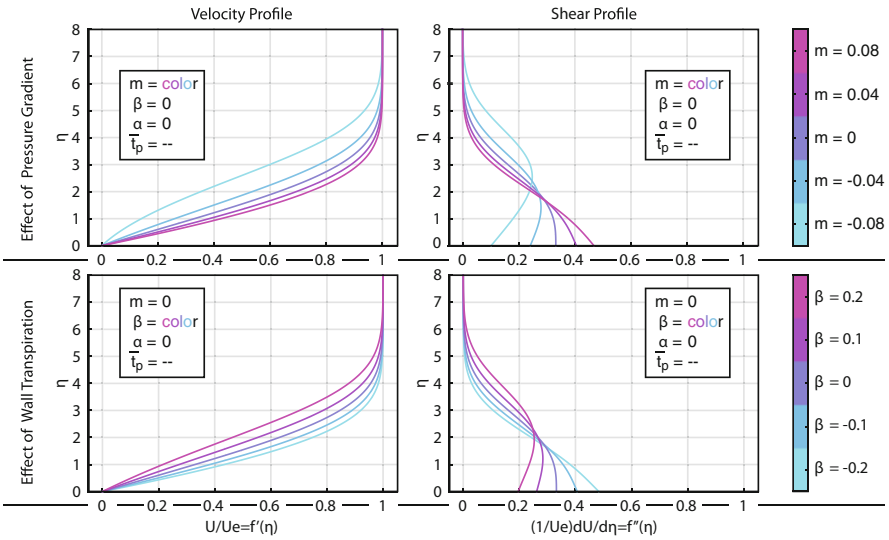

Fig. 3 Numerical solutions of Eq. (21) for different pressure gradient (m) and suction parameters (β)

thinner plasma force fields seem to have a smaller effect than thicker plasma force fields.

5 Applications and Future Research

The developed similarity equation is expected to contribute to the improvement of viscous-inviscid airfoil analysis codes like Xfoil [45] and Rfoil [34, 44]. Doing so will enhance the ability of aerodynamicists to taylor the design of airfoils for reaping the greatest possible benefits from active flow control technologies [24, 33].

Many questions regarding the similarity equation (21) proposed in this paper remain. A serious numerical analysis is still to be done and solution existence or uniqueness conditions aren't defined yet. It is well known [36, 37] that the Falkner-Skan equation (17) ceases to have unique solutions for $m < -0.9040$ but no similar criteria have been established for the plasma strength parameter (20).

References

1. Prandtl, L: Motion of fluids with very little viscosity. NACA Technical Memorandum 452 (1928)
2. Meier, G.E.A., Sreenivasan, K.R., Heinemann, H.J. (eds.): IUTAM Symposium on One Hundred Years of Boundary Layer Research. Springer, Dordrecht (2006)

3. Saint-Venant, A.B.: Mémoire sur la théorie de la résistance des fluides. C. R. Hebd. Seances Acad. Sci. **24** 243–246 (1847)
4. Eckert, M.: The Dawn of Fluid Dynamics. Wiley, Weinheim (2006)
5. Gad-el-Hak, M., Pollard, A., Bonnet, J.P. (ed.): Flow Control Fundamentals and Practices. Springer, Berlin (1998)
6. Griffith, A., Meredith, F.W.: Possible improvement in aircraft performance due to use of boundary layer suction. British ARC R&M 2315 (1935)
7. Richards, E.J., Walker, W.S., Taylor, C.R.: Wind-tunnel tests on a 30 per cent. suction wing. British ARC R&M 2149 (1945)
8. Sage, A., Sargent, R.F.: Design of suction slots. British ARC R&M 2127 (1944)
9. Kay, J.M.: Boundary layer flow along a flat plate with uniform suction. ARC R&M 2628 (1948)
10. Thwaites, B.: An exact solution of the boundary-layer equations under particular conditions of porous surface suction. British ARC R&M 2241 (1946)
11. Watson, E.J.: Asymptotic solution of a boundary layer suction problem. British ARC R&M 2298 (1950)
12. Watson, E.J.: The asymptotic theory of boundary layer flow with suction. British ARC R&M 2619 (1952)
13. Blasius, H.: The boundary layers in fluids with little friction. NACA Technical Memorandum 1256 (1950)
14. Falkner V.M., Skan S.W.: Solutions of the boundary layer equations. Philos. Mag. **12**(80), 865–896 (1931)
15. Tennekes, H.: Similarity laws for turbulent boundary layers with suction or injection. J. Fluid Mech. **21**(4), 689–703 (1965)
16. van Ingen, J., Kotsonis, M.: A two-parameter method for e^N transition prediction. AIAA 2011-3928 (2011)
17. Greenblatt, D., Paschal, K.B., Yao, C.S., Harris, J., Schaeffler, N.W., Washburn, E.: Experimental investigation of separation control part 1: baseline and steady suction. AIAA J. **44**(12), 2820–2830 (2006)
18. Chen, C., Seele, R., Wygnanski, I.: Flow control on a thick airfoil using suction compared to blowing. AIAA J. **51**(6), 1462–1472 (2013)
19. Boermans, L.M.M.: Practical implementations of boundary layer suction for drag reduction and lift enhancement at low speed. Presentation at KATnet II Workshop, Ascot UK (2008)
20. Blackner, A.M.: Jet engine and method for reducing jet engine noise by reducing nacelle boundary layer thickness. U.S. Patent 6,094,907, 5 Jun 1996
21. Actiflow BV: Ferrari 599x with Active BLS. http://www.actiflow.nl/ (2009). Accessed 20 Oct 2016
22. Grife, R., Darabai, A., Wygnanski, I.: Download reduction on a three dimensional V-22 model using active flow control. AIAA 2002-3071 (2002)
23. Moreau, E.: Airflow control by non-thermal plasma actuators. J. Phys. D. Appl. Phys. **40**(3), 605–636 (2007)
24. van Rooij, R.P.J.O.M, Timmer, W.A.: Roughness sensitivity considerations for thick rotor blade airfoils. Trans. ASME **125**, 468–478 (2003)
25. Smith, F.T.: Theorethical prediction and design for vortex generators in turbulent boundary layers. J. Fluid Mech. **270**, 91–131 (1994)
26. Seifert, A., Bachar, D., Koss, D., Shepshelovich, M., Wygnanski, I.: Oscillatory blowing: a tool to delay boundary-layer separation. AIAA J. **31**(11), 2052–2060 (1993)
27. Corke, C.T., Lon Enloe, C., Wilkinson, S.P.: Dielectric barrier discharge plasma actuators for flow control. Ann. Rev. Fluid Mech. **42**, 505–529 (2010)
28. Kotsonis, M.: Diagnostics for characterisation of plasma actuators. Meas. Sci. Technol. **26**, 092001 (2015)
29. Benard, N., Moreau, E.: Electrical and mechanical characteristics of surface AC dielectric barrier discharge plasma actuators applied to airflow control. Exp. Fluids **55** 1846 (2014)
30. Orlov, D.M.: Modelling and simulation of single dielectric barrier discharge plasma actuators. Dissertation, University of Notre-Dame (2006)

31. Grundmann, S., Tropea, C.: Experimental transition delay using glow-discharge plasma actuators. Exp. Fluids **42**, 653–657 (2007)
32. Grundmann, S., Tropea, C.: Active cancellation of artificially introduced Tollmien-Schlichting waves using plasma actuators. Exp. Fluids **44**, 795–806 (2008)
33. Pereira, R., Timmer, W.A., de Oliveira, G., van Bussel, G.J.W.: Design of HAWT airfoils tailored for active flow control. Wind Energy **20**, 1569–1583 (2017)
34. Oliveira G., Pereira, R., Ragni, D., Kotsonis, M.: Modeling DBD plasma actuators in integral boundary layer formulation for application in panel methods. AIAA 2015-3367 (2015)
35. Asghar, A., Jumper, E., Corke, C.T.: On the use of Reynolds number as the scaling parameter for the performance of plasma actuator in a weakly compressible flow. AIAA 2006-0170 (2006)
36. Acheson, D.J.: Elementary Fluid Dynamics. Clarendon Press, Oxford (1990)
37. Batchelor, G.K.: An Introduction to Fluid Dynamics, 3rd edn. Cambridge University Press, Cambridge (2000)
38. Pereira, R., Ragni, D., Kotsonis, M.: Effect of external flow velocity on momentum transfer of dielectric barrier discharge plasma actuators. J. Appl. Phys. **116**(10), 103301 (2014)
39. Pereira, R., Kotsonis, M., de Oliveira, G., Ragni, D.: Analysis of local frequency response of flow to actuation: application to the dielectric barrier discharge plasma actuator. J. Appl. Phys. **118**(15), 153301 (2015)
40. White, F.M.: Viscous Fluid Flow, 2nd edn. McGraw-Hill, New York (1991)
41. Kundu, P.K., Cohen, I.M, Dowling D.R.: Fluid Mechanics, 6th edn. Elsevier, Burlington (2015)
42. Schlichting, H., Gersten, K.: Boundary Layer Theory, 8th edn. Springer, Berlin (2003)
43. Oleinik, O.A., Samokhin, V.N.: Mathematical Models in Boundary Layer Theory. Chapman & Hall/CRC Press, Boca Raton (1999)
44. van Rooij, R.P.J.O.M.: Modification of the boundary layer calculation in RFOIL for improved airfoil stall prediction. DUWIND Report IW-96087R (1996)
45. Drela, M., Giles, M.B.: Viscous-inviscid analysis of transonic and low Reynolds number airfoils. AIAA J. **25**(10), 1347–1355 (1987)
46. Hartree, D.R.: On an equation occurring in Falkner and Skan's approximate treatment of the equations of the boundary layer. Math. Proc. Camb. Philos. Soc. **33**(2), 233–239 (1937)
47. Brown, S.N.: Hartree's solutions of the Falkner-Skan equation. AIAA J. **4**(12), 2215–2216 (1966)
48. Cebeci, T., Keller, H.B.: Shooting and parallel shooting methods for solving the Falkner-Skan boundary layer equation. J. Comput. Phys. **7**, 289–300 (1971)
49. Asaitambi, A.: A finite-difference method for the Falkner-Skan equation. Appl. Math. Comput. **92**, 135–141 (1998)
50. Farell, P.A., Hegarty, A.F., Miller, J.J.H., O'riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman & Hall/CRC Press, Boca Raton (2000)
51. Elgazery, N.S.: Numerical solution for the Falkner-Skan equation. Chaos Solitons Fractals **35**, 738–746 (2006)
52. Kierzenka, J., Shampine, L.F.: A BVP solver based on residual control and the Matlab PSE. ACM Trans. Math. Softw. **27**(3), 299–316 (2001)

On Robust Error Estimation for Singularly Perturbed Fourth-Order Problems

Sebastian Franz and Hans-Görg Roos

Abstract Recently, several classes of fourth order singularly perturbed problems were considered and uniform convergence in the associated energy norm as well as in a balanced norm was proved. In this proceedings paper we will extend some results by looking into L^∞-bounds and postprocessing.

1 Introduction

In [3] several classes of fourth order problems were considered. In this proceedings paper we want to cover some extensions to it.

Consider the singularly perturbed plate bending problem for a clamped plate, given by the fourth-order differential equation

$$\varepsilon^2 \Delta^2 u - b\Delta u + (c \cdot \nabla)u + du = f \quad \text{in } \Omega := (0,1)^2, \tag{1a}$$

$$u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma := \partial\Omega \tag{1b}$$

where $b \geq b_0 > 1$, $d - \frac{1}{2}(\operatorname{div} c + \Delta b) \geq \delta > 0$ and $f \in L^2(\Omega)$ are smooth functions. In the given rectangular domain we have $u \in H_0^2(\Omega) \cap H^4(\Omega)$, see [1].

An alternative representation of above model is obtained by substituting $w = \varepsilon\Delta u \in H^2(\Omega)$ in order to obtain the system

$$w - \varepsilon\Delta u = 0,$$

$$\varepsilon\Delta w - bw + (c \cdot \nabla)u + du = f.$$

S. Franz (✉)
Institut Numerical Mathematics, Technische Universität Dresden, 01062 Dresden, Germany

Institut für Mathematik, BTU Cottbus, 03046 Cottbus, Germany
e-mail: sebastian.franz@tu-dresden.de

H.-G. Roos
Institut Numerical Mathematics, Technische Universität Dresden, 01062 Dresden, Germany
e-mail: hans-goerg.roos@tu-dresden.de

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_6

A weak formulation by using appropriate function spaces and the notation $\mathbf{c} = c + \nabla b$ is given by the following mixed method:

Find $(u, w) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$\varepsilon\langle\nabla u, \nabla\phi\rangle + \langle w, \phi\rangle = 0, \qquad \text{for all } \phi \in H^1(\Omega), \tag{2a}$$

$$\langle b\nabla u, \nabla\psi\rangle + \langle \mathbf{c}\cdot\nabla u + du, \psi\rangle - \varepsilon\langle\nabla w, \nabla\psi\rangle = \langle f, \psi\rangle, \quad \text{for all } \psi \in H_0^1(\Omega). \tag{2b}$$

In our paper we use the standard notation of Sobolev spaces, where $\|\cdot\|_0$ is the L^2-norm, $|\cdot|_k$ the seminorm in H^k and $\|\cdot\|_k$ the full H^k-norm. Furthermore, we denote by $\langle u, v\rangle_D$ the L^2-scalar product over a domain $D \subset \Omega$. If $D = \Omega$ we drop the subscript.

This weak formulation corresponds to the bilinear form $a(\cdot,\cdot)$ defined by

$$\begin{aligned} &a((u, w), (\psi, \phi)) \\ &\quad = \varepsilon\langle\nabla u, \nabla\phi\rangle + \langle w, \phi\rangle + \langle b\nabla u, \nabla\psi\rangle + \langle \mathbf{c}\cdot\nabla u + du, \psi\rangle - \varepsilon\langle\nabla w, \nabla\psi\rangle \end{aligned}$$

and (2) can be rewritten as: Find $(u, w) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$a((u, w), (\psi, \phi)) = \langle f, \psi\rangle \quad \text{for all } (\psi, \phi) \in H_0^1(\Omega) \times H^1(\Omega).$$

We will consider in this paper the mixed FEM of [3] applied to (2) and prove in Sect. 2 uniform convergence rates of this method for the component u in L^∞. Furthermore, the supercloseness result of [3] is used to facilitate a postprocessing approach in order to improve the convergence rate in the energy-norm. Section 3 provides an example supporting the theoretical results.

2 Numerical Analysis

Let us start by defining the energy norm

$$|||(u, w)|||^2 := \|w\|_0^2 + b_0\|\nabla u\|_0^2 + \delta\|u\|_0^2.$$

By [3, Lemma 3.1] we immediately have coercivity

$$a((u, w), (u, w)) \geq |||(u, w)|||^2$$

and therefore the uniqueness of the solution of (2). In order to derive robust error estimates we propose an assumption on the solution in the next section.

2.1 Solution Decomposition and Meshes

Let us assume a decomposition of the solution u into a smooth part S, boundary layers E_k with $k = 1, 2, 3, 4$ and corner layers E_k with $k = 12, 23, 34, 41$. More precisely, we assume for $0 \le i, j \le p + 2$

$$|\partial_x^i \partial_y^j S(x, y)| \le C, \qquad |\partial_x^i \partial_y^j E_1(x, y)| \le C\varepsilon^{1-i} \mathrm{e}^{-x/\varepsilon},$$
$$|\partial_x^i \partial_y^j E_2(x, y)| \le C\varepsilon^{1-j} \mathrm{e}^{-y/\varepsilon}, \qquad |\partial_x^i \partial_y^j E_{12}(x, y)| \le C\varepsilon^{1-i-j} \mathrm{e}^{-x/\varepsilon} \mathrm{e}^{-y/\varepsilon},$$

and similarly for the other components of the decomposition. Then we can construct a layer-adapted Shishkin mesh. (We could also use the generalisation of S-type meshes [7], that can give better numerical results, but in order to simplify the notation of the paper we stick to Shishkin meshes.) We define the so-called transition point

$$\lambda = \min \left\{ \sigma \varepsilon \ln N, \frac{1}{4} \right\}.$$

Note that it follows for this point

$$|E_1(\lambda, y)| \le C\varepsilon N^{-\sigma}.$$

These layers are therefore called *weak layers* as their influence vanishes with decreasing ε in a pointwise sense contrary to solutions of second order problems.

The interval $[0, 1]$ is now partitioned with a piecewise equidistant mesh, that is constructed by equidistantly dividing $[0, \lambda]$ into $N/4$ subintervals, $[\lambda, 1 - \lambda]$ into $N/2$ and $[1 - \lambda, 1]$ into $N/4$ subintervals again. The tensor product of two such 1d-meshes gives the Shishkin mesh.

On these meshes we consider the discrete space

$$V^{\mathcal{Q}} := \{v \in H^1(\Omega) : v|_\tau \in \mathcal{Q}_p(\tau) \forall \tau \in T_N\}, \quad V_0^{\mathcal{Q}} := V^{\mathcal{Q}} \cap H_0^1(\Omega).$$

Here $\mathcal{Q}_p(\tau)$ is the polynomial space on τ, with polynomial degrees at most p in each direction.

Now the discrete problem reads: Find $(u_h, w_h) \in V_0^{\mathcal{Q}} \times V^{\mathcal{Q}}$ such that

$$a((u_h, w_h), (\psi, \phi)) = \langle f, \psi \rangle \quad \text{for all } \phi \in V^{\mathcal{Q}},\ \psi \in V_0^{\mathcal{Q}}. \tag{3}$$

2.2 Error Estimation in L^∞

In [3, Theorem 3.6 and Lemma 3.7] we find the following error bound for the discrete error $|||(Iu - u_h, Jw - w_h)|||$, where I and J are standard interpolation

operators: If $\sigma \geq p+2$ we have

$$|||(Iu - u_h, Jw - w_h)||| \leq C(N^{-1} \ln N)^{p+1}. \tag{4}$$

From this estimate and the interpolation error estimate

$$\|Iu - u\|_{L^\infty(\Omega)} \leq C(N^{-1} \ln N)^{p+1}, \tag{5}$$

that can be proved similarly to the L^2-norm estimates in [3, Lemma 3.3], we obtain the following pointwise convergence result.

Theorem 1 *For $\sigma \geq p+2$ and $u_h \in V_0^Q$ it holds*

$$\|u - u_h\|_{L^\infty(\Omega)} \leq CK(N, \varepsilon)(N^{-1} \ln N)^{p+1},$$

where

$$K(N, \varepsilon) := (\ln N)^{1/2} + \min \left\{ N^{1/2}, \left(\ln \frac{N}{\varepsilon \ln N} \right)^{1/2} \right\}.$$

Proof We start with the triangle inequality and obtain

$$\|u - u_h\|_{L^\infty(\Omega)} \leq \|u - Iu\|_{L^\infty(\Omega)} + \|Iu - u_h\|_{L^\infty(\Omega)}.$$

The first term is already estimated by (5). For the second one we split the domain Ω into the layer regions and the non-layer region, see e.g. [3]. More precisely, we use the splitting

$$\begin{aligned}
\Omega_x^f &:= ([0, \lambda] \cup [1-\lambda, 1]) \times [\lambda, 1-\lambda], \\
\Omega_y^f &:= [\lambda, 1-\lambda] \times ([0, \lambda] \cup [1-\lambda, 1]), \\
\Omega^c &:= [\lambda, 1-\lambda]^2, \quad \Omega_{cor} := \Omega \setminus (\Omega_x^f \cup \Omega_y^f \cup \Omega^c).
\end{aligned}$$

Then we have

$$\|Iu - u_h\|_{L^\infty(\Omega)} = \max \left\{ \|Iu - u_h\|_{L^\infty(\Omega^c)}, \|Iu - u_h\|_{L^\infty(\Omega_x^f \cup \Omega_y^f)}, \|Iu - u_h\|_{L^\infty(\Omega_{cor})} \right\}.$$

For the first two terms we apply the discrete Sobolev inequality, see [5]. Let us start with Ω^c, where all cells have a diameter of order N^{-1}. It follows

$$\|Iu - u_h\|_{L^\infty(\Omega^c)} \leq C(1 + (\ln N)^{1/2}) \|\nabla(Iu - u_h)\|_{0,\Omega^c}.$$

In $\Omega_x^f \cup \Omega_y^f$ all cells have a diameter of order N^{-1} too, although they are anisotropic. We obtain

$$\|Iu - u_h\|_{L^\infty(\Omega_x^f \cup \Omega_y^f)} \leq C(1 + (\ln N)^{1/2}) \|\nabla(Iu - u_h)\|_{0,\Omega_x^f \cup \Omega_y^f}.$$

For the remaining norm over Ω_{cor} we can apply the same reasoning. Here the cells have a diameter of order $\varepsilon N^{-1} \ln N$ and therefore we obtain

$$\|Iu - u_h\|_{L^\infty(\Omega_{cor})} \leq C \left(1 + \left(\ln \frac{N}{\varepsilon \ln N}\right)^{1/2}\right) \|\nabla(Iu - u_h)\|_{0,\Omega_{cor}}.$$

There is an alternative way which we will show on the domain $\Omega_{cor}^1 = [0, \lambda]^2$ but which can be applied on the other parts of Ω_{cor} too. Due to $(Iu - u_h)|_\Gamma = 0$ we also have

$$\begin{aligned}
\|Iu - u_h\|_{L^\infty(\Omega_{cor}^1)} &\leq \sup_{y \in [0,\lambda]} \int_0^\lambda |\partial_x (Iu - u_h)(\xi, y)| \, \mathrm{d}\xi \\
&\leq C \left(\lambda \frac{N}{\lambda}\right)^{1/2} \|\partial_x (Iu - u_h)\|_{0,\Omega_{cor}^1} \\
&\leq C N^{1/2} \|\partial_x (Iu - u_h)\|_{0,\Omega_{cor}^1},
\end{aligned}$$

where an inverse inequality in y-direction was used. Combining all the previous estimates we obtain with $\|\nabla(Iu - u_h)\|_0 \leq |||Iu - u_h, Jw - w_h|||$

$$\begin{aligned}
\|Iu - u_h\|_{L^\infty(\Omega)} &\leq C \left((\ln N)^{1/2} + \min \left\{ N^{1/2}, \left(\ln \frac{N}{\varepsilon \ln N}\right)^{1/2} \right\}\right) |||Iu - u_h, Jw - w_h||| \\
&\leq C \left((\ln N)^{1/2} + \min \left\{ N^{1/2}, \left(\ln \frac{N}{\varepsilon \ln N}\right)^{1/2} \right\}\right) (N^{-1} \ln N)^{p+1}.
\end{aligned}$$

□

Note that we have the ε-uniform bounds

$$(\ln N)^{1/2} \leq K(N, \varepsilon) \leq C N^{1/2}$$

and the dependence of $K(N, \varepsilon)$ on ε is very weak. Actually, for $\varepsilon \geq 10^{-100}$ we have

$$K(N, \varepsilon) \leq 7.55 (\ln N)^{1/2}.$$

2.3 Postprocessing

The supercloseness result (4) can be used to define a better numerical solution, following the lines of [2, 6, 8]. The interpolation operators I and J are chosen as in [3].

As postprocessing operator we use an interpolation operator on a macro mesh. For that, suppose N is divisible by 4. We construct a coarser macro mesh $\tilde{T}^{N/2}$ composed of macro rectangles M, each consisting of four rectangles of T^N. The construction of these macro elements M is done such that the union on them covers Ω and none of them crosses the transition lines at λ_x and $1-\lambda_y$ for x or y.

The precise definition of the operator can be done in different ways. In [6] an operator is described that maps into $\mathcal{Q}_{p+2}$, thus increases the polynomial degree by 2. A minor modification is given in [8] for $p \geq 3$ to map into $\mathcal{Q}_{p+1}$. We will present another modification that is defined differently for even and odd values of $p \geq 1$ and maps always into $\mathcal{Q}_{p+1}$.

We describe the interpolation operator in 1d on the reference interval $[-1,1]$. The full operator is then a tensor product of two 1d interpolators mapped onto a macro cell M. For this purpose let $\hat{v}$ be the mapped function v on $[-1,1]$. Then $\widehat{P} : C[-1,1] \to \mathcal{P}_{p+1}[-1,1]$ is defined for odd p by

$$\widehat{P}\hat{v}(-1) = \hat{v}(-1), \qquad \widehat{P}\hat{v}(0) = \hat{v}(0), \qquad \widehat{P}\hat{v}(1) = \hat{v}(1),$$

and if $p \geq 3$

$$\int_{-1}^{1} (\widehat{P}\hat{v} - \hat{v})q = 0, \quad q \in \mathcal{P}_{p-2}([-1,1]).$$

For even $p \geq 2$ we change the definition to

$$\begin{aligned} \widehat{P}\hat{v}(-1) = \hat{v}(-1), \qquad \widehat{P}\hat{v}(1) = \hat{v}(1), \\ \int_{-1}^{0} (\widehat{P}\hat{v} - \hat{v}) = 0, \quad \int_{0}^{1} (\widehat{P}\hat{v} - \hat{v}) = 0, \\ \int_{-1}^{1} (\widehat{P}\hat{v} - \hat{v})q = 0, \quad q \in \mathcal{P}_{p-2}([-1,1]) \setminus \mathcal{P}_0([-1,1]). \end{aligned}$$

Finally, we set

$$P_M v = \widehat{P}_x \widehat{P}_y \hat{v},$$

where the subscript denotes the coordinate direction the 1d operator is applied to, and extend this piecewise projection to a global, continuous function by setting

$$\big(Pv\big)(x,y) := \big(P_M v\big)(x,y) \quad \text{for } (x,y) \in M.$$

Lemma 1 *For the postprocessing operator defined above we have*

$$PIu = Pu, \quad PJw = Pw, \qquad \text{for all } u,\, w \in C(\Omega)$$
$$|||(Pu^N, Pw^N)||| \le C\,|||(u^N, w^N)|||\,, \qquad \text{for all } u^N \in V_0^{\mathcal{Q}},\, w^N \in V^{\mathcal{Q}}.$$

For $\sigma \ge p+2$ it holds furthermore

$$|||(Pu - u, Pw - w)||| \le C(N^{-1}\ln N)^{p+1}.$$

Proof The proof follows the lines of e.g. [2, Lemma 5.1] for the consistency and stability, and [3, Lemma 3.3] for the interpolation error. □

Theorem 2 *We have for $\sigma \ge p+2$*

$$|||(u - Pu_h, w - Pw_h)||| \le C(N^{-1}\ln N)^{p+1}.$$

Proof Using the consistency and stability of P we obtain

$$\begin{aligned}|||(u - Pu_h, w - Pw_h)||| &\le |||(u - Pu, w - Pw)||| + |||(PIu - Pu_h, PJw - Pw_h)|||\\ &\le |||(u - Pu, w - Pw)||| + C\,|||(Iu - u_h, Jw - w_h)|||\,.\end{aligned}$$

Since both terms are already bounded by the right order, the proof is done. □

3 Numerical Experiments

Let us consider a problem, already investigated in [3, 4]. It is given by

$$\varepsilon^2\Delta^2 u - \Delta u = f \text{ in } \Omega = (0,1)^2, \tag{6a}$$
$$u = \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma = \partial\Omega, \tag{6b}$$

where f is given such that the exact solution is

$$u(x,y) = X(x)Y(y) \text{ where}$$
$$X(x) = \frac{1}{2}\left(\sin(\pi x) + \frac{\pi\varepsilon}{1-\mathrm{e}^{-1/\varepsilon}}\left(\mathrm{e}^{-x/\varepsilon} + \mathrm{e}^{(x-1)/\varepsilon} - 1 - \mathrm{e}^{-1/\varepsilon}\right)\right)$$
$$Y(y) = \left(2y(1-y^2) + \varepsilon\left(\ell d(1-2y) - 3\frac{q}{\ell} + \left(\frac{3}{\ell} - d\right)\mathrm{e}^{-y/\varepsilon} + \left(\frac{3}{\ell} + d\right)\mathrm{e}^{(y-1)/\varepsilon}\right)\right),$$

Table 1 Numerical results for example (6)

	N	$\|u-u_h\|_{L^\infty}$		$\|\|\|(u-u_h, v-v_h)\|\|\|$		$\|\|\|(u-Pu_h, v-Pv_h)\|\|\|$	
$\mathcal{Q}_1$	16	9.412e−03		1.125e−01		2.125e−02	
	32	2.590e−03	1.86	5.620e−02	1.00	5.457e−03	1.96
	64	6.837e−04	1.92	2.811e−02	1.00	1.439e−03	1.92
	128	1.756e−04	1.96	1.406e−02	1.00	3.892e−04	1.89
	256	4.445e−05	1.98	7.038e−03	1.00	1.075e−04	1.86
	512	1.117e−05	1.99	3.522e−03	1.00	3.015e−05	1.83
$\mathcal{Q}_2$	16	1.815e−04		4.682e−03		1.911e−03	
	32	2.323e−05	2.97	1.349e−03	1.80	5.822e−04	1.71
	64	2.913e−06	3.00	4.100e−04	1.72	1.498e−04	1.96
	128	3.643e−07	3.00	1.269e−04	1.69	3.227e−05	2.21
	256	4.543e−08	3.00	3.920e−05	1.69	6.213e−06	2.38
	512	5.663e−09	3.00	1.200e−05	1.71	1.117e−06	2.47
$\mathcal{Q}_3$	16	4.436e−06		6.650e−04		8.930e−04	
	32	3.740e−07	3.57	1.968e−04	1.76	2.308e−04	1.95
	64	7.101e−08	2.40	4.700e−05	2.07	4.162e−05	2.47
	128	1.034e−08	2.78	9.721e−06	2.27	5.766e−06	2.85
	256	1.261e−09	3.04	1.841e−06	2.40	7.007e−07	3.04
	512	1.364e−10	3.21	3.293e−07	2.48	8.530e−08	3.04

with $\ell = 1 - \mathrm{e}^{-1/\varepsilon}$, $q = 2 - \ell$ and $d = 1/(q - 2\varepsilon\ell)$. As parameter for the Shishkin mesh we set $\sigma = p + 2$ and fix $\varepsilon = 10^{-4}$ (uniformity in ε was already investigated in [3]).

Table 1 shows the results for $p = 1$ and $p = 2$. The observed results in L^∞ are even better than expected, as a convergence of $N^{-(p+1)}$ can be seen instead of the bound from Theorem 1. The last column shows quite nicely the improvement in the energy norm by postprocessing the numerical solution. The table also shows the results for $p = 3$, where the rates in L^∞ and for the postprocessed solution are not as good as expected. A possible explanation is, that the solution to the given problem is not smooth enough for the assumption on the solution decomposition to hold for high derivatives. Although the derivatives of the given solution u fulfil the *sum* of the estimates, it is not enough, as the *existence* of a decomposition is not clear. It is an open question, which compatibility and regularity conditions are needed, such that a solution decomposition like the one presented in Sect. 2.1 exists.

References

1. Blum, H., Rannacher, R.: On the boundary value problem of the biharmonic operator on domains with angular corners. Math. Methods Appl. Sci. **2**(4), 556–581 (1980)
2. Franz, S.: Superconvergence using pointwise interpolation in convection-diffusion problems. Appl. Numer. Math. **76**, 132–144 (2014)

3. Franz, S., Roos, H.-G.: Robust error estimation in energy and balanced norms for singularly perturbed fourth order problems. Comput. Math. Appl. **72**, 233–247 (2016)
4. Franz, S., Roos, H.-G., Wachtel, A.: A C^0 interior penalty method for a singularly-perturbed fourth-order elliptic problem on a layer-adapted mesh. Numer. Methods Partial Differ. Equ. **30**(3), 838–861 (2014)
5. Kopteva, N.: The two-dimensional Sobolev inequality in the case of an arbitrary grid. Zh. Vychisl. Mat. Mat. Fiz. **38**(4), 596–599 (1998)
6. Lin, Q., Yan, N., Zhou, A.: A rectangle test for interpolated element analysis. In: Proc. Syst. Sci. Eng., pp. 217–229. Great Wall (H.K.) Culture Publish Co. (1991)
7. Roos, H.-G., Linß, T.: Sufficient conditions for uniform convergence on layer-adapted grids. Computing **63**, 27–45 (1999)
8. Tobiska, L.: Analysis of a new stabilized higher order finite element method for advection-diffusion equations. Comput. Methods Appl. Mech. Eng. **196**, 538–550 (2006)

Singularly Perturbed Initial-Boundary Value Problems with a Pulse in the Initial Condition

José Luis Gracia and Eugene O'Riordan

Abstract A singularly perturbed parabolic equation of reaction-diffusion type is examined. Initially the solution approximates a concentrated source, which causes an interior layer to form within the solution for all future times. Combining a classical finite difference operator with a layer-adapted mesh, parameter-uniform convergence is established. Numerical results are presented to illustrate the theoretical error bounds.

1 Introduction

In [3], a singularly perturbed parabolic problem, of convection diffusion type,

$$-\varepsilon u_{xx} + a u_x + b u + c u_t = f, \quad \varepsilon, a(x,t), b(x,t), c(x,t) > 0,$$

with a layer (having a Gaussian profile) present in the initial condition $u(x,0) = \phi(x;\varepsilon)$, was examined. The initial layer induced an interior layer in the solution of the parabolic problem. To establish that the numerical method (constructed in [3]) was parameter-uniform [2], the scale of the initial layer was set to be of order $O(\sqrt{\varepsilon})$; in other words, the scale of the initial layer corresponded to the scale of any interior layer present in the solution. In this paper, we examine the possibility of an initial layer of a different scale being transported through time. To simplify the matter, we consider a parabolic problem with no convection present. In order to

The research of the author "J.L. Gracia" was partly supported by the Institute of Mathematics and Applications (IUMA), the project MTM2016-75139-R and the Diputación General de Aragón.

J.L. Gracia (✉)
Department of Applied Mathematics, University of Zaragoza, Zaragoza, Spain
e-mail: jlgracia@unizar.es

E. O'Riordan
School of Mathematical Sciences, Dublin City University, Dublin, Ireland
e-mail: eugene.oriordan@dcu.ie

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_7

retain parameter-uniform convergence, it is established below that the layer width in the initial condition can have a scale wider than the scale induced by the differential equation. However, if the scale of the initial layer is significantly thinner than the scale of any interior layer then the rate of convergence is adversely effected by the presence of such an excessively thin layer in the initial condition, when a uniform mesh in time is utilized in the numerical method. Numerical results for a numerical method utilizing a particular piecewise-uniform mesh in both space and time suggest a potential improvement in the convergence rate in the case of a very thin pulse. In this paper C denotes a generic constant that is independent of the parameter ε and the mesh parameters N and M. For any function z, we set $\|z\|_{\bar{G}} := \max_{(x,t)\in\bar{G}} |z(x,t)|$.

2 Reaction-Diffusion Problem

Consider the following singularly perturbed parabolic problem of reaction-diffusion type : Find u such that

$$Lu := (-\varepsilon u_{xx} + bu + cu_t)(x,t) = f(x,t),\ (x,t) \in \Omega := (-1,1)\times(0,T], \quad \text{(1a)}$$

$$u(x,0) = g_1(x) + g_2(x)e^{-\theta\frac{x^2}{\varepsilon}}, -1 \le x \le 1, \quad \theta > 0; \quad \text{(1b)}$$

$$u(-1,t) = \phi_L(t),\ u(1,t) = \phi_R(t),\ 0 < t \le T, \quad b(x,t) \ge 0,\ c(x,t) > 0; \quad \text{(1c)}$$

$$g_2^{(i)}(-1) = g_2^{(i)}(1) = 0, \qquad i = 0,1,2; \quad \text{(1d)}$$

where $b(x,t), c(x,t), f(x,t), g_1(x), g_2(x)$ are sufficiently smooth functions. In this problem, in contrast to the case of a convection-diffusion problem [3], there are no immediate restrictions on the final time T, as the interior layer will not interact with the boundaries of the domain. However, the bounds in the final error estimate given in Theorem 2 do depend on $e^{\theta T}$ and hence, for $\theta > 1$, these bounds become large as T increases.

To highlight the interplay between the width of the pulse and the scale of the layers emanating from the presence of the singular perturbation parameter in the differential equation, we consider the following simple problem

$$-\varepsilon u_{xx} + u_t = 0, \quad (x,t) \in Q := (-1,1)\times(0,0.5], \quad \text{(2a)}$$

$$u(x,0) = (1-x^2)^2 e^{-\frac{\theta x^2}{\varepsilon}},\ -1 \le x \le 1; \quad u(-1,t) = u(1,t) = 0,\ 0 < t \le 0.5. \quad \text{(2b)}$$

A closed form representation of the solution of this problem is

$$u(x,t) = \frac{1}{2\sqrt{\varepsilon\pi t}} e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}} \int_{s=-1}^{1} (1-s^2)^2 e^{-\frac{(1+4\theta t)}{4\varepsilon t}\left(\frac{x}{1+4\theta t}-s\right)^2} ds.$$

As $\frac{\varepsilon}{\theta} \to 0^+$, we note that the solution behaves like

$$u(x,t) \to \frac{1}{\sqrt{1+4\theta t}} e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}} \left(1 - \frac{x^2}{(1+4\theta t)^2}\right)^2.$$

Observe that the layer width initially is visible in the bound

$$e^{-\frac{\theta x^2}{\varepsilon}} \le e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}} \le e^{-\frac{\theta x^2}{2\varepsilon}}, \qquad 0 < t \le \frac{1}{4\theta};$$

and the range of applicability for this inequality increases as the value of θ decreases. For any $\theta \ge 1$ we also have, for intermediate values of time, that

$$e^{-\frac{x^2}{\varepsilon}} \le e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}} \le e^{-\frac{x^2}{2\varepsilon}}, \qquad \frac{1}{4} - \frac{1}{4\theta} \le t \le \frac{1}{2} - \frac{1}{4\theta}, \qquad \theta \ge 1.$$

The layer width associated with the function $e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}}$ evolves from an initial width of $\mathcal{O}(\sqrt{\varepsilon/\theta})$ to a width of $\mathcal{O}(\sqrt{\varepsilon})$ as time increases, in the case where $\theta \ge 1$. In the case where $\theta < 1$, we simply have

$$e^{-\frac{x^2}{5t\varepsilon}} \le e^{-\frac{\theta x^2}{\varepsilon(1+4\theta t)}} \le e^{-\frac{x^2}{6t\varepsilon}}, \qquad \frac{1}{2\theta} \le t \le \frac{1}{\theta}.$$

Over a finite time range $4T\theta \le 1$, there will be no significant change in the layer width for $\theta < 1$. In both cases $\theta < 1, \theta \ge 1$, note that the amplitude of the pulse at $x = 0$ decreases with time and with respect to θ. This effect is illustrated in the two figures displayed in Fig. 1. Finally, note that $u_t(0,0) < 0$, but $u_t(x,0) > 0,\ x \in$

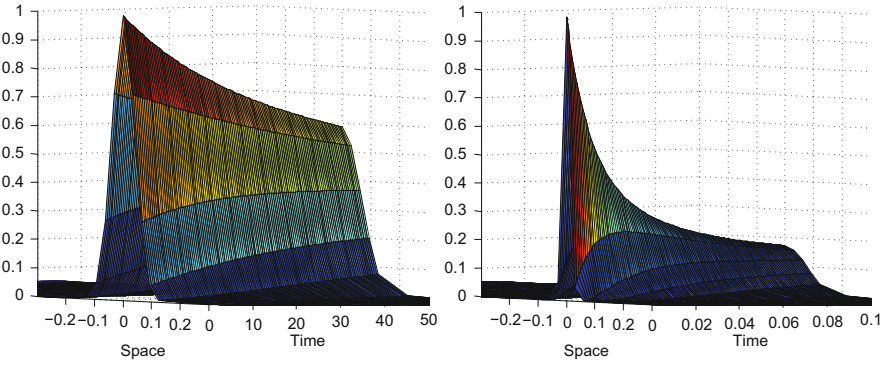

Fig. 1 Problem (2): zoom into the interior layer region of the computed solution $U^{N,M}$ generated by the numerical scheme (5) for $N = 32, M = 128$. In the *left figure*, $\theta = 0.01, T = 50$ and $\varepsilon = 2^{-15}$ and in the *right figure* $\theta = 100,\ T = 0.1$ and $\varepsilon = 2^{-5}$

$(-1,1) \setminus (-C\sqrt{\varepsilon/\theta}, C\sqrt{\varepsilon/\theta})$. Hence, the initial time derivative has different signs within and outside the layer region.

In our subsequent numerical analysis, we shall see that a piecewise-uniform Shishkin mesh, with a transition point related to the width of the pulse, coupled with a uniform mesh in time, suffices to obtain parameter-uniform convergence only in the case where $\theta \leq 1$.

3 Bounds on the Derivatives of the Continuous Solution

The general solution of (1) can be decomposed into the sum $u = v + w_L + w_R + z$, where $v, w_L, w_R, z \in \mathscr{C}^{4+\gamma}(\bar{\Omega})$ and

$$\begin{aligned}
Lv &= f, \;\; v(x,0) = g_1(x), \quad v(-1,t), v(1,t) \quad \text{suitably chosen;} \\
Lw_L &= 0, \;\; w_L(-1,t) = (u-v)(-1,t), \quad w_L(1,t) = 0, \; w_L(x,0) = 0; \\
Lw_R &= 0, \;\; w_R(1,t) = (u-v)(1,t), \quad w_R(-1,t) = 0, \; w_R(x,0) = 0; \\
Lz &= 0, \;\; z(x,0) = g_2(x) e^{-\theta \frac{x^2}{\varepsilon}}, \quad z(-1,t) = z(1,t) = 0.
\end{aligned}$$

For the regular and boundary layer components the following bounds can be established, as in [5]: For $0 \leq j + 2m =: n \leq 4$,

$$\left\| \frac{\partial^{j+m} v}{\partial x^j \partial t^m} \right\|_{\Omega} \leq C \left(1 + \varepsilon^{1-j/2}\right), \tag{3a}$$

$$\left| \frac{\partial^{j+m} w_L}{\partial x^j \partial t^m}(x,t) \right| \leq C\varepsilon^{-j/2} e^{-\frac{(1+x)}{\sqrt{\varepsilon}}}, \qquad \left| \frac{\partial^{j+m} w_R}{\partial x^j \partial t^m}(x,t) \right| \leq C\varepsilon^{-j/2} e^{-\frac{(1-x)}{\sqrt{\varepsilon}}}. \tag{3b}$$

In passing we note that the interior layer component z is smoother than in the case of the convection-diffusion problem [3] as $[z](0,t) = [z_x](0,t) = 0$.

Theorem 1 *Assume that $\theta T \leq C$ and $\theta \geq C\varepsilon$. For $0 \leq j + 2m =: n \leq 4$,*

$$|z(x,t)| \leq C e^{\theta t/c_0} e^{-\frac{\sqrt{\theta}|x|}{\sqrt{\varepsilon}}}, \tag{4a}$$

$$\left\| \frac{\partial^{j+m} z}{\partial x^j \partial t^m} \right\|_{\Omega} \leq C e^{\theta T/c_0} (1 + \theta^{n/2}) \varepsilon^{-j/2}, \tag{4b}$$

where $c_0 := \min c(x,t)$. In addition, if $C\varepsilon \leq \theta \leq 1$, then

$$\left\| \frac{\partial^{j+m} z}{\partial x^j \partial t^m} \right\|_{\Omega} \leq C \varepsilon^{-j/2} \theta^{j/2}. \tag{4c}$$

Proof From the maximum principle, $|z(x,t)| \le Ce^{\frac{t}{c_0}}$, $\forall (x,t) \in \Omega$. Note that for all s and any $\kappa > 0$

$$e^{-\kappa s^2} \le e^{\frac{1}{4}} e^{-\sqrt{\kappa}|s|}.$$

Then using the obvious barrier functions, we establish the bounds (4a) on z separately on $\Omega^- = [-1,0] \times [0,T]$ and $\Omega^+ = [0,1] \times [0,T]$, while noting that $|z(0,t)| \le Ce^{\frac{t}{c_0}}$ has been already established. In order to obtain parameter-explicit bounds on the derivatives of z in the entire region Ω, and to deal with the cases of $c\varepsilon \le \theta < 1$ and $\theta \ge 1$ together, we introduce the stretched variables

$$\eta := \frac{\sqrt{\theta_*}x}{\sqrt{\varepsilon}}, \quad \tau = \theta_* t \text{ with } \theta_* := \max\{1, \theta\} \text{ and } \breve{u}(\eta,\tau) := u(s,t).$$

Hence the differential equation can be written in the form

$$-\breve{z}_{\eta\eta} + \frac{\breve{b}}{\theta_*}\breve{z} + \breve{c}\,\breve{z}_\tau = 0, \quad (\eta,\tau) \in \left(-\sqrt{\theta_*}/\sqrt{\varepsilon}, \sqrt{\theta_*}/\sqrt{\varepsilon}\right) \times (0, \theta_* T],$$

with zero boundary conditions and an initial condition of the form

$$\breve{z}(\eta,0) = g_2\left(\sqrt{\varepsilon}\eta/\sqrt{\theta}\right)e^{-\eta^2} \text{ if } \theta_* = \theta, \text{ and } \breve{z}(\eta,0) = g_2(\sqrt{\varepsilon}\eta)e^{-\theta\eta^2} \text{ if } \theta_* = 1.$$

To obtain bounds on the derivatives of the interior layer, we now use the interior estimates from [4, p. 352], to deduce that

$$\left|\frac{\partial^{j+m}\breve{z}}{\partial\eta^j\partial\tau^m}(\eta,\tau)\right| \le Ce^{\theta T/c_0} + C\left|\frac{\partial^j \breve{z}}{\partial\eta^j}(\eta,0)\right|.$$

Returning to the original variables, we get

$$\left\|\frac{\partial^{j+m}z}{\partial x^j\partial t^m}\right\|_\Omega \le Ce^{\theta T/c_0}\theta^m\left(\frac{\theta}{\varepsilon}\right)^{j/2}\left(1+\sqrt{\frac{\varepsilon}{\theta}}\right)^j \le Ce^{\theta T/c_0}\theta^{(j+2m)/2}\varepsilon^{-j/2}, \quad \text{if } \theta_* = \theta,$$

$$\left\|\frac{\partial^{j+m}z}{\partial x^j\partial t^m}\right\|_\Omega \le Ce^{\theta T/c_0}\varepsilon^{-j/2}\left(\sqrt{\varepsilon}+\sqrt{\theta}\right)^j \le Ce^{\theta T/c_0}\varepsilon^{-j/2}\theta^{j/2}, \quad \text{if } \theta_* = 1.$$

In the last inequality we have used the fact that $C\varepsilon \le \theta$.

Remark 1 Note that in the case where $\theta \le C\varepsilon$, then in the above proof, we can replace θ_* by θ and consequently deduce that all the partial derivatives of z are uniformly bounded. Thus, when $\theta \le C\varepsilon$, there is no interior layer present.

4 Numerical Method and Error Analysis

We employ a classical fully implicit finite difference operator on a piecewise-uniform Shishkin mesh [2]. The finite difference scheme is given by

$$L^{N,M}U := -\varepsilon\delta_x^2 U + bU + cD_t^- U = f(x_i,t_j), \quad (x_i,t_j) \in \Omega^{N,M}, \tag{5a}$$

$$U(x_i,0) = u(x_i,0), \quad U(-1,t_j) = u(-1,t_j),\ U(1,t_j) = u(1,t_j). \tag{5b}$$

Note that nothing special is required at the mesh points $(0,t_j)$, where the interior layer is located. Based on the above bounds on the layer components of the solution, we split the space domain $[-1,1]$ into the five subintervals

$$[-1,-1+\sigma_R] \cup [-1+\sigma_R,-\tau] \cup [-\tau,\tau] \cup [\tau,1-\sigma_R] \cup [1-\sigma_R,1], \tag{5c}$$

$$\tau := \min\left\{\frac{1}{8}, 2\frac{\sqrt{\varepsilon}}{\sqrt{\theta}}\ln N\right\},\ \sigma_R := \min\left\{\frac{1}{8}, 2\sqrt{\varepsilon}\ln N\right\}, \tag{5d}$$

where N is the spatial discretisation parameter. The grid points, in space, are uniformly distributed within each subinterval such that

$$-x_0 = x_N = 1,\ -x_{N/8} = x_{7N/8} = 1-\sigma_R,\ -x_{3N/8} = x_{5N/8} = \tau,\ x_{N/2} = 0.$$

We use M mesh elements uniformly distributed in time and the mesh $\bar{\Omega}^{N,M}$ is the tensor product of the spatial and time meshes.

The discrete counterparts of the components v, w_L, w_R and z are denoted by V, W_L, W_R and Z, which are defined in a standard way. Bounds on the errors in the discrete regular component (V), boundary layers components (W_L and W_R) follow from a standard truncation argument and suitable barrier functions [1, 5]. Thus, we have that

$$\|V-v\|_{\bar{\Omega}^{N,M}} \le C(N^{-1}\ln N)^2 + CM^{-1}, \quad \|W_L - w_L\|_{\bar{\Omega}^{N,M}} \le CN^{-2} + CM^{-1}, \tag{6}$$

and the boundary layer component W_R satisfies similar error estimates as W_L. For $M \ge \mathcal{O}(\ln N)$, the discrete interior layer function satisfies the bounds

$$(a) \qquad |Z(x_i,t_j)| \le Ce^{\theta T/c_0} \prod_{k=i}^{N/2}\left(1+\frac{\sqrt{\theta}h_k}{2\sqrt{\varepsilon}}\right), \quad x_i \le 0,$$

$$(b) \qquad |Z(x_i,t_j)| \le Ce^{\theta T/c_0} \prod_{k=N/2}^{i}\left(1+\frac{\sqrt{\theta}h_k}{2\sqrt{\varepsilon}}\right)^{-1}, \quad x_i \ge 0,$$

where $h_k := x_k - x_{k-1}$, for $k = 1, 2, \ldots, N$. From these bounds we establish that when $8\tau < 1$

$$|Z(x_i, t_j)| \leq Ce^{\theta T/c_0} N^{-2}, \quad x_i \in (-1, 1) \setminus (-\tau, \tau).$$

In addition, for $x_i \in (-\tau, \tau)$

$$|L^{N,M}(Z - z)(x_i, t_j)| \leq \begin{cases} Ce^{\theta T/c_0} \left(\theta (N^{-1} \ln N)^2 + \theta^2 M^{-1}\right), & \text{if } 1 \leq \theta, \\ Ce^{\theta T/c_0} \left(\theta (N^{-1} \ln N)^2 + M^{-1}\right), & \text{if } \theta < 1. \end{cases}$$

Using a suitable barrier function we deduce that

$$\|Z - z\|_{(-\tau,\tau)} \leq \begin{cases} Ce^{\theta T/c_0} \left(\theta (N^{-1} \ln N)^2 + \theta^2 M^{-1}\right), & \text{if } 1 \leq \theta, \\ Ce^{\theta T/c_0} \left(\theta (N^{-1} \ln N)^2 + M^{-1}\right), & \text{if } \theta < 1. \end{cases} \tag{7}$$

From the error bounds (6) and (7), the following nodal error bound follows by the triangular inequality.

Theorem 2 *Assume $M \geq \mathcal{O}(\ln N)$. Let be U the solution of the discrete problem (5) and u the solution of the continuous problem (1). Then,*

$$\|U - u\|_{\bar{\Omega}^{N,M}} \leq \begin{cases} Ce^{\theta T/c_0} \left(\theta (N^{-1} \ln N)^2 + \theta^2 M^{-1}\right), & \text{if } 1 \leq \theta, \\ Ce^{\theta T/c_0} \left((N^{-1} \ln N)^2 + M^{-1}\right), & \text{if } \theta < 1. \end{cases}$$

5 Numerical Experiments

Consider the following test problem

$$-\varepsilon u_{xx} + u + u_t = (1 - x^2)t, \quad (x, t) \in Q := (-1, 1) \times (0, 1], \tag{8}$$

$$u(x, 0) = (1 - x^2)^2 (1 + x)^2 e^{-\frac{\theta x^2}{\varepsilon}}, \ -1 \leq x \leq 1, \ u(-1, t) = u(1, t) = 0, \ 0 < t \leq 1,$$

where we shall consider some sample values for the parameter θ. In Figs. 2 and 3 we display the computed solutions generated by the numerical scheme (5) with $\varepsilon = 2^{-5}, 2^{-10}$ and $N = M = 32$. The values of the parameter θ in the initial condition are $\theta = 1$ and $\theta = 100$. We observe again the influence of this parameter in the profile of the solution. Note that the time derivative $|u_t(0, 0)|$ increases with θ.

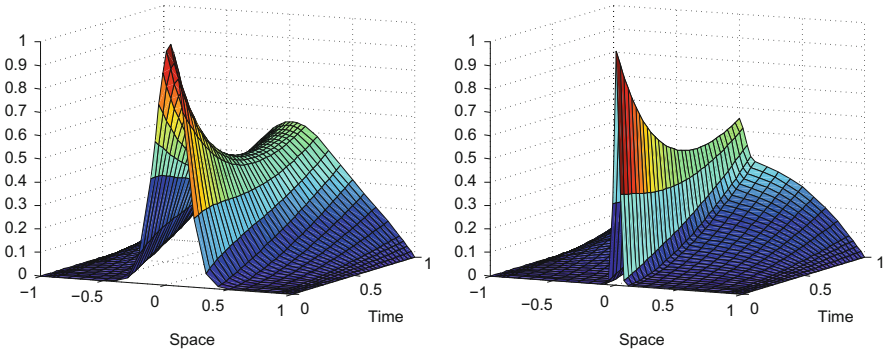

Fig. 2 Test problem (8): computed solution $U^{N,M}$ generated by the numerical scheme (5) for $N = M = 32$, $\theta = 1$ and $\varepsilon = 2^{-5}$ (*left figure*) and $\varepsilon = 2^{-10}$ (*right figure*)

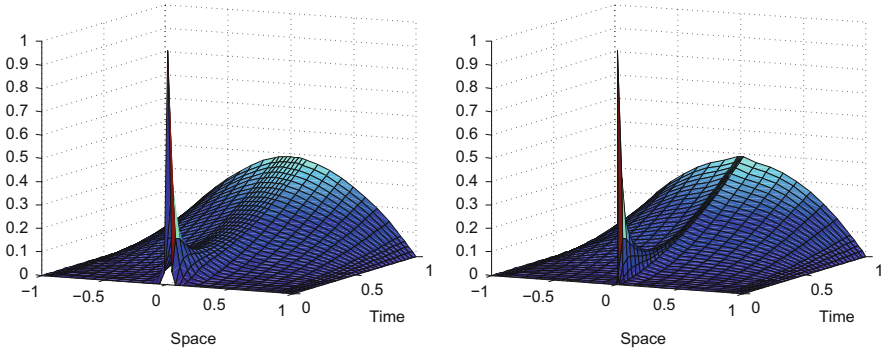

Fig. 3 Test problem (8): computed solution $U^{N,M}$ generated by the numerical scheme (5) for $N = M = 32$, $\theta = 100$ and $\varepsilon = 2^{-5}$ (*left figure*) and $\varepsilon = 2^{-10}$ (*right figure*)

The exact solution of this problem is unknown and we use the two-mesh principle [2] to estimate the orders of convergence by first computing the two-mesh differences

$$F_{\varepsilon}^{N,M} := \max\left\{\|U^{N,M} - \bar{U}^{2N,2M}\|_{\bar{\Omega}^{N,M}}, \|\bar{U}^{N,M} - U^{2N,2M}\|_{\bar{\Omega}^{2N,2M}}\right\},$$

where $\bar{U}^{N,M}$ denotes the bilinear interpolant of the solution. These values are used to compute the approximate orders of global convergence using

$$Q_{\varepsilon}^{N,M} := \log_2(F_{\varepsilon}^{N,M}/F_{\varepsilon}^{2N,2M}).$$

Table 1 Numerical method (5): computed two-mesh differences $F_{\varepsilon}^{N,M}$ and uniform differences $F^{N,M}$ with their corresponding orders of convergence $Q_{\varepsilon}^{N,M}$, $Q^{N,M}$ for problem (8) with $\theta = 1$

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\varepsilon = 2^{0}$	0.742E−01	0.528E−01	0.339E−01	0.197E−01	0.107E−01	0.557E−02
	0.492	0.639	0.785	0.882	0.939	
$\varepsilon = 2^{-2}$	0.558E−01	0.326E−01	0.179E−01	0.944E−02	0.486E−02	0.247E−02
	0.777	0.867	0.920	0.957	0.978	
$\varepsilon = 2^{-4}$	0.373E−01	0.225E−01	0.125E−01	0.656E−02	0.337E−02	0.171E−02
	0.727	0.855	0.925	0.961	0.980	
$\varepsilon = 2^{-6}$	0.556E−01	0.228E−01	0.118E−01	0.602E−02	0.304E−02	0.153E−02
	1.284	0.949	0.974	0.987	0.993	
$\varepsilon = 2^{-8}$	0.698E−01	0.317E−01	0.141E−01	0.652E−02	0.312E−02	0.152E−02
	1.140	1.164	1.117	1.064	1.035	
$\varepsilon = 2^{-10}$	0.110E+00	0.615E−01	0.239E−01	0.922E−02	0.381E−02	0.169E−02
	0.836	1.360	1.376	1.277	1.173	
$\varepsilon = 2^{-12}$	0.861E−01	0.103E+00	0.562E−01	0.196E−01	0.660E−02	0.241E−02
	−.264	0.881	1.522	1.568	1.456	
$\varepsilon = 2^{-14}$	0.827E−01	0.847E−01	0.672E−01	0.304E−01	0.116E−01	0.427E−02
	−.033	0.332	1.145	1.387	1.445	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\varepsilon = 2^{-30}$	0.827E−01	0.816E−01	0.662E−01	0.302E−01	0.116E−01	0.426E−02
	0.020	0.301	1.134	1.381	1.442	
$F^{N,M}$	0.110E+00	0.103E+00	0.677E−01	0.311E−01	0.116E−01	0.557E−02
$Q^{N,M}$	0.085	0.612	1.121	1.420	1.063	

The uniform global orders of convergence are estimated by computing

$$F^{N,M} := \max_{\varepsilon \in S} F_{\varepsilon}^{N,M}, \quad Q^{N,M} := \log_2(F^{N,M}/F^{2N,2M}),$$

with $S = \{2^{0}, 2^{-1}, 2^{-2}, \ldots, 2^{-30}\}$. The numerical results presented in Tables 1 and 2 are in line with our theoretical findings. In Table 3 we fix the value of the singular perturbation parameter to $\varepsilon = 2^{-16}$ and we take different values of $\theta = 2^{-20}, 2^{-18}, \ldots, 2^{10}$. We observe that the method is convergent but the maximum two-mesh differences are greater as θ increases and the orders of convergence have deteriorated.

Table 2 Numerical method (5): computed two-mesh differences $F_{\varepsilon}^{N,M}$ and uniform differences $F^{N,M}$ with their corresponding orders of convergence $Q_{\varepsilon}^{N,M}$, $Q^{N,M}$ for problem (8) with $\theta = 100$

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\varepsilon = 2^{0}$	0.587E−01	0.526E−01	0.436E−01	0.461E−01	0.467E−01	0.399E−01
	0.156	0.273	−.080	−.020	0.226	
$\varepsilon = 2^{-2}$	0.557E−01	0.517E−01	0.428E−01	0.462E−01	0.463E−01	0.394E−01
	0.108	0.271	−.108	−.005	0.236	
$\varepsilon = 2^{-4}$	0.781E−01	0.572E−01	0.444E−01	0.470E−01	0.466E−01	0.393E−01
	0.451	0.363	−.080	0.011	0.244	
$\varepsilon = 2^{-6}$	0.122E+00	0.703E−01	0.503E−01	0.501E−01	0.478E−01	0.398E−01
	0.800	0.484	0.006	0.067	0.265	
$\varepsilon = 2^{-8}$	0.122E+00	0.703E−01	0.497E−01	0.508E−01	0.486E−01	0.403E−01
	0.799	0.502	−.033	0.064	0.270	
$\varepsilon = 2^{-10}$	0.122E+00	0.703E−01	0.497E−01	0.508E−01	0.486E−01	0.403E−01
	0.798	0.502	−.032	0.064	0.270	
$\varepsilon = 2^{-12}$	0.122E+00	0.703E−01	0.497E−01	0.508E−01	0.486E−01	0.403E−01
	0.797	0.502	−.031	0.064	0.269	
$\varepsilon = 2^{-14}$	0.122E+00	0.703E−01	0.497E−01	0.507E−01	0.485E−01	0.403E−01
	0.797	0.502	−.031	0.064	0.269	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\varepsilon = 2^{-30}$	0.122E+00	0.703E−01	0.497E−01	0.507E−01	0.485E−01	0.403E−01
	0.797	0.502	−.030	0.064	0.269	
$F^{N,M}$	0.127E+00	0.703E−01	0.503E−01	0.509E−01	0.486E−01	0.404E−01
$Q^{N,M}$	0.849	0.484	−.016	0.065	0.269	

Remark 2 Given the initial large time derivatives visible in Fig. 3 and also present in the bounds (4b) on the time derivatives of the solution, it is natural to consider a piecewise uniform mesh in time where the transition parameter is taken to be

$$\tau_t := \min\{T/2, (1/\theta)\ln M\}, \tag{9}$$

and to distribute uniformly $M/2+1$ points in the time subdomains $[0, \tau_t]$ and $[\tau_t, T]$. We repeat only two of the previous Tables 2 (where $\theta = 100$) and 3 (where $\varepsilon = 2^{-16}$); their companion tables are Tables 4 and 5. We observe an improvement in the numerical results compared to using a uniform mesh in time. The question of whether the inclusion of a piecewise-uniform mesh in time produces a parameter-uniform (with respect to both ε and θ) remains an open question.

Table 3 Numerical method (5): computed two-mesh differences $F_{\varepsilon}^{N,M}$ and their corresponding orders of convergence $Q_{\varepsilon}^{N,M}$ for problem (8) with $\varepsilon = 2^{-16}$

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\theta = 2^{-16}$	0.317E−01	0.130E−01	0.576E−02	0.269E−02	0.130E−02	0.638E−03
	1.290	1.171	1.096	1.051	1.027	
$\theta = 2^{-12}$	0.266E−01	0.134E−01	0.557E−02	0.236E−02	0.106E−02	0.517E−03
	0.988	1.265	1.237	1.156	1.037	
$\theta = 2^{-8}$	0.656E−01	0.225E−01	0.771E−02	0.292E−02	0.123E−02	0.553E−03
	1.547	1.542	1.402	1.252	1.146	
$\theta = 2^{-4}$	0.904E−01	0.105E+00	0.544E−01	0.170E−01	0.494E−02	0.152E−02
	−.209	0.943	1.681	1.778	1.701	
$\theta = 2^{0}$	0.827E−01	0.831E−01	0.667E−01	0.303E−01	0.116E−01	0.427E−02
	−.007	0.317	1.140	1.384	1.443	
$\theta = 2^{2}$	0.113E+00	0.829E−01	0.683E−01	0.355E−01	0.159E−01	0.685E−02
	0.449	0.280	0.946	1.158	1.214	
$\theta = 2^{4}$	0.127E+00	0.772E−01	0.690E−01	0.470E−01	0.281E−01	0.154E−01
	0.714	0.162	0.553	0.744	0.869	
$\theta = 2^{6}$	0.132E+00	0.694E−01	0.562E−01	0.536E−01	0.456E−01	0.339E−01
	0.932	0.306	0.069	0.231	0.430	
$\theta = 2^{8}$	0.890E−01	0.640E−01	0.570E−01	0.479E−01	0.449E−01	0.473E−01
	0.476	0.167	0.251	0.093	−.075	
$\theta = 2^{10}$	0.437E−01	0.411E−01	0.493E−01	0.532E−01	0.523E−01	0.459E−01
	0.090	−.263	−.108	0.023	0.189	

Table 4 Finite difference scheme (5) coupled with a piecewise-uniform mesh in time (9): computed two-mesh differences $F_{\varepsilon}^{N,M}$ and uniform differences $F^{N,M}$ with their corresponding orders of convergence $Q_{\varepsilon}^{N,M}$, $Q^{N,M}$ for problem (8) with $\theta = 100$

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\varepsilon = 2^{0}$	0.844E−01	0.565E−01	0.339E−01	0.183E−01	0.912E−02	0.513E−02
	0.579	0.738	0.886	1.006	0.830	
$\varepsilon = 2^{-2}$	0.917E−01	0.593E−01	0.347E−01	0.185E−01	0.919E−02	0.506E−02
	0.630	0.772	0.904	1.014	0.860	
$\varepsilon = 2^{-4}$	0.881E−01	0.745E−01	0.369E−01	0.185E−01	0.975E−02	0.529E−02
	0.243	1.012	0.996	0.925	0.881	
$\varepsilon = 2^{-6}$	0.127E+00	0.840E−01	0.699E−01	0.317E−01	0.136E−01	0.634E−02
	0.594	0.266	1.139	1.220	1.104	
$\varepsilon = 2^{-8}$	0.127E+00	0.819E−01	0.666E−01	0.347E−01	0.163E−01	0.758E−02
	0.632	0.298	0.941	1.093	1.101	
$\varepsilon = 2^{-10}$	0.127E+00	0.808E−01	0.662E−01	0.346E−01	0.162E−01	0.757E−02
	0.651	0.287	0.938	1.092	1.100	

(continued)

Table 4 (continued)

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\varepsilon = 2^{-12}$	0.127E+00	0.802E−01	0.660E−01	0.345E−01	0.162E−01	0.756E−02
	0.661	0.281	0.936	1.091	1.100	
$\varepsilon = 2^{-14}$	0.127E+00	0.800E−01	0.659E−01	0.345E−01	0.162E−01	0.756E−02
	0.665	0.278	0.935	1.090	1.099	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\varepsilon = 2^{-30}$	0.127E+00	0.797E−01	0.658E−01	0.345E−01	0.162E−01	0.756E−02
	0.670	0.275	0.934	1.090	1.099	
$F^{N,M}$	0.141E+00	0.953E−01	0.699E−01	0.348E−01	0.163E−01	0.761E−02
$Q^{N,M}$	0.565	0.448	1.007	1.095	1.096	

Table 5 Finite difference scheme (5) coupled with a piecewise-uniform mesh in time (9): computed two-mesh differences $F_{\varepsilon}^{N,M}$ and their corresponding orders of convergence $Q_{\varepsilon}^{N,M}$ for problem (8) with $\varepsilon = 2^{-16}$

	N=M=32	N=M=64	N=M=128	N=M=256	N=M=512	N=M=1024
$\theta = 2^2$	0.113E+00	0.829E−01	0.683E−01	0.355E−01	0.159E−01	0.685E−02
	0.449	0.280	0.946	1.158	1.214	
$\theta = 2^4$	0.124E+00	0.805E−01	0.687E−01	0.432E−01	0.250E−01	0.140E−01
	0.628	0.227	0.671	0.789	0.837	
$\theta = 2^6$	0.127E+00	0.794E−01	0.678E−01	0.425E−01	0.245E−01	0.137E−01
	0.679	0.228	0.675	0.793	0.840	
$\theta = 2^8$	0.128E+00	0.791E−01	0.676E−01	0.423E−01	0.244E−01	0.136E−01
	0.693	0.227	0.675	0.794	0.840	
$\theta = 2^{10}$	0.128E+00	0.790E−01	0.675E−01	0.423E−01	0.244E−01	0.136E−041
	0.697	0.227	0.675	0.794	0.841	

References

1. de Falco, C., O'Riordan, E.: Interior layers in a reaction–diffusion equation with a discontinuous diffusion coefficient. Int. J. Numer. Anal. Model. **7**(3), 444–461 (2010)
2. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman and Hall/CRC Press, Boca Raton (2000)
3. Gracia, J.L., O'Riordan, E.: A singularly perturbed convection-diffusion problem with a moving pulse. J. Comput. Appl. Math. **321**, 371–388 (2017)
4. Ladyzhenskaya, O.A., Solonnikov, V.A., Ural'tseva, N.N.: In: Linear and Quasilinear Equations of Parabolic Type. Transactions of Mathematical Monographs, vol. 23. American Mathematical Society, Providence (1968)
5. Miller, J.J.H., O'Riordan, E., Shishkin, G.I., Shishkina, L.P.: Fitted mesh methods for problems with parabolic boundary layers. In: Mathematical Proceedings of the Royal Irish Academy, vol. 98A, pp. 173–190 (1998)

Numerical Results for Singularly Perturbed Convection-Diffusion Problems on an Annulus

Alan F. Hegarty and Eugene O'Riordan

Abstract Numerical methods for singularly perturbed convection-diffusion problems posed on annular domains are constructed and their performance is examined for a range of small values of the singular perturbation parameter. A standard polar coordinate transformation leads to a transformed elliptic operator containing no mixed second order derivative and the transformed problem is then posed on a rectangular domain. In the radial direction, a piecewise-uniform Shishkin mesh is used. This mesh captures any boundary layer appearing near the outflow boundary. The performance of such a method is examined in the presence or absence of compatibility constraints at characteristic points, which are associated with the reduced problem.

1 Introduction

Our interest is in the design of parameter-uniform globally pointwise accurate numerical methods [3] for a wide class of singularly perturbed problems of the form

$$-\varepsilon\triangle u + \mathbf{a}\nabla u + bu = f(x,y),\ b(x,y) \geq 0, \quad (x,y) \in \ \Omega; \quad u = f_B,\ (x,y) \in \partial\Omega. \tag{1}$$

The work of the first author "A.F. Hegarty" was supported by MACSI, the Mathematics Applications Consortium for Science and Industry (www.macsi.ul.ie), funded by the Science Foundation Ireland Investigator Award 12/IA/1683.

A.F. Hegarty
University of Limerick, Limerick, Ireland
e-mail: alan.hegarty@ul.ie

E. O'Riordan (✉)
Dublin City University, Dublin, Ireland
e-mail: eugene.oriordan@dcu.ie

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_8

Moreover, we are only interested in methods that generate numerical approximations which are guaranteed to be free of spurious oscillations. That is, our interest is in numerical methods where the associated system matrix is a monotone matrix. In this paper, the focus will be on singularly perturbed problems posed on non-rectangular domains Ω.

In general, if the elliptic problem (1) on a non-rectangular domain is transformed to a problem posed on a rectangular domain (which acts as the computational domain), a mixed derivative will appear in the transformed differential equation. Monotone discretizations of a mixed second order derivative term normally impose a constraint on the mesh aspect ratio of the form $h_x = Ch_y$, $C = O(1)$ [2, 9], where h_x and h_y are the local mesh steps in the two orthogonal directions. However, as we are interested in global pointwise accuracy for any approximate numerical solutions of (1), we wish to use anisotropic layer-adapted meshes, which are fine along the normal direction to the outflow boundary and coarse in the direction parallel to the boundary. In order that we are free to use layer-adapted meshes, where at certain points in the domain $h_x/h_y = O(1/\varepsilon)$, and also retain a monotone numerical method, we consider specific domains for which a transformation (to a rectangular domain) exists which does not generate a mixed second order derivative term.

In [5] the following problem (posed on the unit disk $\tilde{\Omega}_C := \{(x,y)|x^2+y^2<1\}$)

$$-\varepsilon\triangle\tilde{u}+\tilde{a}(x,y)\tilde{u}_y=\tilde{f},\ \text{in}\ \tilde{\Omega}_C,\quad \tilde{u}=0,\ \text{on}\ \partial\tilde{\Omega}_C;\qquad \tilde{a}>\alpha>0; \tag{2}$$

was examined. The reduced solution $\tilde{r}$ of (2) is defined as the solution of the problem

$$\tilde{a}\tilde{r}_y=\tilde{f}(x,y),\quad (x,y)\in\{(x,y)|x^2+y^2\le 1\}\setminus\tilde{\Gamma}_I^C,$$
$$\tilde{r}(x,y)=0,\quad (x,y)\in\tilde{\Gamma}_I^C:=\{(x,y)|-1\le x\le 1,\ y=-\sqrt{1-x^2}\}.$$

A boundary layer will form in the vicinity of the outflow boundary $\partial\tilde{\Omega}_C\setminus\tilde{\Gamma}_I^C$ and there will be no layer near the inflow boundary $\tilde{\Gamma}_I^C$. Restrictions need to be placed on the admissible $\tilde{f}$ in the vicinity of the characteristic points $(\pm1,0)$ to avoid additional singularities appearing in the reduced solution. In [8] compatibility conditions of level m

$$\frac{\partial^{i+j}\tilde{f}}{\partial x^i\partial y^j}(\pm1,0)=0,\quad 0\le 2i+j\le m,\qquad m\ge 0, \tag{3}$$

are identified, which prevent singularities appearing in some partial derivatives of the reduced solution, in the neighbourhoods of the characteristic points $(\pm1,0)$. In order to prove a parameter-uniform error bound in [5], it was required that $\tilde{f}$ was identically zero in a δ-neighbourhood (where $\delta>0$ is independent of ε) of the points $(\pm1,0)$. This is a more stringent constraint on the data than that given in (3). However, in [6], numerical results suggested that not even the weaker compatibility conditions (3) of level zero are required (in practice) to observe parameter-uniform

convergence for a stable numerical method on an appropriate Shishkin mesh [3, 10], in the case of the particular problem (2). In this paper, the methodology adopted in [5, 6] is extended to a convection-diffusion problem posed on an annulus and we also investigate numerically whether the regularity of the data at the *characteristic* boundary points affects the accuracy of the numerical approximations.

2 Continuous Problem

Consider the problem posed on $\tilde{\Omega}_A := \{(x,y)|R_1^2 < x^2+y^2 < R_2^2\}$: Find $\tilde{u}(x,y)$ s.t.

$$-\varepsilon\triangle\tilde{u} + \tilde{u}_x = 0,\ (x,y) \in \tilde{\Omega}_A;\ \tilde{u} = 0,\ \text{if } x^2+y^2 = R_2^2;\ \tilde{u} = \tilde{g},\ \text{if } x^2+y^2 = R_1^2. \tag{4}$$

For this problem, the reduced solution r solves the first order problem

$$\tilde{r}_x = 0, \quad (x,y) \in \{(x,y)|R_1^2 \le x^2+y^2 \le R_2^2\} \setminus \tilde{\Gamma}_I^A, \quad \tilde{r}(x,y) = 0,\ (x,y) \in \tilde{\Gamma}_I^A; \tag{5}$$

where the inflow boundary is given by

$$\tilde{\Gamma}_I^A := \{(x,y)|y \in [-R_2, R_2],\ x = -\sqrt{R_2^2-y^2}\} \cup \{(x,y)|y \in (-R_1, R_1),\ x = \sqrt{R_1^2-y^2}\}.$$

Unlike problem (2), the inflow boundary $\tilde{\Gamma}_I^A$ is now a disconnected set. The reduced solution of problem (5) is identically zero except downstream, where

$$\tilde{r}(x,y) = g(\sqrt{R_1^2-y^2}, y),\ \forall y \in (-R_1, R_1) \quad (\text{and } x \in \left[\sqrt{R_1^2-y^2}, \sqrt{R_2^2-y^2}\right]).$$

Assuming g is smooth, additional compatibility constraints need to be imposed on the boundary function $g(x,y)$ at the characteristic points $(0, \pm R_1)$, if the partial derivatives of order one are to exist across the two line segments $\{(x, \pm R_1),\ 0 \le x \le \sqrt{R_2^2-R_1^2}\}$.

As in [5, 6] polar coordinates ($x = r\cos(\theta), y = r\sin(\theta)$) are a natural coordinate system associated with both problem (2) and problem (4). Using this map, the annular domain $\tilde{\Omega}_A$ is mapped to a rectangular domain $\Omega_P := \{(r,\theta)|0 < \theta < 2\pi, R_1 < r < R_2\}$. The continuous problem (4) transforms into: Find $u(r,\theta)$ such that

$$\begin{aligned} -\varepsilon r^{-2}u_{\theta,\theta} - \varepsilon u_{rr} + a_1 u_r + a_2 u_\theta &= f, \quad (r,\theta) \in \Omega_P, \\ u(R_1,\theta) = g(\theta), \quad u(R_2,\theta) &= 0, \quad 0 \le \theta \le 2\pi, \\ \text{and} \quad u(r,2\pi) = u(r,0), \quad u_\theta(r,2\pi) &= u_\theta(r,0), \quad R_1 < r < R_2; \end{aligned}$$

where the convective coefficients are

$$a_1(r,\theta) := \cos(\theta) - \varepsilon r^{-1} \quad \text{and} \quad a_2(r,\theta) := -\sin(\theta) r^{-1}.$$

3 Discrete Problem and Numerical Results

The absence of a mixed second order derivative term in the transformed problem permits the use of simple upwinding to approximate the first order derivative terms, so that the discrete operator is then inverse monotone. To capture the layers at the outflow boundary, we use one of two possible piecewise uniform Shishkin meshes (which we denote, respectively, by $\bar{\omega}_i,\ i = 1,2$), in the radial direction and a uniform mesh $\bar{\omega}_T := \{\theta_j = iK,\ j = 0,1,\ldots,N,\ K = \frac{2\pi}{N}\}$ in the angular direction. If we just refine around the inner boundary where $r = R_1$ then the radial mesh $\bar{\omega}_1$ is defined by subdividing the interval $[R_1,R_2] = [R_1,R_1+\sigma_1]\cup[R_1+\sigma_1,R_2]$, and then subsequently subdividing each of these two subintervals into $N/2$ mesh elements. On the other hand, if we refine the radial mesh near both boundaries $r = R_1$ and $r = R_2$ then the radial mesh ω_2^N is defined by dividing the interval into three subintervals via $[R_1,R_2] = [R_1,R_1+\sigma_2]\cup[R_1+\sigma_2,R_2-\sigma_2]\cup[R_2-\sigma_2,R_2]$ and then subsequently subdividing each of these subintervals in the ratio $N/4 : N/2 : N/4$ mesh elements. The particular choice for the transition parameter $\sigma_i, i = 1,2$ is discussed below. Hence we define two potential meshes $\bar{\Omega}_i^N := \bar{\omega}_i \times \bar{\omega}_T,\ i = 1,2$. The discrete problem(s) are then of the form[1]: Find $U_p^N,\ p = 1,2$ such that

$$-\frac{\varepsilon}{r_i^2}\delta_\theta^2 U_p^N - \varepsilon\delta_r^2 U_p + a_1 D_r^{\pm} U_p^N + a_2 D_\theta^{\pm} U_p^N = f,\ r_i \in \omega_p, \theta_j \in \omega_T;$$

$$U_p^N(r_1,\theta) = f(\theta), \quad U_p^N(r_2,\theta_j) = 0, \quad 0 \le \theta_j \le 2\pi,$$

$$U_p^N(r_i,2\pi) = U_p^N(r_i,0), \quad D_\theta^- U_p^N(r_i,2\pi) = D_\theta^+ U_p^N(r_i,0), \quad R_1 < r_i < R_2.$$

We examine the performance of this numerical method applied to two particular problems, which satisfy either some level of compatibility or none at all. In order to prove a parameter-uniform error bound, the choice of the transition parameter in [5] depended on stringent compatibility being imposed on the data. For the problems considered in this current paper, we simply override these theoretical constraints and set

$$\sigma_i := \min\{\frac{R_2 - R_1}{2i}, 2\varepsilon \ln N\}, \quad i = 1,2.$$

[1] $aD_r^{\pm} := 0.5\Big((|a|+a)D_r^- + (|a|-a)D_r^+\Big)$, $D_r^+ U(r_i,\theta_j) := D_r^- U(r_{i+1},\theta_j)$; where $D_r^- U(r_i,\theta_j) := \frac{U(r_i,\theta_j)-U(r_{i-1},\theta_j)}{(r_i - r_{i-1})}$, $\delta_r^2 U(r_i,\theta_j) := 2\frac{(D_r^+ - D_r^-)U(r_i,\theta_j)}{(r_{i+1}-r_{i-1})}$.

Consider the following two particular examples of problem (4) where

$$R_1 = 1,\ R_2 = 4, \qquad g \equiv x^6, \quad \text{on} \quad x^2 + y^2 = 1; \tag{6}$$

$$\text{and} \qquad R_1 = 1,\ R_2 = 4, \qquad g \equiv 1, \quad \text{on} \quad x^2 + y^2 = 1. \tag{7}$$

For both of these problems we employ the Shishkin mesh ω_2, which refines around both boundaries $r = R_1$ and $r = R_2$. Sample computed solutions for both problems are displayed in Figs. 1 and 2. The reduced solution for the first test problem (4, 6) is continuous and it is discontinuous in the second test problem (4, 7). To investigate the convergence of the scheme, we estimate the order of convergence p_ε^N for each particular choice of ε and the uniform order of convergence p^N over a range of values for $\varepsilon \in [2^{-20}, 1]$, using the two mesh differences [3, pg. 166]. The maximum two–mesh differences D_ε^N and the parameter–uniform maximum two–mesh differences D^N, are computed from

$$D_\varepsilon^N := ||U^N - \overline{U}^{2N}||_{\Omega^N,\infty}, \qquad D^N := \max_{\varepsilon \in \{2^{-j}\}_0^{20}} D_\varepsilon^N.$$

Approximations p_ε^N to the local order of convergence and approximations p^N to the parameter–uniform order of local convergence are subsequently computed from

$$p_\varepsilon^N := \log_2 \frac{D_\varepsilon^N}{D_\varepsilon^{2N}} \qquad p^N := \log_2 \frac{D^N}{D^{2N}}.$$

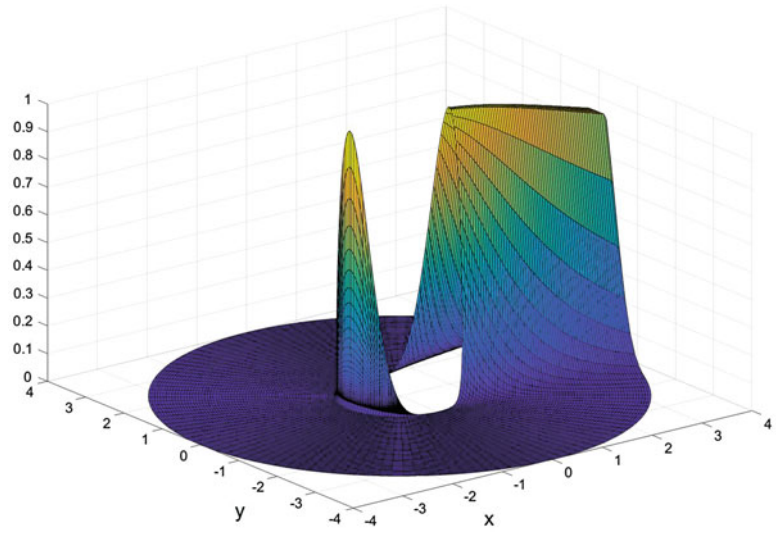

Fig. 1 Numerical solution $\bar{U}_2^{128}$ of problem (4, 6) with $\varepsilon = 2^{-20}$

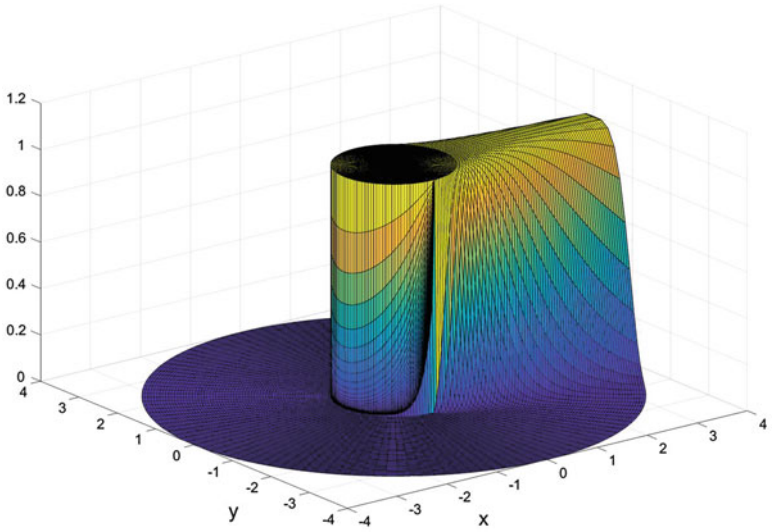

Fig. 2 Numerical solution $\bar{U}_2^{128}$ of problem (4, 7) with $\varepsilon = 2^{-20}$

Table 1 Computed orders of nodal convergence p_ε^N and ε-uniform order p^N for problem (4, 6)

ε	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	0.65	1.37	1.51	1.14	1.06	1.03	1.01
2^{-2}	0.63	1.47	1.37	1.09	1.04	1.02	1.01
2^{-4}	0.72	1.24	0.91	0.93	0.91	0.89	0.98
2^{-6}	0.78	0.63	0.61	0.97	1.03	0.99	0.96
2^{-8}	0.76	0.39	0.40	0.97	1.24	0.85	0.84
2^{-10}	0.76	0.31	0.31	0.95	1.31	0.69	0.79
2^{-12}	0.76	0.29	0.29	0.93	1.31	0.67	0.76
2^{-14}	0.76	0.29	0.28	0.93	1.31	0.66	0.76
2^{-16}	0.76	0.28	0.28	0.93	1.31	0.66	0.75
2^{-18}	0.76	0.28	0.28	0.93	1.31	0.66	0.75
2^{-20}	0.76	0.28	0.28	0.93	1.31	0.66	0.75
p^N	0.76	0.28	0.28	0.93	1.31	0.66	0.75

In Table 1, we observe the orders of convergence have stabilized for sufficiently small values of ε and we observe parameter uniform convergence for the compatible problem (6). In Table 2, we observe a lack of convergence when the method is applied to the incompatible problem (7). In [6], we also examined numerically an incompatible problem posed on the interior of the unit disc and observed minimal adverse effect on the performance of the numerical method. However, in the case of the problem (4, 7) we observe in Fig. 3 (and not in Fig. 4) a spike in the

Table 2 Computed orders of nodal convergence p_ε^N for problem (4, 7)

ε	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	1.27	1.26	1.15	0.97	0.98	0.99	1.00
2^{-2}	0.81	1.05	1.08	1.05	1.02	1.01	1.01
2^{-4}	0.33	0.75	1.00	0.95	0.82	0.80	0.98
2^{-6}	0.10	0.58	0.98	0.50	0.63	0.77	0.87
2^{-8}	0.04	0.50	0.71	0.29	0.34	0.47	0.62
2^{-10}	0.03	0.47	0.55	0.25	0.19	0.22	0.32
2^{-12}	0.02	0.47	0.34	0.17	0.32	0.18	0.14
2^{-14}	0.02	0.47	0.29	0.07	0.09	0.27	0.30
2^{-16}	0.02	0.47	0.28	0.05	0.02	0.09	0.19
2^{-18}	0.02	0.47	0.28	0.04	0.00	0.04	0.06
2^{-20}	0.02	0.47	0.28	0.04	−0.00	0.03	0.03

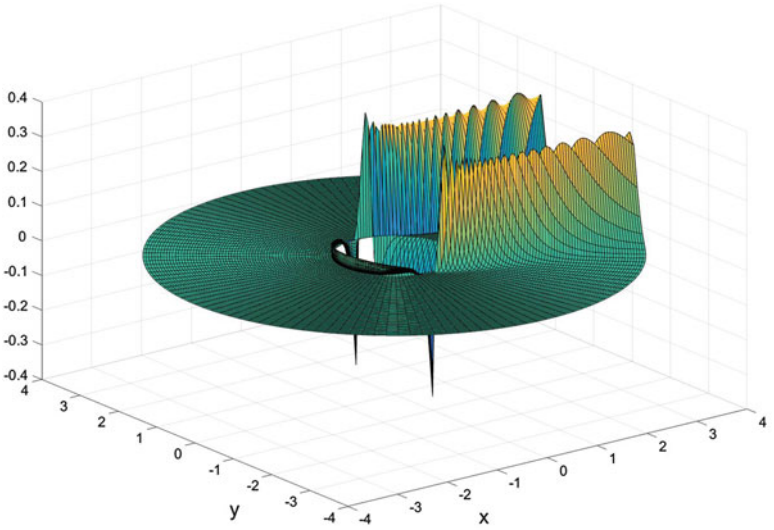

Fig. 3 Fine mesh comparison $\bar{U}_2^{128} - \bar{U}_2^{1024}$ for problem (4, 7) with $\varepsilon = 2^{-20}$

error caused by the lack of compatibility.In a future publication, the authors will investigate parameter-uniform convergence of the above numerical method applied to the annulus problem (4), under the additional assumption that there exists some $\delta > 0$ such that $g(x, y) \equiv 0$ when $|R_1 - y| \leq \delta$.

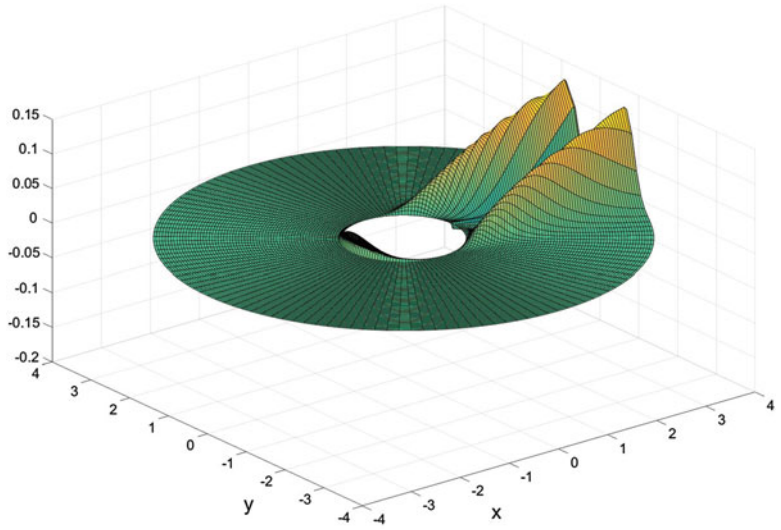

Fig. 4 Fine mesh comparison $\bar{U}_2^{128} - \bar{U}_2^{1024}$ for problem (4, 6) with $\varepsilon = 2^{-20}$

4 The Hemker Problem

The test problem (4, 7) is linked to the singularly perturbed problem

$$-\varepsilon \Delta u + u_x = 0, \quad \Omega := \{(x,y) | x^2 + y^2 > 1\}, \tag{8}$$

$$u(x,y) = 1 \quad \text{for} \quad x^2 + y^2 = 1, \quad u(x,y) = 0 \quad \text{for} \quad x^2 + y^2 \to \infty; \tag{9}$$

which was first proposed by Hemker [7] as a benchmark problem. This problem is a singularly perturbed elliptic problem (of convection-diffusion type) posed on an unbounded domain exterior to the unit disc. See [1, 4] (and the references therein) for some computational approaches to solving this benchmark problem. If we wish to generate accurate approximate solutions to this problem, then we can take the computational domain as an annulus where $R_2 \to \infty$. However, our results on such a domain for problem (4, 7) have not been impressive. So we now consider using a different computational domain for $x > 0$. Moreover, we formulate an alternative problem to (8, 9), which we believe retains some of the key difficulties within the Hemker problem. Consider the singularly perturbed problem: Find u such that

$$-\varepsilon \Delta u + u_x = 0, \quad (x,y) \in D \tag{10}$$

$$u(x,y) = 1, \quad \text{for} \quad x^2 + y^2 = 1, \quad u_x(L,y) = 0, \quad \text{for} \quad -R_2 < y < R_2, \tag{11}$$

$$u(x,y) = 0, \ (x,y) \in (\bar{D} \setminus D) \setminus (\{(x,y) | x = L, -R_2 < y < R_2\}); \tag{12}$$

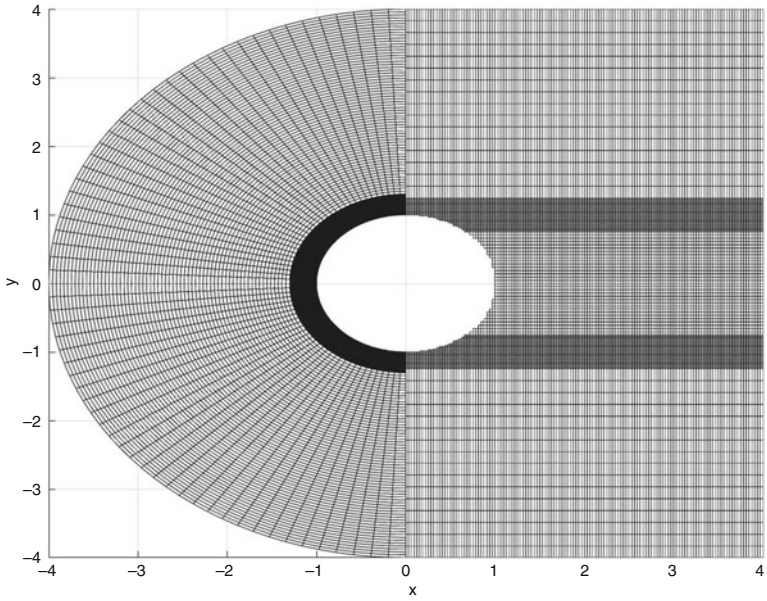

Fig. 5 The overlay mesh $\Omega_1^N \cap \tilde{\Omega}_3^N$ with $\varepsilon = 2^{-5},\ N = 128,\ L = 4, R_2 = 4$

posed on the bounded composite domain $D := (C_2 \cup Q) \setminus \bar{C}_1$ (see Fig. 5), where

$$C_i := \{(x,y) | x^2 + y^2 < R_i^2\},\ i = 1,2, R_1 := 1;\ \ Q := (0,L) \times (-R_2, R_2),\ L > R_2.$$

We compute a numerical approximation to the solution of this problem in two stages. We first compute an approximate solution of problem (4, 7) using the mesh $\tilde{\Omega}_1^N$, which only refines near the inner boundary $R_1 = 1$. The computed solution will (in polar coordinates) be denoted by $Z_A^N(r_i, \theta_j)$. In the second phase, we solve the following discrete problem over the rectangle Q, using a mesh $\tilde{\Omega}_3^N$: Find $\tilde{Z}_Q^N$ s.t.

$$\tilde{Z}_Q^N(x_i, y_j) \equiv 1, \quad (x_i, y_j) \in \Omega_3^N \cap \bar{C}_1;$$
$$\left(-\varepsilon \delta_x^2 - \varepsilon \delta_y^2 + D_x^-\right) \tilde{Z}_Q^N(x_i, y_j) = 0, \quad (x_i, y_j) \in \Omega_3^N \setminus \bar{C}_1;$$

with the remaining boundary values computed from the equations

$$D_x^- \tilde{Z}_Q^N(L, y_j) = 0,\ -R_2 < y_j < R_2, \quad \tilde{Z}_Q^N(x_i, -R_2) = \tilde{Z}_Q^N(x_i, R_2) = 0,\ x_i \in [0, L];$$
$$\tilde{Z}_Q^N(0, y_j) = \bar{Z}_A^N(0, y_j), \quad y_j \in (-R_2, -1) \cup (1, R_2).$$

Here $\bar{Z}_A^N(0, y_j)$ is a linear interpolant of the values $Z_A^N(r_i, \pi/2)$ and $Z_A^N(r_i, 3\pi/2)$ along the line $x = 0$. Also the mesh $\tilde{\Omega}_3^N := \omega_u \times \omega_3$ is a tensor product mesh of a uniform mesh $\omega_u := \{x_i | x_i = iL/N, 0 < i < N\}$ in the horizontal direction and

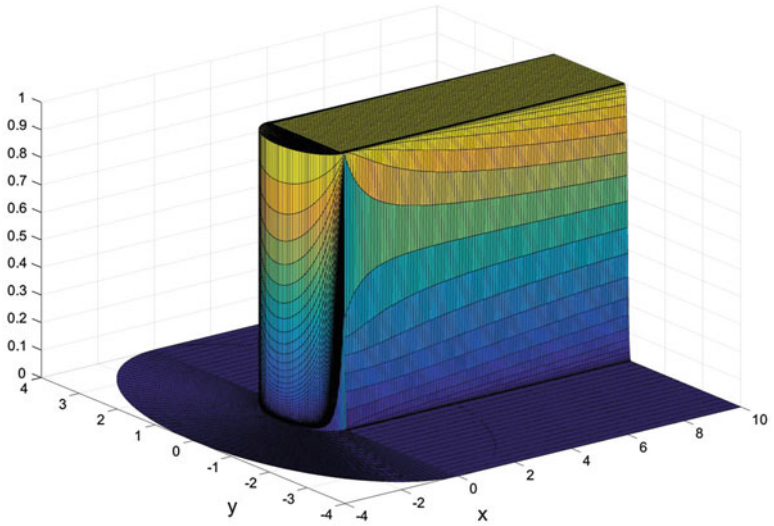

Fig. 6 Computed solution of problem (10, 11, 12) on the mesh $\Omega_1^N \cap \tilde{\Omega}_3^N$ with $\varepsilon = 2^{-15}$, $N = 128$, $R_2 = 4$, $L = 10$

the vertical mesh ω_3 is a Shishkin mesh which refines in the region of the interior characteristic layers, emanating from the characteristic points $(0, \pm 1)$. This mesh ω_3 is generated by splitting the vertical interval $[-R_2, R_2]$ into the five subregions

$$[-R_2, -1-\tau_2] \cup [-1-\tau_2, -1+\tau_1] \cup [-1+\tau_1, 1-\tau_1] \cup [1-\tau_1, 1+\tau_2] \cup [1+\tau_2, R_2]$$

and distributing the mesh elements in the ratio $N/8 : N/4 : N/4 : N/4 : N/8$. The Shishkin transition parameters[2] are taken to be

$$\tau_1 := \min\{\frac{1}{2}, \sqrt{\varepsilon} \ln N\}; \quad \tau_2 := \min\{\frac{R_2 - 1}{2}, \sqrt{\varepsilon} \ln N\}.$$

The resulting mesh from this construction is illustrated in Fig. 5. The computed approximation U^N to the solution of problem (10, 11, 12) is given as

$$U^N := Z_A^N(r_i, \theta_j),\ (r_i, \theta_j) \in \bar{\Omega}_1^N \setminus (\{x \geq 0\}), \quad \tilde{U}^N := \tilde{Z}_Q^N(x_i, y_j),\ (x_i, y_j) \in \bar{\Omega}_3^N \setminus \bar{C}_1.$$

A sample computed solution U^N is displayed in Fig. 6. A plot of the approximate errors in Fig. 7 indicates that the spike in the error in the vicinity of the characteristic

[2]The choice of $\sqrt{\varepsilon}$ for the scaling in these interior layers is motivated by the following heuristic argument: the reduced solution of (8, 9) is discontinuous along the half-lines $y = \pm R_1, x > 0$. Assuming that the vertical derivatives dominate (in scale) the horizontal derivatives in the neighbourhood of these half-lines, the solution is approximated by the solution of a heat equation of the form $-\varepsilon s_{yy} + s_x = 0$, which suggests the scaling of $\zeta = y/\sqrt{\varepsilon}$ for the mesh.

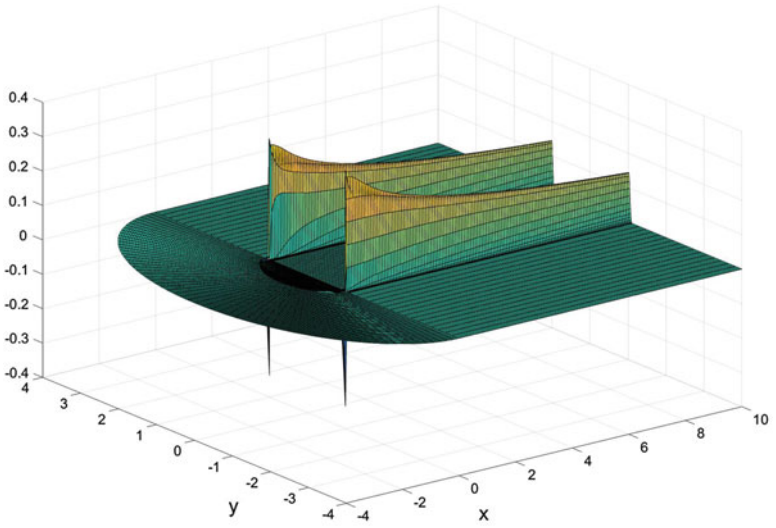

Fig. 7 Fine mesh comparison $\bar{U}^{128} - \bar{U}^{1024}$ for problem (10, 11, 12) with $\varepsilon = 2^{-15}$, $N = 128$, $R_2 = 4$, $L = 10$

points pollutes the approximate solution downwind where $x \geq 0$. However, in Fig. 7 we see that the loss in accuracy appears to be restricted to the parabolic layers in the vicinity of the lines $y = \pm 1$. This leads to the conjecture that, if the numerical method could be corrected near the points $(0, \pm 1)$ then this may correct the approximate solution at all points, where the accuracy is currently lost.

5 Conclusions

Unlike for the problem posed on the interior of the unit disk [6], the absence of compatibility conditions at characteristic points has an adverse effect on the convergence of numerical approximations (generated by the numerical methods in this paper) in the case of the problem posed on the exterior of the unit disc. Nevertheless, as a computational approach to solving the Hemker problem, upwinding on appropriate Shishkin meshes in a patched computational domain appears to yield reasonable (oscillation-free) approximations away from the interior parabolic layer regions near the half-lines $y = \pm 1$, $x \geq 0$, which emanate from the characteristic points $(0, \pm 1)$. The question of whether it is possible (or not) to design a fitted mesh and, perhaps, combine this with a fitted operator to obtain parameter-uniform pointwise accuracy throughout the entire domain for problems like (10, 11, 12) remains open to further investigation.

References

1. Augustin, M., Caiazzo, A., Fiebach, A., Fuhrmann, J., John, V., Linke, A., Umla, R.: An assessment of discretizations for convection-dominated convection-diffusion equations. Comput. Methods Appl. Mech. Eng. **200**(47–48), 3395-3409 (2011)
2. Dunne, R.K., O' Riordan, E., Shishkin, G.I.: Fitted mesh numerical methods for singularly perturbed elliptic problems with mixed derivatives. IMA J. Numer. Anal. **29**, 712–730 (2009)
3. Farrell, P.A., Hegarty, A.F., Miller, J. J. H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman and Hall/CRC Press, Boca Raton (2000)
4. Han, H., Huang, Z., Kellogg, R.B.: A tailored finite point method for a singular perturbation problem on an unbounded domain. J. Sci. Comput. **36**(2), 243-261 (2008)
5. Hegarty, A.F., O'Riordan, E.: Parameter-uniform numerical method for singularly perturbed convection-diffusion problem on a circular domain. Adv. Comput. Math. doi:10.1007/s10444-016-9510-z
6. Hegarty, A.F., O'Riordan, E.: Numerical solution of a singularly perturbed elliptic problem on a circular domain. Model. Anal. Inf. Syst. **23**(3), 349–356 (2016)
7. Hemker, P.W.: A singularly perturbed model problem for numerical computation. J. Comput. Appl. Math. **76**, 277–285 (1996)
8. Jung, C.-Y., Temam, R.: Convection-diffusion equations in a circle: the compatible case. J. Math. Pures Appl. **96**, 88–107 (2011)
9. Matus, P.: The maximum principle and some of its applications. Comput. Methods Appl. Math. **2**(1), 50–91 (2002)
10. Shishkin, G.I.: Discrete approximation of singularly perturbed elliptic and parabolic equations. Russian Academy of Sciences, Ural section, Ekaterinburg (1992, in Russian)

Numerical Calculation of Aerodynamic Noise Generated from an Aircraft in Low Mach Number Flight

Vladimir Jazarević and Boško Rašuo

Abstract The paper describes numerical prediction of aerodynamic noise generated from the aircraft. It focuses on the simulation of turbulent flow around rectified flap on the wing represented in 2D. Simulation of turbulent flow is modeled using the stabilized orthogonal subgrid scale (OSGS) method with dynamical subscales. It is shown how the stabilization method can perform simulation of turbulent flow affecting the prediction of acoustic sources calculated applying Lighthill's analogy. Acoustic sources are used in inhomogeneous Helmholtz equation to simulate pressure wave propagation in the domain closing the circle of three main steps required for simulating aeroacoustics phenomena. It is shown that OSGS with dynamical subscales gives better representation of the spectrum. Overall, better prediction of energy transfer across large and small eddies provides better allocation and presentation of acoustics sources. These sources change wave propagation of the pressure in acoustic field.

1 Introduction

As it is known, the Navier-Stokes partial differential equation describes the behaviour of fluid flow, but there are two problems that mathematicians have not been able to solve to the present day. The first one is uniqueness and smoothness of the solution of the Navier-Stokes equation. The second problem is turbulence, because fluid continually generates features at decreasing scale eddies. One of the most difficult challenges in numerical algorithms of turbulent flow is how to model these small scale eddies and their effect on large scale eddies. Also, how to properly define energy distribution between these small scale and large

V. Jazarević • B. Rašuo (✉)
Faculty of Mechanical Engineering, University of Belgrade, Kraljice Marije 16, Belgrade, Serbia
e-mail: vlada.jazarevic@gmail.com; brasuo@mas.bg.ac.rs

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_9

scale eddies. Turbulent flow around bodies that travel fast through the air makes fluctuations of the pressure that our ear recognizes as sound. It is this kind of physical phenomena that Aeroacoustics [1] deals with. With constant growth of capabilities of personal computers, a new field of computational mechanics has also emerged: Computational Aeroacoustics (CAA) [2]. The objective of this work is to present the stabilized subgrid scale finite element method with variation of orthogonal subgrid scale method with dynamical subscale for the approximation of incompressible Navier-Stokes equation which beside of stabilization also can model turbulent flow and how this modeling of turbulent flow affects calculation of Lighthill's [3] tensor. In this work, 2D simulation is performed although it is well know that in 2D it is impossible to recover natural behavior of turbulent flow. The idea is to show how for the same mesh and same computer resources the proposed methodology shows better results than for the usual approach which uses LES[4] methodology for modeling turbulent flow. In addition to the stabilization of proposed methodology it will be shown how the use of orthogonal projection on finite element space for the approximation of small scales is a good choice to separate energy bounds for two scales. Dynamical scales in the end provide the opportunity to model the backscatter of energy between small and large eddies and possibility for the energy to go in both directions from large to small eddies and vice versa. Finally, the velocity vector as a solution of Navier-Stokes equation is used in calculating of Lighthill's tensor which presents the acoustic sources and these acoustic sources are the right side of inhomogenous Helmholtz equation that describes acoustic pressure wave propagation in the domain. Here, we are interested in small wave numbers ($k \leq 15$) avoiding the problems known as pollution error for large wave numbers.

2 Proposed Methodology to Simulate Aerodynamic Noise

In order to simulate aerodynamic noise around the components of the aircraft as rectified flaps shown in Fig. 1 the simulation will be divided into three main steps. The first one is the simulation of turbulent flow with CFD. The numerical problem consists in solving partial differential equation in the domain $\Omega \subset \mathbb{R}^d$ with the boundary condition $\Gamma = \Gamma_N \cup \Gamma_D = \partial\Omega$ and prescribed initial condition and boundary.

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega, \quad t > 0, \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad \Omega, \quad t > 0, \tag{2}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \quad \text{in} \quad \Omega, \quad t = 0, \tag{3}$$

$$\mathbf{u}(\mathbf{x}, t) = 0 \quad \text{on} \quad \Gamma_D, \quad t > 0, \tag{4}$$

$$\mathbf{n} \cdot \sigma(x, t) = \mathbf{t}_N(x, t) \quad \text{on} \quad \Gamma_N, \quad t > 0. \tag{5}$$

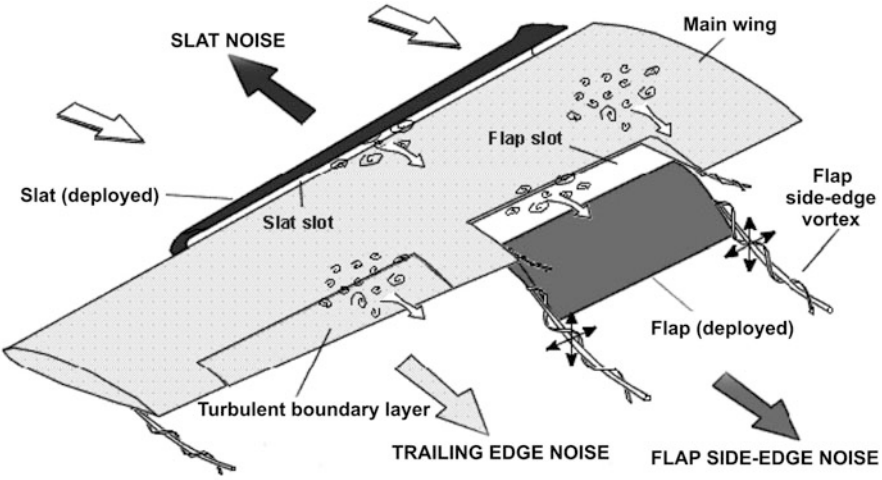

Fig. 1 Sources of sound generation on the wing of the aircraft with stress on deployed flaps

As it is known, the simulation of turbulent flow demands the appropriate design of numerical schemes (LES, Orthogonal SGS with dynamical subscales) in order to catch the physical behaviour of turbulence. The goal of turbulent flow simulation is to obtain the velocity vector $\mathbf{u}$ with exact productions of wakes.

The second part consists of calculating the acoustics sources using the method proposed by Lighthill, which means calculation of Reynolds tensor.

$$\begin{aligned}(\nabla \otimes \nabla) : \mathbf{T} \approx \rho_0(\nabla \otimes \nabla) : (\mathbf{u} \otimes \mathbf{u}) &= \rho_0 \nabla[(\nabla \otimes \mathbf{u}) \cdot \mathbf{u} + \mathbf{u}(\nabla \cdot \mathbf{u})] \\ = \rho \nabla \cdot [(\nabla \otimes \mathbf{u}) \cdot \mathbf{u}] &= \rho \mathbf{u} \cdot \nabla(\nabla \cdot \mathbf{u}) + \rho(\nabla \otimes \mathbf{u}) : (\nabla \otimes \mathbf{u})^{\top} \\ = \rho(\nabla \otimes \mathbf{u}) : (\nabla \otimes \mathbf{u})^{\top} &= s(\mathbf{x}, t) \qquad (6)\end{aligned}$$

In order to keep the advantages of using C^0-class finite elements, Reynolds tensor is expressed in the form $\rho_0(\nabla \otimes \mathbf{u}) : (\nabla \otimes \mathbf{u})^{\top} = s(\mathbf{x}, t)$. After the source term in time domain is obtained, it has to be transformed to the frequency domain, using Direct Fourier Transform (DFT).

The third part consists of solving inhomogeneous Helmholtz equation using the source term obtained in the previous step in order to simulate pressure wave propagation in the domain, which is known as simulation of acoustic part. The same domain is used as in the CFD step. Mathematical problem involves finding the pressure p_H in the domain $\Omega \subset R^d$ with the boundary $\Gamma_N \cup \Gamma_\infty = \partial\Omega$.

$$\triangle p_H(x, \omega) + k_0^2 p_H(x, \omega) = s(x, \omega) \quad \text{in} \quad \Omega, \qquad (7)$$

$$\nabla p_H(x, \omega) \cdot \mathbf{n} = 0 \quad \text{on} \quad \Gamma_N, \qquad (8)$$

$$\nabla p_H(x, \omega) \cdot \mathbf{n} = ik_0 p_H \quad \text{on} \quad \Gamma_\infty. \qquad (9)$$

First, time domain is decomposed in sub intervals and in each time step the Navier-Stokes equation is solved than is calculated Lighthill's tensor for that time step and with direct Fourier transform translated to frequency domain for chosen number of frequencies. After all time steps are calculated, the appropriate acoustic sources calculated in a previous step are chosen. These acoustic sources are the right hand side of the inhomogeneous Helmholtz equation which is then solved.

3 Orthogonal Subgrid Scale Method with Dynamical Subscales

The idea behind SGS method is also to decompose the velocity and velocity test function to resolvable (capture with FEM mesh) or large scales and non-resolvable or small scales. The decomposition of $\mathbf{u} = \mathbf{u}_* = \mathbf{u}_h + \tilde{\mathbf{u}}$, $\mathbf{v} = \mathbf{v}_h + \tilde{\mathbf{v}}$ refers to space splitting $V_0^d = V_{h,0}^d + \tilde{V}_h^d$. The velocity time derivation can be split as $\partial_t \mathbf{u} = \partial_t \mathbf{u}_h + \partial_t \tilde{\mathbf{u}}$ where the second term is saved because it is chosen to deal with dynamical subscales [5]. Enforcing the subscales to be L^2 orthogonal to the finite element or, in other words, $\tilde{V}_0^d$ is taken as a subspace of this solution leads to the separate energy bounds for the two different scales. The separation of the scales is only proper if they are orthogonal in the sense that the total kinetic energy is the sum of the kinetic energy of $\mathbf{u}_h$ plus the kinetic energy of small scales. Also, the pressure and pressure test function are decomposed as $p = p_h + \tilde{p}$, $q = q_h + \tilde{q}$ corresponding to the space splitting $Q_0 = Q_{h,0} + \tilde{Q}_0$. In the formulation the pressure subscales are not used. Let us consider finite element partition K of the computational domain Ω. Applying the ideas in Eqs. (1)–(5) it is formulated:

$$\begin{aligned}
&(\partial_t \mathbf{u}_h, \mathbf{v}_h) + \langle \mathbf{u}_* \cdot \nabla \mathbf{u}_h, \mathbf{v}_h \rangle + \nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h) \\
&\quad + (\partial_t \tilde{\mathbf{u}}, \mathbf{v}_h) - \sum_K \langle \tilde{\mathbf{u}}, \mathbf{u}_* \cdot \nabla \mathbf{v}_h + \nu \triangle \mathbf{v}_h + \nabla q_h \rangle_K \\
&\quad + \sum_K \langle \tilde{\mathbf{u}}, \nu \mathbf{n} \cdot \nabla \mathbf{v}_h + q_h \mathbf{n} \rangle_{\partial K} = \langle \mathbf{f}, \mathbf{v}_h \rangle
\end{aligned} \tag{10}$$

$$\begin{aligned}
&(\partial_t \tilde{\mathbf{u}}, \tilde{\mathbf{v}}) + \sum_K \langle \mathbf{u}_* \cdot \nabla \tilde{\mathbf{u}} - \nu \triangle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_K + \sum_K \langle \nu \mathbf{n} \cdot \nabla \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_{\partial_K} \\
&\quad + \sum_K \langle \partial_t \mathbf{u}_h + \mathbf{u}_* \cdot \nabla \mathbf{u}_h - \nu \triangle \mathbf{u}_h + \nabla p_h, \tilde{\mathbf{v}} \rangle_K \\
&\quad + \sum_K \langle \nu \mathbf{n} \cdot \nabla \mathbf{u}_h - p_h \mathbf{n}, \tilde{\mathbf{v}} \rangle_{\partial K} = \langle \mathbf{f}, \tilde{\mathbf{v}} \rangle
\end{aligned} \tag{11}$$

These discrete variational must hold for all test functions $[\mathbf{v}_h, q_h] \in V_h \times Q_h$ and $\tilde{\mathbf{v}} \in \tilde{V}$, where $\tilde{V}$ is the space of subscales to be defined. It is observed that some terms have been integrated by parts within each elements. Equation (10) defines large scales and Eq. (11) defines small scales. Apart from taking zero the pressure subscale, no approximations have been done to arrive to Eqs. (10)–(11). Different approximations will lead to different formulations within the same framework. It would be consider that $\partial_t \approx 0$ and takes $\mathbf{u}_* \approx \mathbf{u}_h$ as advection velocity. Firstly, let describe the space of subscales $\tilde{V}$, that is the space where $\tilde{u}$ belongs for t fixed. A particular feature of our approach is to take it L^2 orthogonal to the finite element space or it could be said that $\tilde{V}$ is subspace of $V_h^{\perp}$. The first approximation will be is to take $\tilde{u} \approx 0$ on ∂K for each element in domain K of the finite element partition. That can be understood as approximating the velocity subscale by a space of bubble function. However, the heuristic Fourier arguments proposed in [6] also allows us to explain why the effect of the subscales on the element boundaries can be neglected compared to the effect in the element interiors. Nevertheless, this approximation can be relaxed following the ideas suggested in [7].

$$\sum_K \langle \mathbf{u}_* \cdot \nabla \tilde{\mathbf{u}} - \nu \triangle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_K \approx \sum_K \tau_K^{-1} \langle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_K \tag{12}$$

where τ_K is algebraic parameter calculated for every element of computational domain

$$\tau_K^{-1} = \frac{c_1 \nu}{h_k^2} + \frac{c_2 \|\mathbf{u}_*\|_{L^\infty(K)}}{h_K} \tag{13}$$

where: $c_1 = 4$ and $c_2 = 2$—for linear triangular elements, h_K—characteristic mesh element size.

The next goal is to find the solution of small scales in Eq. (11) as a function of large scales and then to put back that result in Eq. (10). Equation (11) could be written in a differential form:

$$\delta_t \tilde{\mathbf{u}}^n + (\mathbf{u}_h^{n+\alpha} + \tilde{\mathbf{u}}^{n+\alpha}) \cdot \nabla \tilde{\mathbf{u}}^{n+\alpha} - \nu \triangle \tilde{\mathbf{u}}^{n+\alpha} + \nabla \tilde{p}^{n+1} = \mathbf{r}_{u,h}^{n+\alpha} \tag{14}$$

$$\nabla \cdot \tilde{\mathbf{u}}^{n+\alpha} = \mathbf{r}_{p,h}^{n+\alpha} \tag{15}$$

where $\mathbf{r}_{u,h}^{n+\alpha}$ and $\mathbf{r}_{p,h}^{n+\alpha}$ represent residuals obtained with FEM method $\mathbf{u}_h$ and p_h given as,

$$\mathbf{r}_{u,h}^{n+\alpha} = -P[\delta_t \mathbf{u}^n + (\mathbf{u}_h^{n+\alpha} + \mathbf{u}^{n+\alpha}) \cdot \nabla \mathbf{u}^{n+\alpha} - \nu \triangle \mathbf{u}^{n+\alpha} + \nabla \tilde{p}^{n+1} - \mathbf{f}] \tag{16}$$

$$\mathbf{r}_{p,h}^{n+\alpha} = -P[\nabla \cdot \mathbf{u}^{n+\alpha}] \tag{17}$$

where $P = I - \Pi_h$, and Π_h is L^2 projection on finite element space which leads to the approach known as the Orthogonal subgrid scale stabilization [6]. The formulation of orthogonal subgrid scale with dynamical subscales can be formulated as

$$(\partial_t \mathbf{u}_h, \mathbf{v}_h) + \langle \mathbf{u}_* \cdot \nabla \mathbf{u}_h, \mathbf{v}_h \rangle + \nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h) - \sum_K \langle \tilde{\mathbf{u}}, \mathbf{u}_* \cdot \nabla \mathbf{v}_h + \nu \triangle \mathbf{v}_h + \nabla q_h \rangle_K = \langle \mathbf{f}, \mathbf{v}_h \rangle \tag{18}$$

$$(\partial_t \tilde{\mathbf{u}}, \tilde{\mathbf{v}}) + \sum_K \tau_K^{-1} \langle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_K + \sum_K \langle \mathbf{u}_* \cdot \nabla \mathbf{u}_h - \nu \triangle \mathbf{u}_h + \nabla p_h, \tilde{\mathbf{v}} \rangle_K = \langle \mathbf{f}, \tilde{\mathbf{v}} \rangle \tag{19}$$

where the second equation is calculated first and then the approximation of small scales is incorporate in a first equation. The model such as it is, suitable for implementation in computer code.

4 Model and Simulation Setup

The domain produced around airfoil 23012 is of dimensions of 300 m × 300 m. The chord dimension of airfoil is 1m and the flap is deployed at 20°. The mesh of linear triangular elements for the hall domain is produced where the density of the mesh is finer around the airfoil, and becoming thicker with departure from the airfoil. The number of triangular elements is 128,307 and there is 64,680 computational nodes. Figure 2 shows the nodes of the mesh where time tracking of velocity components and pressure is done. From the literature is known that the maximum aerodynamic noise is generated in rift between the airfoil body and the flap, and on the trailing edge of the flap. Because of that the position of tracking nodes looks as shown

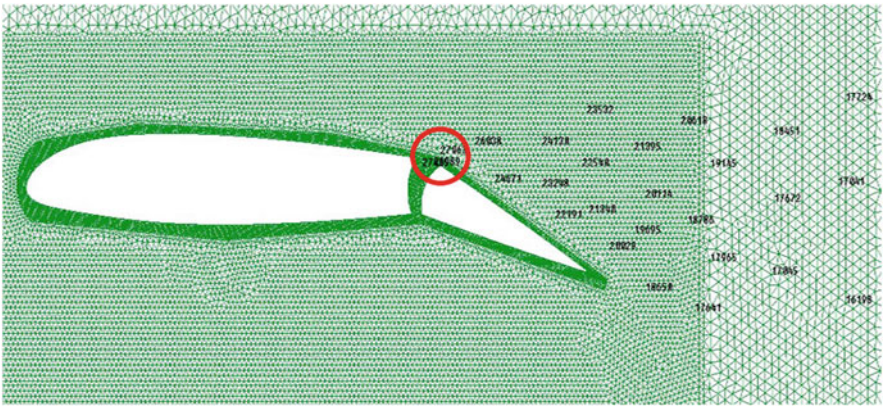

Fig. 2 Position of time tracking nodes for velocity components and pressure

in Fig. 2. In the end it is prescribed the Dirichlet boundary condition of $v = 0$ on the boundary of the airfoil and flap. Also, it is prescribed $v = 100\,\text{m/s}$ on the left side of the domain. The angle of attack of the airfoil is $\alpha_n = 10°$. On the rest of the boundaries of the domain Neumann boundary condition is prescribed to avoid problems with numerical instabilities which can produce wakes generated from the trailing edge. The LES-Smagorinsky method and Orthogonal stabilized subgrid scale method with dynamical subscale are used for simulation of turbulence. The hall time interval is 0.35 s and the time step is 0.0005 s, where second order Adams-Bashforth method is used for time discretization. Smagorinsky parameter is 0.004 and Newton-Raphson method is used for linearization of convective term.

5 Comparison of Results for Two Turbulent Models and Their Effect on Aeroacoustic Sources and Acoustic Propagation

It is noticeable from Figs. 3 and 4 that both approaches of modelling turbulent flow give different representation of the velocity field around the airfoil. Even with the same model, mesh and boundary conditions, the difference in turbulent pattern around the airfoil is noticeable on the first side. LES approach presents a fixed pattern of the behaviour of velocity giving more dissipative representation [8], capturing only large eddies and giving poor representation of small scales and their energy influence on large scales. It is noticeable that LES methodology didn't recover the influence of the rift on the flow around the airfoil leading to bad results for acoustic sources which are produced inside the rift. On the other hand, the orthogonal SGS method with dynamical subscales has some features which give better presentation of turbulent flow. Even the flow inside the rift is reproduced with the influence on around flow.

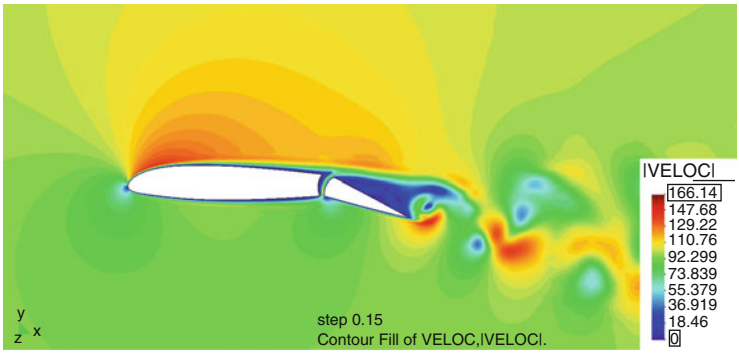

Fig. 3 Velocity profile obtained with LES-Smagorinsky model around airfoil after 0.15 s

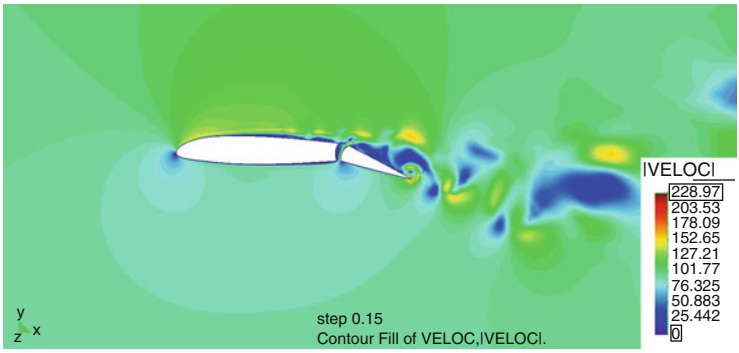

Fig. 4 Velocity profile obtained with Orthogonal SGS stabilized method with dynamical subscales after 0.15 s

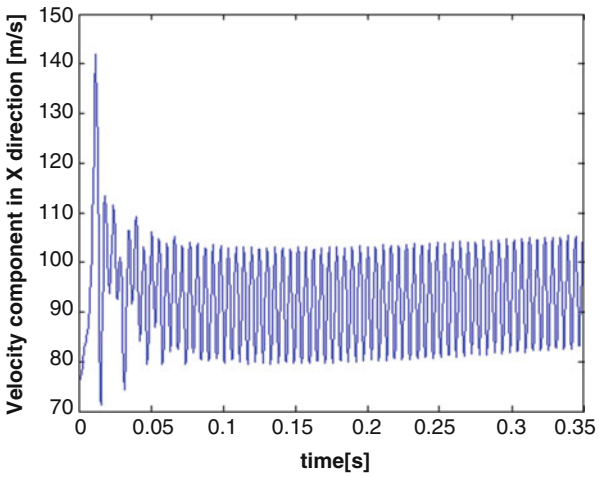

Fig. 5 Time tracking of velocity in X direction using LES-Smagorinsky method

First, forcing L^2 projection of small scales on the velocity finite element space leads to orthogonal subgrid scale, which gives proper scale separation in the sense that total kinetic energy is the sum of kinetic energy of solvable (grid) scale plus kinetic energy of small (non-grid) scales. Second, modelling of dynamical subscales leads to correct behaviour of time discretization schemes and better accuracy. On the other side, dynamical tracking of subscales gives the opportunity to model backscatter [9] that gives right energy transfer between the large and the small scales.

In Figs. 5, 6, 7, 8 the dissipative structure of the LES approach is even more visible, giving poor spectral analysis. Recovering only large scales, giving good evidence this approach can capture turbulent characteristics to some point on the Kolmogorov scale diagram.

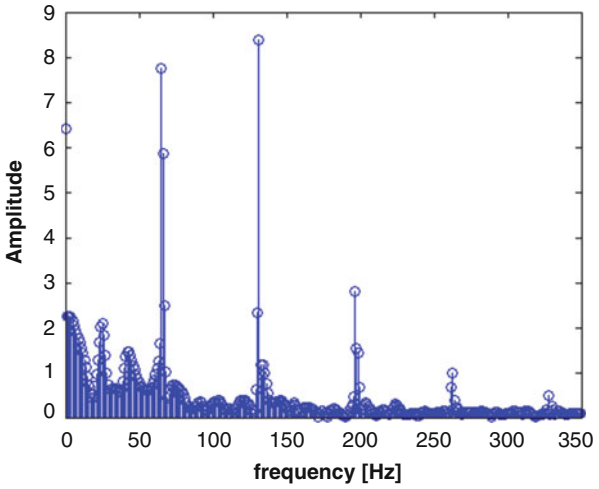

Fig. 6 FFT of X component velocity function using LES

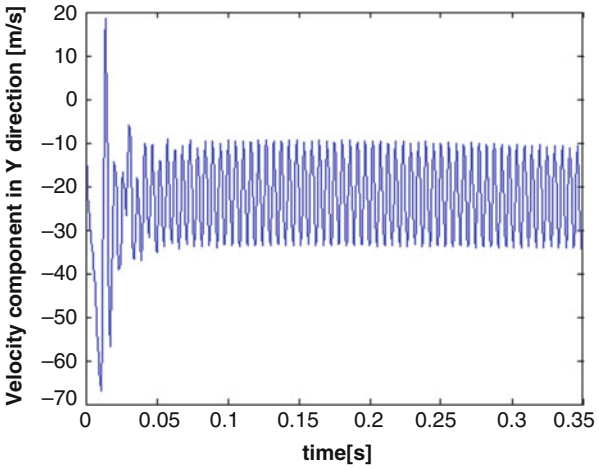

Fig. 7 Time tracking of velocity in Y direction using LES-Smagorinsky method

In Figs. 9, 10, 11, 12 the Orthogonal SGS with dynamical subscale, on the other hand, recovers much richer spectral diagram, modelling both small and large scales and their energy interchange [10, 11]. It is important to highlight that representation in the figures is done for the same points shown in Fig. 2 for both methods. Spectral analysis in these figures gives rich representation, provides good evidence of recovering small scales and their influence on large scales. It is evident from Figs. 13 and 14 that the proposed methodology of orthogonal SGS with dynamical subscales gives stronger and richer presentation of acoustic sources, giving smaller dipoles that come from small scales and their extra modeling [12, 13]. Good

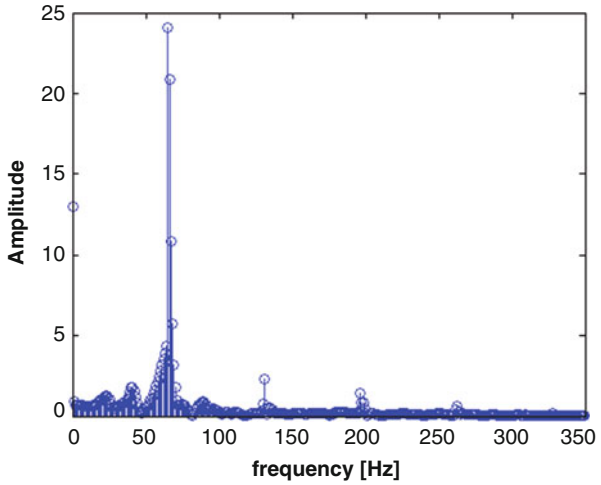

Fig. 8 FFT of Y component velocity function using LES

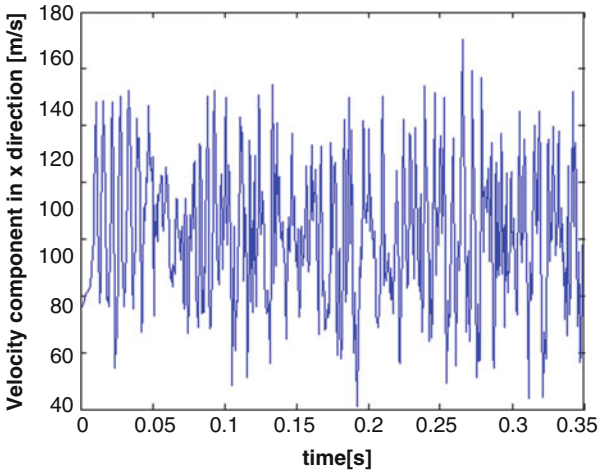

Fig. 9 Time tracking of velocity in X direction using Orthog. SGS method with dyn. subscales

modeling of energy transfer in flow between the large and the small eddies recovers the fluctuating nature, which gives nice accent on gradients in the field around the airfoil, which is shown best in the rift between the airfoil body and the flap. Both methods recover to the some point trailing edge generation of the noise, but LES model is very dissipative for the rift flow [14]. For acoustic wave propagation we are interested in small wave numbers, because large values produce stabilization problems, known as pollution error. Because of that, only the implementation of the Galerkin method will show dependencies of different modeling of turbulent flow on calculation of acoustic wave propagation. It is important to mention that

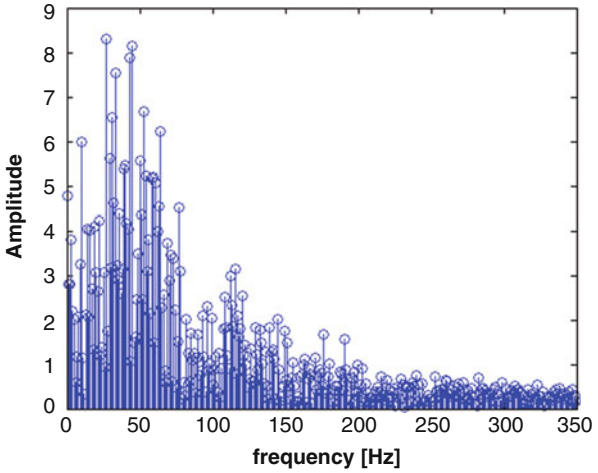

Fig. 10 FFT of X component velocity function using Orthogonal SGS with dynamical subscales

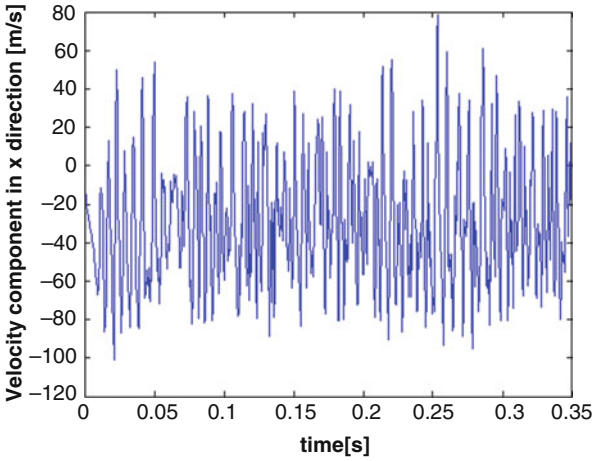

Fig. 11 Time tracking of velocity in Y direction using Orthogonal SGS method with dynamical subscales

the CFD domain is the same for acoustic field and wave numbers are same for all simulations, and only different is the acoustic source term that comes from different modelling of turbulent flow. Better modelling of turbulent flow and richer approximation of acoustic sources also affect wave propagation of pressure and solution of inhomogeneous Helmholtz equation, as represented in Figs. 15 and 16 [15]. It is clear that the Orthogonal SGS method with dynamical subscales gives stronger waves in the field.

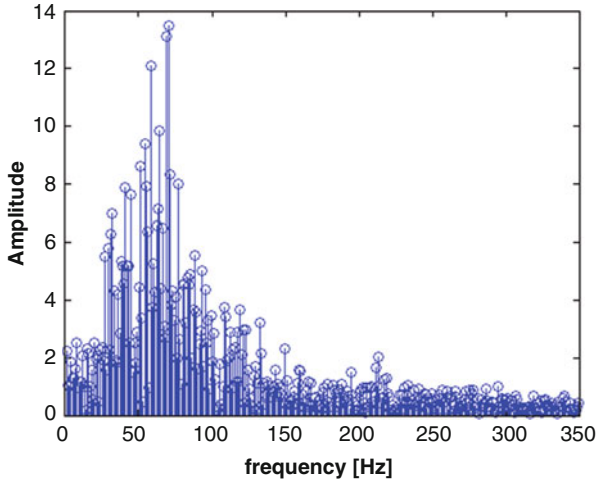

Fig. 12 FFT of Y component velocity function using Orthogonal SGS with dynamical subscales

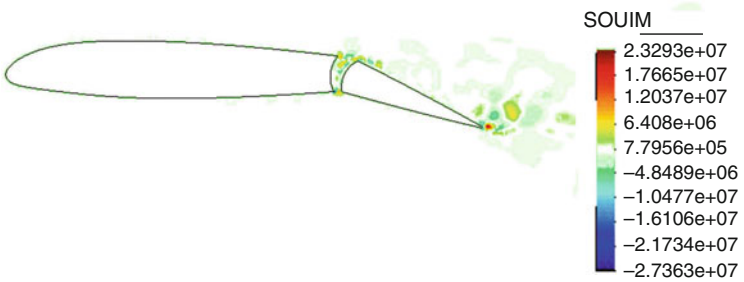

Fig. 13 Imaginary solution of acoustic sources where is used LES-Smagorinsky model for turbulance

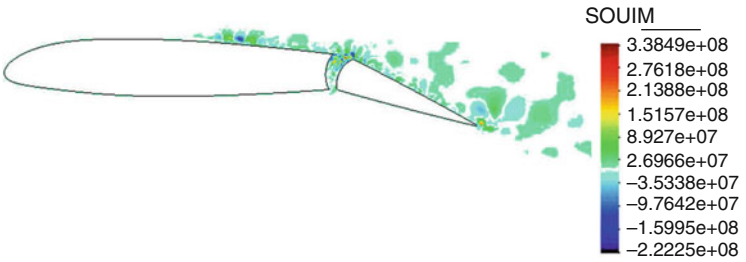

Fig. 14 Imaginary solution of acoustic sources where is used Orthogonal SGS method with dynamical subscale model for turbulance

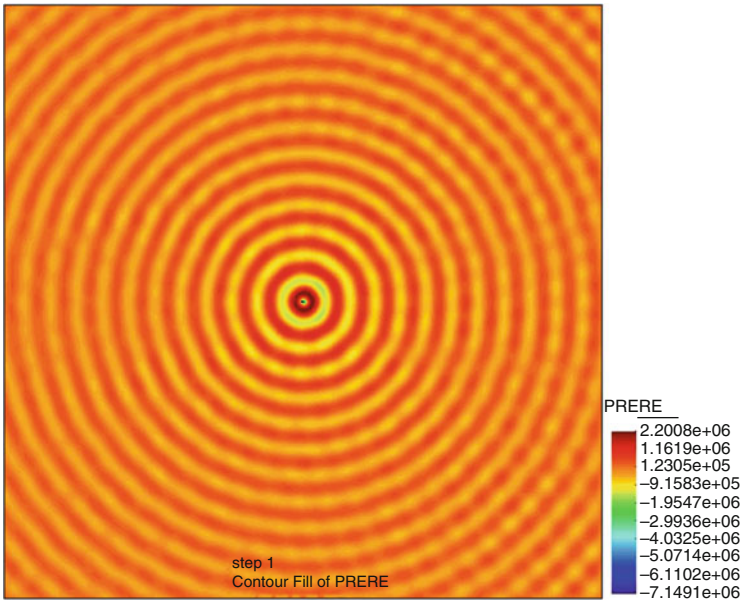

Fig. 15 Real solution of acoustic pressure field for Orthogonal SGS method with dynamical subscale approach of turbulent flow

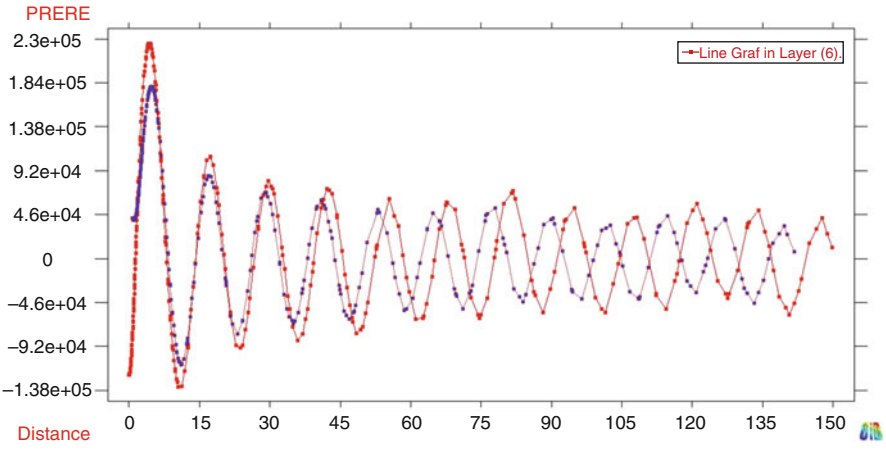

Fig. 16 Pressure distribution in the cutting plane of Fig. 15

6 Discussion and Conclusion

It is obvious that for the airfoil with deployed flap, the method of Orthogonal SGS with dynamical scale which provides stabilization and model turbulent flow gives richer and stronger presentations of aeroacoustics sources, which leads to the same conclusion in propagation of acoustics waves in acoustic domain. The point to be highlighted is in a different approach of modelling small scales and their filtration in the solution of resolvable scale captured by the finite element mesh recovering good energy distribution in area of small eddies. Also, using dynamic subscales, the method gives the opportunity of modeling backscatter across large and small eddies giving energy flow across them. In addition, the model is less dissipative as clearly shown in spectral diagrams exhibiting the possibility to recover a wide range of frequencies arising from small scales and their energy somehow lost in the LES approach. Comparison of time tracking in the same point in mesh for both methods is clearly distinctive between them and their modelling of turbulent flow. Modelling of turbulent flow is directly affect Lighthill's tensor producing different results for both methods. It is clear that the proposed method recovers richer distribution of acoustic sources and also their strength. Different modeling of small eddies directly influences the distribution of dipoles in the near field of the airfoil with deployed flap. Previously obtained results also affect inhomogeneous Helmholtz equation. Richer and stronger source term arising from the proposed method shows stronger waves pressure in the calculated domain. Besides, the area near the airfoil is clearly distinguished, where there is a concentration of acoustic sources.

References

1. Ffowcs Williams, J.E., Hawkings, D.L.: Sound generated by turbulence and surfaces in arbitrary motion. Phil. Trans. R. Soc. Lond. **264**, 321–342 (1969)
2. Hardin, J., Hussaini, M.: Computational Aeroacoustics. Springer, Berlin (1993)
3. Lyrintzis, A.: Review of advances in aeroacoustics (in honor of Professor Geoffrey M. Lilley). AIAA J. **49**, 2334–2335 (2011)
4. Canuto, V.: Large eddy simulation of turbulence: a subgrid scale model including shear, vorticity, rotation and buoyancy. Astrophys. J. **428**, 729–758 (1994)
5. Codina, R., Principe, J., Guasch, O., Badia, S.: Time dependent subscales in the stabilized finite element approximation of incompressible flow problems. Comput. Methods Appl. Mech. Eng. **196**, 2413–2430 (2007)
6. Codina, R.: Stabilized finite element approximation of transient incompressible flows using orthogonal subscales. Comput. Methods Appl. Mech. Eng. **191**, 4295–4321 (2002)
7. Codina, R., Principe, J., Baiges, J.: Subscales on the element boundaries in the variational two-scale finite element method. Comput. Methods Appl. Mech. Eng. **198**, 838–852 (2009)
8. Principe, J., Codina, R., Henke, F.: The dissipative structure of variational multiscale methods for incompressible flows. Comput. Methods Appl. Mech. Eng. **199**, 791–801 (2010)
9. Badia, S., Codina, R., Gutierrez-Santacreu, J.V.: Long-term stability estimates and existence of a global attractor in a finite element approximation of a global attractor in a finite element approximation of the Navier-Stokes equations with numerical subgrid scale modelling. SIAM J. Numer. Anal. **48**, 1013–1037 (2010)

10. Jazarević, V., Rašuo, B.: Computation of acoustic sources for the landing gear during the takeoff and landing. FME Trans. **41**(3), 180–188 (2013)
11. Rašuo, B., Jazarević, V.: Numerical calculation of acoustic sources for the landing gear of aeroplane during take-off and landing. Proc. Appl. Math. Mech. **15**(1), 529–530 (2015)
12. Rašuo, B., Jazarević, V.: Numerical calculation of acoustic sources for the landing gear of aeroplane during take-off and landing. In: GAMM 86th Annual Scientific Conference, Lecce. Book of Abstract, p. 476 (2015)
13. Jazarević, V.: Optimization of aeroacoustic phenomena around aerodynamic surfaces. Doctoral Thesis (in Serbian), Aeronautical Department, Faculty of Mechanical Engineering, University of Belgrade (2016)
14. Rašuo, B., Jazarević, V.: Numerical calculation of aerodynamic noise generated from an aircraft in low Mach number flight. In: BAIL 2016, Beijing, Book of Abstract, pp. 24–25 (2016)
15. Jazarević V., Rašuo, B.: Numerical prediction of aerodynamic noise generated from missile for low mach number flows. Technical Gazette **24**(3), 663–670 (2017)

On the Discrete Maximum Principle for Algebraic Flux Correction Schemes with Limiters of Upwind Type

Petr Knobloch

Abstract Algebraic flux correction (AFC) schemes are applied to the numerical solution of scalar steady-state convection-diffusion-reaction equations. A general result on the discrete maximum principle (DMP) is established under a weak assumption on the limiters and used for proving the DMP for a particular limiter of upwind type under an assumption that may hold also on non-Delaunay meshes. Moreover, a simple modification of this limiter is proposed that guarantees the validity of the DMP on arbitrary simplicial meshes. Furthermore, it is shown that AFC schemes do not provide sharp approximations of boundary layers if meshes do not respect the convection direction in an appropriate way.

1 Introduction

The aim of this paper is the numerical solution of the scalar steady-state convection-diffusion-reaction equation

$$-\varepsilon\,\Delta u + \mathbf{b}\cdot\nabla u + c\,u = g \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega\,, \tag{1}$$

defined in a bounded d-dimensional domain Ω ($d = 2, 3$) having a polygonal (resp. polyhedral) Lipschitz-continuous boundary $\partial\Omega$. We assume that $\varepsilon > 0$ is constant, and $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, and $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$ are given functions satisfying $\nabla\cdot\mathbf{b} = 0$ and $c \geq 0$ a.e. in Ω.

Problem (1) is discretized by means of the finite element method and stabilized using algebraic flux correction (AFC), following the ideas in [3–6]. The first rigorous analysis of the AFC scheme considered in this paper was published in [1] where, in particular, the validity of the discrete maximum principle (DMP) for the limiter of [3] was proved for Delaunay meshes. In [2], another limiter was proposed

P. Knobloch (✉)
Faculty of Mathematics and Physics, Department of Numerical Mathematics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_10

for which the AFC scheme is linearity preserving and satisfies the DMP on arbitrary simplicial meshes. The design of this limiter was based on a general result on the DMP that is, however, not applicable to the limiter of [3] due to its upwind character.

In what follows, we first formulate the AFC scheme and then, in Sect. 3, we prove a local DMP under a more general assumption on the limiters than in [2]. This enables us to prove in Sect. 4 that, for the limiter of [3], the DMP is satisfied under a condition that may hold also on non-Delaunay meshes. Moreover, the result of Sect. 3 makes it possible to introduce a limiter of upwind type for which the DMP is satisfied on arbitrary simplicial meshes. Finally, in Sect. 5, we discuss the upwind character of the limiters and show that AFC schemes provide sharp approximations of boundary layers only when used on appropriate meshes.

2 An Algebraic Flux Correction Scheme

In this section we formulate the AFC scheme investigated in this paper. First we introduce a finite element discretization of the problem (1), which is of Galerkin type and hence unstable in the convection-dominated regime. Then we apply the algebraic flux correction to enforce the discrete maximum principle.

To discretize the problem (1), we introduce finite element spaces

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in P_1(T)\ \forall\, T \in \mathscr{T}_h\}\,, \qquad V_h = W_h \cap H_0^1(\Omega)\,,$$

where $\mathscr{T}_h$ is a simplicial triangulation of Ω and $P_1(T)$ is the space of linear polynomials on T. We denote by $x_1, \ldots, x_N$ the vertices of $\mathscr{T}_h$ and assume that M of them ($0 < M < N$) are interior vertices which are numbered first, i.e., $x_1, \ldots, x_M \in \Omega$ and $x_{M+1}, \ldots, x_N \in \partial\Omega$. We denote by $\varphi_1, \ldots, \varphi_N$ the standard basis functions of W_h, i.e., one has $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol. Then the functions $\varphi_1, \ldots, \varphi_M$ form a basis in V_h. For any $v_h \in W_h$, we denote by $\{v_i\}_{i=1}^N$ the uniquely determined coefficients of v_h with respect to the above basis of W_h, i.e., $v_h = \sum_{i=1}^N v_i\,\varphi_i$. Let us introduce the following discretization of (1).

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a_h(u_h, v_h) = (g, v_h) \qquad \forall\ v_h \in V_h\,, \tag{2}$$

where

$$a_h(u_h, v_h) = \varepsilon\,(\nabla u_h, \nabla v_h) + (\mathbf{b}\cdot\nabla u_h, v_h) + \sum_{i=1}^{M} (c, \varphi_i)\, u_i\, v_i \qquad \forall\ u_h \in W_h,\ v_h \in V_h\,,$$

and $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. Note that the usual reaction term $(c\,u_h, v_h)$ is replaced by a diagonal approximation, analogous to a mass lumping in discretizations of transient problems.

We denote $a_{ij} = a_h(\varphi_j, \varphi_i)$ for $i,j = 1,\ldots,N$, $g_i = (g, \varphi_i)$ for $i = 1,\ldots,M$, and $u_i^b = u_b(x_i)$ for $i = M+1,\ldots,N$. Then u_h is a solution of (2) if and only if the corresponding coefficient vector $U = (u_1,\ldots,u_N)$ satisfies the linear system

$$\sum_{j=1}^{N} a_{ij}\,u_j = g_i\,, \qquad i = 1,\ldots,M\,, \tag{3}$$

$$u_i = u_i^b\,, \qquad i = M+1,\ldots,N\,. \tag{4}$$

Since $(\mathbf{b}\cdot\nabla v, v) = 0$ for any $v \in H_0^1(\Omega)$ and $\sum_{i=1}^N \varphi_i = 1$ in Ω, it is easy to see that

$$a_{ii} > 0\,, \qquad \sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1,\ldots,M\,. \tag{5}$$

Moreover, one deduces that the matrix $(a_{ij})_{i,j=1}^M$ is positive definite so that the linear system (3), (4) (and hence also the discrete problem (2)) has a unique solution.

To stabilize the discretization (2), the algebraic flux correction is applied as described in [1]. This leads to the following system of nonlinear equations:

$$\sum_{j=1}^{N} a_{ij}\,u_j + \sum_{j=1}^{N} (1-\alpha_{ij}(U))\,d_{ij}\,(u_j - u_i) = g_i\,, \qquad i = 1,\ldots,M\,, \tag{6}$$

$$u_i = u_i^b\,, \qquad i = M+1,\ldots,N\,, \tag{7}$$

where $\alpha_{ij}(U) \in [0,1]$, $i,j = 1,\ldots,N$, are limiters that depend on the solution $U = (u_1,\ldots,u_N)$ and form a symmetric matrix, and d_{ij} are entries of an artificial diffusion matrix defined by

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\ i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j\neq i} d_{ij}\,.$$

Moreover, for any $i,j \in \{1,\ldots,N\}$, we assume that $\alpha_{ij}(U)(u_j - u_i)$ is a continuous function of $U = (u_1,\ldots,u_N)$. This property makes it possible to prove that the nonlinear problem (6), (7) has a solution, see [1].

In the following sections we shall formulate general properties and particular examples of limiters α_{ij} for which the AFC scheme (6), (7) satisfies the DMP. It can be verified (see [1]) that the limiters presented in this paper possess the above-mentioned continuity property.

3 A General Result on the Discrete Maximum Principle

In this section we prove a local DMP for the AFC scheme (6), (7). This result will be established under a general assumption on the limiters α_{ij}, which is weaker than the one considered in [2].

Given $i \in \{1, \ldots, M\}$, the DMP will be formulated locally, with respect to an index set $S_i \subset \{1, \ldots, N\} \setminus \{i\}$. We assume that

$$S_i \supset \{j \in \{1, \ldots, N\} \setminus \{i\} \,:\, a_{ij} \neq 0 \ \text{ or } \ a_{ji} > 0\}\,, \qquad i = 1, \ldots, M\,. \tag{8}$$

We shall investigate the validity of the DMP in the following sense.

Definition 1 The AFC scheme (6), (7) satisfies the local DMP if, for any vector $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$ satisfying (6) and any $i \in \{1, \ldots, M\}$, the implications

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j \in S_i} u_j^+\,, \qquad\qquad g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j \in S_i} u_j^- \tag{9}$$

hold true, with $u_j^+ = \max\{0, u_j\}$ and $u_j^- = \min\{0, u_j\}$. If $\sum_{j=1}^N a_{ij} = 0$, it is also required that

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j \in S_i} u_j\,, \qquad\qquad g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j \in S_i} u_j\,. \tag{10}$$

To prove the local DMP, we make the following assumption.

Assumption (A) Consider any $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$ and any $i \in \{1, \ldots, M\}$. If u_i is a strict local extremum of U with respect to S_i, i.e.,

$$u_i > u_j \quad \forall\, j \in S_i \qquad \text{or} \qquad u_i < u_j \quad \forall\, j \in S_i\,,$$

then

$$a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij} \leq 0 \qquad \forall\, j \in S_i\,.$$

In [2], the local DMP was proved under the assumption that $\{\alpha_{ij}(U)\, d_{ij}\}_{j \in S_i}$ vanish if u_i is a strict local extremum of U with respect to S_i. This assumption is obviously stronger than the present Assumption (A) since $a_{ij} + d_{ij} \leq 0$ for any $i \neq j$.

Theorem 1 *The AFC scheme (6), (7) with limiters α_{ij} satisfying Assumption (A) satisfies the local DMP.*

Proof Let $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$ satisfy (6). Consider any $i \in \{1, \ldots, M\}$ and let $g_i \leq 0$. Since $d_{ij} = 0$ for any $j \notin S_i \cup \{i\}$, it follows from (6) that

$$A_i\, u_i + \sum_{j \in S_i} [a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij}]\, (u_j - u_i) = g_i\,, \tag{11}$$

where $A_i = \sum_{j=1}^{N} a_{ij}$. If $A_i > 0$, it suffices to consider $u_i > 0$ since otherwise the first implication in (9) trivially holds. Let us assume that $u_i > u_j$ for all $j \in S_i$. Then Assumption (A) implies that the sum in (11) is non-negative. If $A_i = 0$, then there is $j \in S_i$ such that $a_{ij} < 0$ since $a_{ii} > 0$ (see (5)). As $d_{ij} \le 0$, this implies that the sum in (11) is positive. If $A_i > 0$, then $A_i u_i > 0$. Thus, in both cases, the left-hand side of (11) is positive, which is a contradiction. Therefore, there is $j \in S_i$ such that $u_i \le u_j$, which proves the first implication in (10) and hence also in (9). The statements for $g_i \ge 0$ follow in an analogous way.

4 Validity of the DMP for Particular Limiters

In this section we first present the definition of the limiters α_{ij} proposed in [3]. This choice is often used in computations and was thoroughly investigated in [1]. We prove the validity of the DMP for the AFC scheme (6), (7) with these limiters under a weaker assumption than in [1] but we also show that the DMP cannot be guaranteed on non-Delaunay meshes. Therefore, we introduce a modification of these limiters for which the DMP holds on arbitrary meshes. The results are valid for any index sets S_i satisfying (8).

To define the limiter of [3], one first computes, for $i = 1, \ldots, M$,

$$P_i^+ = \sum_{\substack{j=1 \\ a_{ji} \le a_{ij}}}^{N} f_{ij}^+ , \qquad P_i^- = \sum_{\substack{j=1 \\ a_{ji} \le a_{ij}}}^{N} f_{ij}^- , \qquad Q_i^+ = -\sum_{j=1}^{N} f_{ij}^- , \qquad Q_i^- = -\sum_{j=1}^{N} f_{ij}^+ , \tag{12}$$

where $f_{ij} = d_{ij}(u_j - u_i)$, $f_{ij}^+ = \max\{0, f_{ij}\}$, and $f_{ij}^- = \min\{0, f_{ij}\}$. Then, one defines

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\} , \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\} , \qquad i = 1, \ldots, M .$$

If P_i^+ or P_i^- vanishes, one sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. For $i = M+1$, $\ldots, N$, one defines $R_i^+ = R_i^- = 1$. Furthermore, one sets

$$\widetilde{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 , \\ 1 & \text{if } f_{ij} = 0 , \\ R_i^- & \text{if } f_{ij} < 0 , \end{cases} \qquad i, j = 1, \ldots, N .$$

Finally, one defines

Limiter 1 For any $i, j \in \{1, \ldots, N\}$ with $a_{ji} \le a_{ij}$, one sets $\alpha_{ij} = \alpha_{ji} = \widetilde{\alpha}_{ij}$.

It was proved in [1] that the AFC scheme (6), (7) with Limiter 1 satisfies a local DMP provided that

$$a_{ij} + a_{ji} \le 0 \qquad \forall\ i,j = 1,\dots,N\,,\ i \ne j\,,\ i \le M \text{ or } j \le M\,. \tag{13}$$

As discussed in [1], the validity of (13) is guaranteed if the triangulation $\mathscr{T}_h$ is weakly acute. In the two-dimensional case, (13) holds if and (in principle) only if $\mathscr{T}_h$ is a Delaunay triangulation, i.e., the sum of any pair of angles opposite a common edge is smaller than, or equal to, π. The following theorem shows that the assumption (13) can be weakened.

Theorem 2 *Let*

$$\min\{a_{ij}, a_{ji}\} \le 0 \qquad \forall\ i = 1,\dots,M\,,\ j = 1,\dots,N\,,\ i \ne j\,. \tag{14}$$

Then the AFC scheme (6), (7) *with Limiter 1 satisfies the local DMP.*

Proof It suffices to verify Assumption (A). Consider any $U = (u_1,\dots,u_N) \in \mathbb{R}^N$, $i \in \{1,\dots,M\}$, and $j \in S_i$. Let u_i be a strict local extremum of U with respect to S_i. We want to prove that

$$a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij} \le 0\,. \tag{15}$$

If $a_{ij} \le 0$, then (15) holds since $(1 - \alpha_{ij}(U))\, d_{ij} \le 0$. If $a_{ij} > 0$, then $a_{ji} \le 0$ due to (14) and hence $\alpha_{ij} = \widetilde{\alpha}_{ij}$ and $d_{ij} = -a_{ij} < 0$. If $u_i > u_k$ for any $k \in S_i$, then $f_{ij} > 0$ and $f_{ik} \ge 0$ for $k = 1,\dots,N$, so that $\widetilde{\alpha}_{ij} = R_i^+ = 0$. Similarly, if $u_i < u_k$ for any $k \in S_i$, then $f_{ij} < 0$ and $f_{ik} \le 0$ for $k = 1,\dots,N$, so that $\widetilde{\alpha}_{ij} = R_i^- = 0$. Thus, (15) holds again.

Obviously, the assumption (13) implies (14) and hence, in particular, (14) holds if $\mathscr{T}_h$ is a Delaunay triangulation. However, in contrast to (13), the condition (14) may be satisfied also on non-Delaunay meshes, particularly in the convection-dominated case, since the convection matrix is skew-symmetric. If the condition (14) is not satisfied, then the DMP generally does not hold, as the following result shows.

Theorem 3 *Let there exist* $i \in \{1,\dots,M\}$ *and* $k \in \{1,\dots,N\}$ *such that* $i \ne k$ *and* $0 < a_{ik} < a_{ki}$. *If* $k \in \{1,\dots,M\}$, *let there exist* $l \in S_k \setminus (S_i \cup \{i\})$ *such that* $d_{kl} \ne 0$. *Then the AFC scheme* (6), (7) *with Limiter 1 does not satisfy the local DMP.*

Proof Let i, k, and l satisfy the assumptions of the theorem. Consider any $u_i > 0$ and any $u_j < u_i$ for $j \in S_i \setminus \{k\}$. Let $u_k < u_i$ be such that

$$\sum_{j \in S_i \cup \{i\}} a_{ij}\, u_j + \sum_{j \in S_i \setminus \{k\}} d_{ij}\, (u_j - u_i) \le 0\,.$$

We shall show that the remaining components of $(u_1, \ldots, u_N)$ can be defined in such a way that $\alpha_{ik} = 1$. Then the ith equation of the AFC scheme (6), (7) is satisfied for $g_i \le 0$ but $u_i > \max_{j \in S_i} u_j^+$ so that (9) does not hold. Since $a_{ik} < a_{ki}$ and $f_{ki} < 0$, one deduces that $\alpha_{ik} = \widetilde{\alpha}_{ki} = R_k^-$. Thus, if $k > M$, one always has $\alpha_{ik} = 1$ and u_j can be defined arbitrarily for $j \notin S_i \cup \{i\}$. If $k \le M$, one defines u_j arbitrarily for $j \notin S_i \cup \{i, l\}$. Then, it suffices to choose $u_l < u_k$ in such a way that $Q_k^- \le P_k^-$. This is possible, since $Q_k^- \le d_{kl}\,(u_k - u_l)$, $d_{kl} \ne 0$ and P_k^- is independent of u_l if $u_l < u_k$.

It is easy to define non-Delaunay triangulations of Ω and data $\varepsilon > 0$ and $\mathbf{b} \in W^{1,\infty}(\Omega)^d$ such that the assumptions of Theorem 3 are satisfied. This shows that, on non-Delaunay triangulations, the AFC scheme (6), (7) does not satisfy the DMP in general. Therefore, we introduce the following simple modification of Limiter 1 that assures the validity of the DMP on arbitrary triangulations (the proof is analogous as for Theorem 2).

Limiter 2 For any $i, j \in \{1, \ldots, N\}$, the limiters $\alpha_{ij} = \alpha_{ji}$ are defined as for Limiter 1 if $\min\{a_{ij}, a_{ji}\} \le 0$ and otherwise set to $\min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}$, omitting the condition $a_{ji} \le a_{ij}$ in (12).

Remark 1 The definition of Limiter 1 is ambiguous if $a_{ij} = a_{ji}$. For Limiter 2, this ambiguity is restricted to the case $\min\{a_{ij}, a_{ji}\} \le 0$ so that it does not influence the resulting method. Indeed, if $a_{ij} = a_{ji} \le 0$, then $d_{ij} = 0$ so that the respective α_{ij} does not occur in the nonlinear problem (6), (7), and can be defined arbitrarily.

It is easy to see that the AFC scheme (6), (7) satisfies the local DMP also for

Limiter 3 For any $i, j \in \{1, \ldots, N\}$, one sets $\alpha_{ij} = \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}$, omitting the condition $a_{ji} \le a_{ij}$ in (12).

However, Limiter 3 generally introduces more artificial diffusion than Limiter 2 and hence leads to a more pronounced smearing of layers. Let us illustrate this by means of the following example.

Example 1 We consider the problem (1) defined in $\Omega = (0, 1)^2$ with the data $\varepsilon = 10^{-8}$, $\mathbf{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = 0$, $g = 0$, and the boundary condition

$$u_b(x, y) = \begin{cases} 0 & \text{for } x = 1 \text{ or } y \le 0.7, \\ 1 & \text{else.} \end{cases}$$

Figure 2 shows approximate solutions obtained using the AFC scheme (6), (7) on a triangulation of the type depicted in Fig. 1a consisting of 800 triangles. In this case, the assumption (14) is satisfied and hence Limiters 1 and 2 are identical. One can observe that Limiter 3 leads to a stronger smearing of the layers than Limiter 1.

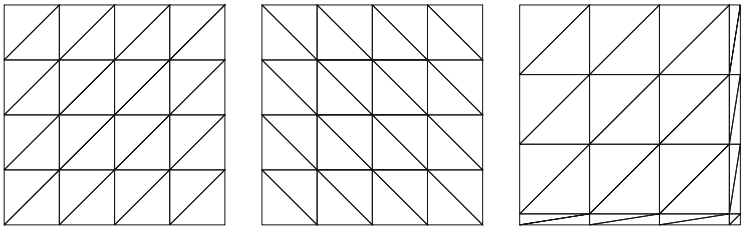

Fig. 1 Types of triangulations used in numerical experiments

5 Upwind Character of the Limiters

In this section we show that Limiter 1 (and hence also the related Limiter 2) can be regarded as a limiter of upwind type. We shall also investigate the question of sharp approximation of layers at outflow boundaries and demonstrate that smearing can be avoided only if the underlying triangulation is defined appropriately.

To define $\alpha_{ij} = \alpha_{ji}$ in Limiter 1 (with $i \neq j$), one chooses either $\widetilde{\alpha}_{ij}$ or $\widetilde{\alpha}_{ji}$, depending on the validity of the inequality $a_{ji} < a_{ij}$ (we do not consider the case $a_{ij} = a_{ji}$, cf. Remark 1). Let us investigate the meaning of this inequality. One may assume that $i \leq M$ or $j \leq M$ since α_{ij} with $i, j \in \{M+1, \ldots, N\}$ does not occur in (6). Then $(\mathbf{b} \cdot \nabla\varphi_j, \varphi_i) = -(\mathbf{b} \cdot \nabla\varphi_i, \varphi_j)$ and hence the inequality $a_{ji} < a_{ij}$ is equivalent to the condition $(\mathbf{b} \cdot \nabla\varphi_j, \varphi_i) > 0$.

Let $\mathbf{b}$ be constant and denote $\mathbf{d}^{ij} = \int_\Omega \varphi_i \nabla\varphi_j \, \mathrm{d}x$. Then $(\mathbf{b} \cdot \nabla\varphi_j, \varphi_i) = \mathbf{b} \cdot \mathbf{d}^{ij}$. Our aim is to analyze the relation of the vector $\mathbf{d}^{ij}$ to the edge E_{ij} with endpoints x_i, x_j (if x_i, x_j are not endpoints of an edge, then $a_{ij} = a_{ji} = 0$). For simplicity, let us consider the two-dimensional case. Let x_k, x_l be the remaining vertices of the two elements of $\mathscr{T}_h$ adjacent to E_{ij} which we denote T_k, T_l, respectively. Furthermore, we denote by E_{jk} and E_{jl} the edges with endpoints x_j, x_k and x_j, x_l, respectively, and introduce the line segment $E_{kl}^\star$ with endpoints x_k, x_l. Finally, let $\mathbf{n}^{ij}$ be the unit normal vector to $E_{kl}^\star$ satisfying $(x_j - x_i) \cdot \mathbf{n}^{ij} > 0$. Denoting by $\mathbf{n}_{\partial(T_k \cup T_l)}$ the unit outward normal vector to the boundary of the set $T_k \cup T_l$ and using the fact that $\nabla\varphi_j$ is piecewise constant, one derives

$$\mathbf{d}^{ij} = \frac{1}{3} \int_{T_k \cup T_l} \nabla\varphi_j \, \mathrm{d}x = \frac{1}{3} \int_{E_{jk} \cup E_{jl}} \varphi_j \, \mathbf{n}_{\partial(T_k \cup T_l)} \, \mathrm{d}s = \frac{1}{6} \int_{E_{kl}^\star} \mathbf{n}^{ij} \, \mathrm{d}s = \frac{|E_{kl}^\star|}{6} \, \mathbf{n}^{ij} \,,$$

where $|E_{kl}^\star|$ denotes the length of $E_{kl}^\star$. Consequently, $(x_j - x_i) \cdot \mathbf{d}^{ij} > 0$. Thus, if $\mathbf{b}$ is aligned with the edge E_{ij}, then $\mathbf{b} \cdot \mathbf{d}^{ij} > 0$ if and only if x_i is the upwind vertex. Moreover, if the line segment $E_{kl}^\star$ is orthogonal to the edge E_{ij}, then $\mathbf{d}^{ij} = \alpha \, (x_j - x_i)$ with $\alpha > 0$ and hence, for any $\mathbf{b} \in \mathbb{R}^2$, it follows that again $\mathbf{b} \cdot \mathbf{d}^{ij} > 0$ if and only if x_i is the upwind vertex. Thus, in these (and many other) cases, the application of the inequality $a_{ji} < a_{ij}$ in the definition

of Limiter 1 causes that $\alpha_{ij} = \alpha_{ji}$ is defined using quantities computed at the upwind vertex. It turns out that this feature has a positive influence on the quality of the approximate solutions and on the convergence of the iterative process for solving the nonlinear problem (6), (7). Nevertheless, one should be aware that, in general, the upwind character of Limiter 1 depends on the used mesh.

It can be expected that the upwind character of the method is also important for a sharp approximation of layers at outflow boundaries. However, this problem is more complicated as we shall show in the following. We shall concentrate on Example 1 for which we observed that the AFC scheme (6), (7) with the limiters from the previous section does not provide a sharp approximation of the layer along the boundary part $\Gamma = \{(1, y)\,;\ y \in (0, 1)\}$ for the considered triangulation. At interior vertices of this triangulation near Γ, the values of the exact solution are indistinguishable from the value 1 in the finite precision arithmetic. A natural question is whether a corresponding mesh function $U = (u_1, \ldots, u_N)$ may satisfy those equations of the AFC scheme (6) which correspond to vertices near Γ. We shall now consider arbitrary limiters α_{ij} satisfying the assumptions of Sect. 2 and Assumption (A) with sets S_i containing only indices corresponding to vertices connected by edges with x_i (which is the standard case). We choose any vertex $x_i \in \Omega$ connected by an edge with a vertex lying on Γ (but not with a vertex on $\partial\Omega \setminus \Gamma$). Then $u_i = 1$ and $u_j \in \{0, 1\}$ for $j \in S_i$. Since u_i is a local extremum of U with respect to S_i, it follows from Assumption (A) and the continuity of $\alpha_{ij}(U)(u_j - u_i)$ that

$$[a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij}]\,(u_j - u_i) \geq 0 \qquad \forall\ j \in S_i\,. \tag{16}$$

For the data of Example 1, the AFC scheme can be written in the form (11) with $A_i = g_i = 0$. Thus, in view of (16), the ith equation of the AFC scheme can be satisfied if and only if (16) holds with equality. This implies that

$$a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij} = 0 \qquad \forall\ j \in S_i^{\partial\Omega} := S_i \cap \{M + 1, \ldots, N\}\,, \tag{17}$$

since $u_j = 0$ for $j \in S_i^{\partial\Omega}$. As $d_{ij} \leq 0$, a necessary condition for the validity of (17) is $a_{ij} \geq 0$ for all $j \in S_i^{\partial\Omega}$. On the other hand, if $a_{ij} \geq 0$ for some $j \in S_i^{\partial\Omega}$, then $a_{ji} \leq a_{ij}$ (since we use a Delaunay triangulation and hence (13) holds) so that $d_{ij} = -a_{ij}$ and (16) implies that $\alpha_{ij}(U)\, d_{ij} = 0$. Consequently, $a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij} = 0$, which means that the condition $a_{ij} \geq 0$ is also sufficient in this case. Note also that the limiters from the previous section then define the value α_{ij} using quantities computed at the interior vertex x_i which again may be interpreted as an upwind feature.

To check whether the condition $a_{ij} \geq 0$ (with i and j as above) is satisfied, one may employ that $a_{ij} = \varepsilon\,(\nabla\varphi_j, \nabla\varphi_i) + \mathbf{b}\cdot\mathbf{d}^{ij}$. Then one easily sees that this condition is not satisfied for one of the two indices $j \in S_i^{\partial\Omega}$ when one considers a triangulation of the type depicted in Fig. 1a. This explains the smearing of the layer along Γ observed in Fig. 2. On the other hand, if one changes the direction of the diagonals in the triangulation, i.e., one considers a triangulation of the type depicted in Fig. 1b, then the desired condition holds and the AFC scheme (6), (7) with any of the limiters

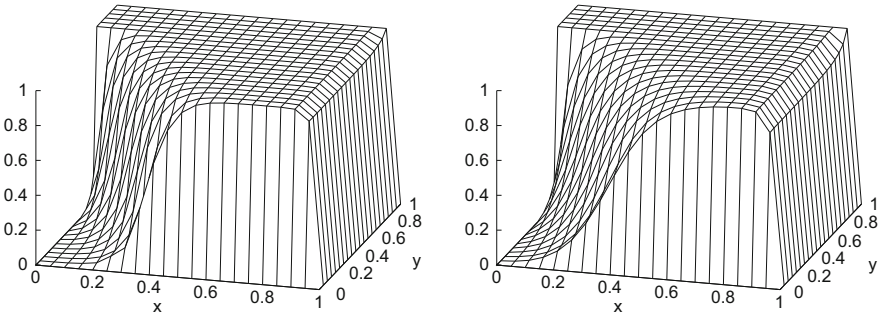

Fig. 2 Example 1: approximate solutions obtained on a triangulation of the type depicted in Fig. 1a using the AFC scheme (6), (7) with Limiter 1 (*left*) and Limiter 3 (*right*)

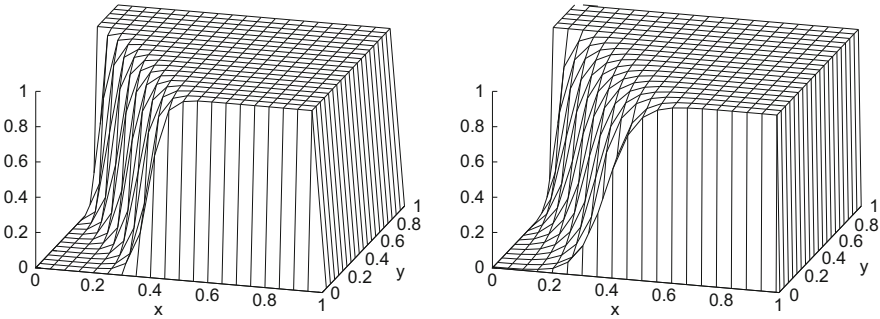

Fig. 3 Example 1: approximate solutions obtained using the AFC scheme (6), (7) with Limiter 1 on triangulations of the type depicted in Fig. 1b (*left*) and in Fig. 1c (*right*)

from Sect. 4 provides a sharp approximation of the layers at the outflow boundaries, see Fig. 3 (left) for the solution obtained with Limiter 1.

It should be stressed that the direction of $\mathbf{d}^{ij}$ may considerably differ from the direction of the edge E_{ij} and hence the suitability of a given mesh depends not only on directions of edges but also on the form of the adjacent elements. For example, instead of changing the direction of the diagonals in the triangulation from Fig. 1a, one may use anisotropic elements along the outflow boundaries, see Fig. 1c. With increasing anisotropy, the vectors $\mathbf{d}^{ij}$ corresponding to edges connecting interior and boundary vertices tend to normal vectors to the boundary and hence the discussed condition $a_{ij} \geq 0$ will be satisfied. Figure 3 (right) shows the approximate solution obtained on a triangulation of the type depicted in Fig. 1c which again consists of 800 triangles (for clarity of visualization, we use only one row of anisotropic elements along the outflow boundary; the width of the anisotropic elements in the orthogonal direction to the boundary is 0.01). Again one can observe a sharp approximation of the outflow boundary layers which are now steeper due to the use of the anisotropic elements. The smearing of the interior layer is larger than in Fig. 3 (left) because of the worse alignment of the triangulation with the convection direction.

The above discussion shows that the AFC scheme alone cannot guarantee sharp approximations of boundary layers and that the use of appropriate meshes is essential.

Acknowledgements This work has been supported through the grant No. 16-03230S of the Czech Science Foundation.

References

1. Barrenechea, G.R., John, V., Knobloch, P.: Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal. **54**(4), 2427–2451 (2016)
2. Barrenechea, G.R., John, V., Knobloch, P.: An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. Math. Models Methods Appl. Sci. **27**(3), 525–548 (2017)
3. Kuzmin, D.: Algebraic flux correction for finite element discretizations of coupled systems. In: Papadrakakis, M., Oñate, E., Schrefler, B. (eds.) Proceedings of the International Conference on Computational Methods for Coupled Problems in Science and Engineering, pp. 1–5. CIMNE, Barcelona (2007)
4. Kuzmin, D.: Explicit and implicit FEM-FCT algorithms with flux linearization. J. Comput. Phys. **228**, 2517–2534 (2009)
5. Kuzmin, D.: Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. J. Comput. Appl. Math. **236**, 2317–2337 (2012)
6. Kuzmin, D., Möller, M.: Algebraic flux correction I. Scalar conservation laws. In: Kuzmin, D., Löhner, R., Turek, S. (eds.) Flux-Corrected Transport. Principles, Algorithms, and Applications, pp. 155–206. Springer, Berlin (2005)

Energy-Norm A Posteriori Error Estimates for Singularly Perturbed Reaction-Diffusion Problems on Anisotropic Meshes: Neumann Boundary Conditions

Natalia Kopteva

Abstract Residual-type a posteriori error estimates in the energy norm are given for singularly perturbed semilinear reaction-diffusion equations posed in polygonal domains. Linear finite elements are considered on anisotropic triangulations. The error constants are independent of the diameters and the aspect ratios of mesh elements and of the small perturbation parameter. The case of the Dirichlet boundary conditions was considered in the recent article (Kopteva, Numer. Math., 2017, Published online 2 May 2017. doi:10.1007/s00211-017-0889-3). Now we extend this analysis to also allow boundary conditions of Neumann type.

1 Introduction

This paper addresses finite element approximations to singularly perturbed semilinear reaction-diffusion equations of the form

$$-\varepsilon^2\triangle u + f(x,y;u) = 0 \ \text{ for } (x,y)\in\Omega, \quad \partial_\nu u = \psi \text{ on } \Gamma_N, \quad u = 0 \text{ on } \Gamma_D, \tag{1}$$

posed in a, possibly non-Lipschitz, polygonal domain $\Omega \subset \mathbb{R}^2$. Here $0 < \varepsilon \le 1$. The boundary segments Γ_D and Γ_N are disjoint with $\bar\Gamma_D \cup \bar\Gamma_N = \partial\Omega$, and ∂_ν denotes the outward normal derivative. The function f is continuous on $\Omega\times\mathbb{R}$ and satisfies $f(\cdot;s)\in L_\infty(\Omega)$ for all $s\in\mathbb{R}$, and the one-sided Lipschitz condition $f(x,y;v) - f(x,y;w) \ge C_f[v-w]$ whenever $v \ge w$, with some constant $C_f \ge 0$ such that $C_f + \varepsilon^2 \ge 1$.

Our goal is to give residual-type a posteriori error estimates on reasonably general anisotropic meshes (such as on Fig. 1, left, and Fig. 2) in the energy norm

N. Kopteva (✉)
Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
e-mail: natalia.kopteva@ul.ie

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_11

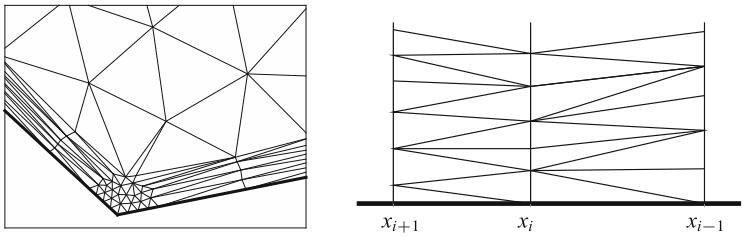

Fig. 1 Example of a mesh considered in [8, 9] (*left*), partially structured anisotropic mesh (*right*)

$|||\cdot|||_{\varepsilon;\Omega}$. The latter is an appropriately scaled $W_2^1(\Omega)$ norm naturally associated with our problem. For any $\mathscr{D} \subseteq \Omega$, it is defined by

$$|||v|||_{\varepsilon;\mathscr{D}} := \left\{\varepsilon^2 \|\nabla v\|^2_{2;\mathscr{D}} + \|v\|^2_{2;\mathscr{D}}\right\}^{1/2}.$$

The case of Dirichlet boundary conditions was considered in the recent article [9]. Now we extend this analysis to also allow boundary conditions of Neumann type. In this preliminary contribution, we shape our treatment of Neumann boundary conditions on anisotropic mesh elements in a simpler setting of partially structured meshes (as on Fig. 1, right). The presented approach will be applied to more general anisotropic meshes, such as addressed in [8, 9], in a forthcoming journal article.

It is worth noting that our estimators in this paper, as well as in [8–10], do not involve the so-called matching functions. (The latter appear in the estimator error constants in [12–14]; they depend on the unknown error and take moderate values only when the grid is either isotropic, or, being anisotropic, is aligned correctly to the solution, while, in general, may be as large as mesh aspect ratios.)

We discretize (1) using linear finite elements. Let $S_h \subset \{v \in H^1(\Omega) \cap C(\bar{\Omega}) : v = 0 \text{ on } \Gamma_D\}$ be a piecewise-linear finite element space relative to a triangulation $\mathscr{T}$, and let the computed solution $u_h \in S_h$ satisfy

$$\varepsilon^2\langle\nabla u_h, \nabla v_h\rangle + \langle f_h^I, v_h\rangle = \int_{\Gamma_N} \varepsilon^2 \psi\, v_h \quad \forall\, v_h \in S_h, \qquad f_h(\cdot) := f(\cdot; u_h). \tag{2}$$

Here $\langle\cdot,\cdot\rangle$ is the $L_2(\Omega)$ inner product, and f_h^I is the standard piecewise-linear Lagrange interpolant of f_h.

To give a flavour of our results, our first estimator reduces to

$$|||u_h - u|||_{\varepsilon;\Omega} \le C \left\{ \sum_{z\in\mathscr{N}} \min\{h_z H_z,\, \varepsilon H_z^2 h_z^{-1}\} \left\|\varepsilon J\right\|^2_{\infty;\gamma_z} + \sum_{z\in\mathscr{N}} |I_z^{\psi}|^2 \right.$$
$$\left. + \sum_{z\in\mathscr{N}} \left\|\min\{1,\, H_z\varepsilon^{-1}\} f_h^I\right\|^2_{2;\omega_z} + \left\|f_h - f_h^I\right\|^2_{2;\Omega} \right\}^{1/2}, \tag{3}$$

where C is independent of the diameters and the aspect ratios of elements in $\mathscr{T}$, and of ε. Here $\mathscr{N}$ is the set of nodes in $\mathscr{T}$, J is the standard jump in the normal derivative of u_h across an interior element edge, while $J := \partial_\nu u_h - \psi$ on Γ_N, ω_z is the patch of elements surrounding any $z \in \mathscr{N}$, γ_z is the set of edges originating at z and lying in $\omega_z \cup \Gamma_N$, $H_z = \operatorname{diam}(\omega_z)$, and $h_z \simeq H_z^{-1}|\omega_z|$. We also obtain a sharper version of (3), in which the interior-residual factors $\min\{1, H_z\varepsilon^{-1}\}$ are replaced by $\min\{1, h_z\varepsilon^{-1}\}$ and a few other terms are included (see (23) and Corollary 4.2).

The presence of Neumann boundary conditions is reflected in J computed on $\gamma_z \cap \Gamma_N$, and in additional (and, perhaps, unexpected) terms I_z^ψ:

$$|I_z^\psi|^2 \le \min\{H_z,\, \varepsilon\}\, H_z \left|\operatorname{osc}(\varepsilon\psi\,;\gamma_z \cap \Gamma_N)\right|^2.$$

In the case of shape-regular triangulations, $\min\{h_z H_z,\, \varepsilon H_z^2 h_z^{-1}\} \simeq \min\{H_z,\, \varepsilon\}\, H_z$, while $\|J\|_{\infty;\gamma_z} \ge \frac{1}{2}\operatorname{osc}(J\,;\gamma_z \cap \Gamma_N) \ge \frac{1}{2}\operatorname{osc}(\psi\,;\gamma_z \cap \Gamma_N)$. Hence, in this case, $\sum |I_z^\psi|^2$ is bounded by the first sum in (3), so may be skipped. For the case $\varepsilon = 1$, this yields a version of the standard estimator [1, §2.2].

To relate (3) to interpolation error bounds, as well as to possible adaptive-mesh construction strategies, note that $|J_z|$ may be interpreted as approximating the diameter of ω_z under the metric induced by the squared Hessian matrix of the exact solution (while f_h^I approximates $\varepsilon^2 \triangle u$).

Our interest in this paper is in general anisotropic meshes, since such meshes, when constructed a priori, have been shown to offer an efficient way of computing reliable numerical approximations of solutions that exhibit sharp boundary and interior layers (see, e.g., [2, 5, 11, 16] and references therein). In the case of shape-regular triangulations, residual-type a posteriori estimates for equations of type (1) were proved in [18] in the energy norm, and more recently in [4] in the maximum norm. The case of anisotropic meshes having a tensor-product structure was addressed in [3, 6, 17]. Above, we briefly discussed anisotropic estimators [12–14].

Note that no attempt will be made in this contribution to derive lower error bounds. For anisotropic meshes, [13] gives such a bound, which includes some of the terms appearing in our estimators. For example, in the case $h_z \le \varepsilon$, the terms in first sum of (3) reduce to $h_z H_z \|\varepsilon J\|_{\infty;\gamma_z}^2$, while the related jump residual terms in the lower error bound [13, (4.2)] can be interpreted as $\sum_{S\subset\gamma_z} h_z|S|\, \|\varepsilon J\|_{\infty;S}^2$.

The paper is organized as follows. In Sect. 2, we make basic triangulation assumptions and recall the anisotropic scaled-trace theorem from [8, 9]. The error is represented in terms of the residual in Sect. 3, while the main results are obtained in Sect. 4. We conclude the paper by presenting some numerical results in Sect. 5.

Notation We write $a \simeq b$ when $a \lesssim b$ and $a \gtrsim b$, and $a \lesssim b$ when $a \le Cb$ with a generic constant C depending on Ω and f, but C does not depend on either ε or the diameters and the aspect ratios of elements in $\mathscr{T}$. Also, for $\mathscr{D} \subset \bar{\Omega}$, $1 \le p \le \infty$, and $k \ge 0$, let $\|\cdot\|_{p;\mathscr{D}} = \|\cdot\|_{L_p(\mathscr{D})}$ and $|\cdot|_{k,p;\mathscr{D}} = |\cdot|_{W_p^k(\mathscr{D})}$, where $|\cdot|_{W_p^k(\mathscr{D})}$ is the standard Sobolev seminorm with integrability index p and smoothness index k.

2 Basic Triangulation Assumptions: Scaled Trace Bounds

We shall use $z = (x_z, y_z)$, S and T to respectively denote particular mesh nodes, edges and elements, while $\mathcal{N}$, $\mathcal{S}$ and $\mathcal{T}$ will respectively denote their sets. For each $T \in \mathcal{T}$, let H_T be the maximum edge length and $h_T := 2H_T^{-1}|T|$ be the minimum height in T. For each $z \in \mathcal{N}$, let ω_z be the patch of elements surrounding any $z \in \mathcal{N}$, $\mathcal{S}_z$ the set of edges originating at z, and

$$H_z := \operatorname{diam}(\omega_z), \quad h_z := \max_{T \subset \omega_z} h_T, \quad \gamma_z := \mathcal{S}_z \setminus \Gamma_D, \quad \mathring{\gamma}_z := \{S \subset \gamma_z : |S| \lesssim h_z\}. \tag{4}$$

Throughout the paper we make the following Triangulation Assumptions (that are automatically satisfied by shape-regular triangulations).

- *Maximum Angle condition.* Let the maximum interior angle in any triangle $T \in \mathcal{T}$ be uniformly bounded by some positive $\alpha_0 < \pi$.
- *Local Element Orientation condition.* For any $z \in \mathcal{N}$, a minimal rectangle $R_z \supset \omega_z$ is such that $|R_z| \simeq |\omega_z|$.
- Also, let the number of triangles containing any node be uniformly bounded.

Our analysis in [8, 9] applies to three node types, which we call (i) anisotropic, (ii) semi-anisotropic, and (iii) isotropic nodes (see Fig. 2). As in this preliminary contribution, only type (i) is considered, we skip the definitions (ii) and (iii).

(i) *Anisotropic Nodes*, whose set is denoted by $\mathcal{N}_{\rm ani}$, are such that

$$h_z < c_0 H_z, \qquad h_T \simeq h_z \ \text{ and } \ H_T \simeq H_z \quad \forall\, T \subset \omega_z, \tag{5}$$

where c_0 is a fixed small constant.

(i*) One typically expects anisotropic elements near Γ_D to be aligned along it. The boundary nodes for which this is not the case form a special set:

$$\mathcal{N}_D^* := \big\{\, z \in \bar{\Gamma}_D \cap \mathcal{N}_{\rm ani} \ : \ |\mathcal{S}_z \cap \Gamma_D| \lesssim h_z \ \text{ or } \ z \notin \bar{\Gamma}_N \cap \bar{\Gamma}_D \text{ is a corner of } \Omega \,\big\}. \tag{6}$$

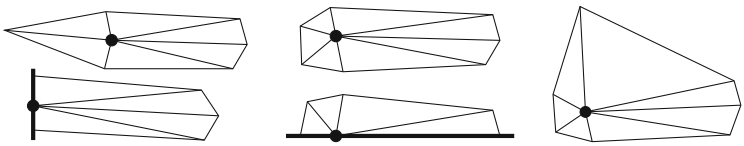

Fig. 2 Examples of anisotropic nodes $z \in \mathcal{N}_{\rm ani}$ (*left*), semi-anisotropic nodes $z \in \mathcal{N}_{\rm s.ani}$ (*centre*), an isotropic node $z \in \mathcal{N}_{\rm iso}$ (*right*), and a node $z \in \mathcal{N}_{\rm ani} \cap \mathcal{N}_D^*$ (*bottom left*); see [8, 9]

Next, we recall a version of the scaled trace theorem for possibly anisotropic nodes using, with $p = 1, 2$, the scaled $W_p^1(\mathscr{D})$ norm

$$|||v|||_{p;\mathscr{D}} := (\mathrm{diam}\mathscr{D})^{-1}\|v\|_{p;\mathscr{D}} + \|\nabla v\|_{p;\mathscr{D}}\,.$$

In particular, in view of $\mathrm{diam}(\omega_z) = H_z$ and $\mathrm{diam}(T) \simeq H_T$,

$$|||v|||_{p;\omega_z} = H_z^{-1}\|v\|_{p;\omega_z} + \|\nabla v\|_{p;\omega_z}\,, \quad |||v|||_{p;T} \simeq H_T^{-1}\|v\|_{p;T} + \|\nabla v\|_{p;T}\,. \tag{7}$$

Lemma 2.1 (Anisotropic Scaled Trace bounds [8, 9]) *For any node $z \in \mathscr{N}$ of type (5), and any function $v \in W_1^1(\omega_z)$, one has*

$$\|v\|_{1;\mathring{\gamma}_z} + \frac{h_z}{H_z}\|v\|_{1;\gamma_z\setminus\mathring{\gamma}_z} + \frac{h_z}{H_z}\|v\|_{1;\bar{S}_z} \lesssim |||v|||_{1;\omega_z}\,, \tag{8}$$

$$\|v\|_{1;\mathring{\gamma}_z} + \frac{h_z}{H_z}\|v\|_{1;\gamma_z\setminus\mathring{\gamma}_z} + \frac{h_z}{H_z}\|v\|_{1;\bar{S}_z} \lesssim \left\{h_z\,\|v\|_{2;\omega_z}\,|||v|||_{2;\omega_z}\right\}^{1/2}, \tag{9}$$

where γ_z and $\mathring{\gamma}_z$ are from (4), while $\bar{S}_z \subset \omega_z$ is any segment that originates at z and satisfies $|\bar{S}_z| \simeq H_z$.

3 Representation of the Error in Terms of the Residual

Using the monotonicity of f and $C_f + \varepsilon^2 \geq 1$, one gets

$$\begin{aligned}|||u_h - u|||_{\varepsilon;\Omega}^2 &\lesssim \varepsilon^2\langle\nabla(u_h - u), \nabla(u_h - u)\rangle + \langle f(\cdot; u_h) - f(\cdot; u), u_h - u\rangle\\ &= \varepsilon^2\langle\nabla u_h, \nabla(u_h - u)\rangle + \langle f(\cdot; u_h), u_h - u\rangle - \int_{\Gamma_N}\varepsilon^2\psi\,(u_h - u),\end{aligned}$$

where we also used (1). Next, assuming $|||u_h - u|||_{\varepsilon;\Omega} > 0$, let

$$G := \frac{u_h - u}{|||u_h - u|||_{\varepsilon;\Omega}} \qquad \Rightarrow \qquad |||G|||_{\varepsilon;\Omega} = 1. \tag{10}$$

So $|||u_h - u|||_{\varepsilon;\Omega} \lesssim \varepsilon^2\langle\nabla u_h, \nabla G\rangle + \langle f(\cdot; u_h), G\rangle - \int_{\Gamma_N}\varepsilon^2\psi G$. So (2) implies, $\forall v_h \in S_h$,

$$|||u_h - u|||_{\varepsilon;\Omega} \lesssim \varepsilon^2\langle\nabla u_h, \nabla(G - v_h)\rangle + \langle f_h^I, G - v_h\rangle - \int_{\Gamma_N}\varepsilon^2\psi\,(G - v_h) + \langle f_h - f_h^I, G\rangle\,. \tag{11}$$

Here $\langle f_h - f_h^I, G\rangle =: \mathscr{E}_{\rm quad}$ is the quadrature error, for which $\|G\|_{2;\Omega} \le 1$ implies

$$|\mathscr{E}_{\rm quad}| \le |\!|\!| f_h - f_h^I |\!|\!|^*_{\varepsilon;\Omega} \le \|f_h - f_h^I\|_{2;\Omega}\,, \tag{12}$$

where the norm $|\!|\!|\cdot|\!|\!|^*_{\varepsilon;\Omega}$ is dual to $|\!|\!|\cdot|\!|\!|_{\varepsilon;\Omega}$; see also [9, Remark 4.1].

Next, let ϕ_z be the standard linear hat function corresponding to $z \in \mathscr{N}$, and $v_h := G_h + \sum_{z\in\mathscr{N}} \bar g_z \phi_z \in S_h$, where $G_h \in S_h$ is some interpolant of G, while $\bar g_z$ is a certain average of $G - G_h$ near z (to be specified later), but $\bar g_z = 0$ for $z \in \Gamma_D$ (so that $v_h \in S_h$). Now, using $g := G - G_h$, one gets $G - v_h = g - \sum_{z\in\mathscr{N}} \bar g_z \phi_z = \sum_{z\in\mathscr{N}} (g - \bar g_z)\phi_z$. Combining this with (11) gives a standard error representation

$$\begin{aligned} |\!|\!| u_h - u |\!|\!|_{\varepsilon;\Omega} &\lesssim \sum_{z\in\mathscr{N}} \varepsilon^2 \int_{\gamma_z} J(g - \bar g_z)\phi_z + \sum_{z\in\mathscr{N}} \int_{\omega_z} f_h^I\,(g - \bar g_z)\phi_z + \mathscr{E}_{\rm quad} \\ &=: I + I\!I + \mathscr{E}_{\rm quad}\,, \end{aligned} \tag{13}$$

which holds for any $G_h \in S_h$ and any $\{\bar g_z\}_{z\in\mathscr{N}}$ such that $\bar g_z = 0$ whenever $z \in \Gamma_D$. Here we use a standard definition for J with $J := \partial_\nu u_h\big|_{T'} + \partial_\nu u_h\big|_{T''}$ on an interior edge $\partial T' \cap \partial T'' \ne \emptyset$ (where $T', T'' \in \mathscr{T}$), and $J := \partial_\nu u_h - \psi$ on Γ_N.

4 Error Analysis for a Partially Structured Anisotropic Mesh

Our ultimate goal is to consider a reasonably general anisotropic mesh such as addressed in [8, 9] (see Fig. 1, left, and Fig. 2). But in this preliminary contribution, to illustrate our approach, we restrict the analysis to a simpler, partially structured, anisotropic mesh in a square domain. To be more precise, let

$$\Omega := (0,1)^2, \qquad \Gamma_N := \{(x,y) \in \partial\Omega :\ x = 1 \text{ or } y = 1\}, \qquad \psi(0,1) = 0. \tag{14}$$

(The condition on ψ is a compatibility condition, as $u(0,y) = 0$ implies $\partial_y u(0,1) = 0$. If it is violated, the mesh node at $(0,1)$ is expected to be isotropic, and a version of our analysis below will apply.) The following triangulation assumptions are made.

A1. Let $\{x_i\}_{i=0}^n$ be an arbitrary mesh on the interval $(0,1)$ in the x direction. Then, let each $T \in \mathscr{T}$, for some i,

   1. have the shortest edge on the line $x = x_i$;
   2. have a vertex on the line $x = x_{i+1}$ or $x = x_{i-1}$ (see Fig. 1, right).

A2. Let $\mathscr{N} = \mathscr{N}_{\rm ani}$, i.e. each mesh node z satisfies (5).

A3. *Quasi-non-obtuse anisotropic elements.* Let the maximum angle in any triangle be bounded by $\frac{\pi}{2} + \alpha_1 \frac{h_T}{H_T}$ for some positive constant α_1.

These conditions essentially imply that all mesh elements are anisotropic and aligned in the x-direction. They also imply that if $x_z = x_i$, then

$$\omega_z \subseteq \omega_z^* := (x_{i-1}, x_{i+1}) \times (y_z^-, y_z^+), \quad y_z^+ - y_z^- \simeq h_z, \quad \operatorname{diam} \omega_z^* \simeq H_z\,, \tag{15}$$

where (y_z^-, y_z^+) is the range of y within ω_z, and we also use $x_{-1} := x_0$ and $x_{n+1} := x_n$.

Remark 4.1 The above conditions (in particular, A3) imply that there is $J \lesssim 1$ such that $\omega_z^* \subset \omega_z^{(J)}$ for all $z \in \mathcal{N}$, with the notation $\omega_z^{(0)} := \omega_z$ and $\omega_z^{(j+1)}$ for the patch of elements in/touching $\omega_z^{(j)}$. (Note that $J = 1$ for any non-obtuse triangulation.)

4.1 Choice of $\bar{g}_z$: Main Results

Following [8, 9], the choice of $\bar{g}_z$ in (13) is related to the orientation of anisotropic elements, and is crucial. Let $\bar{g}_z = 0$ for $z \in \Gamma_D$, and

$$\int_{x_{i-1}}^{x_{i+1}} (g(x, y_z) - \bar{g}_z)\, \varphi_i(x)\, dx = 0 \qquad \text{for } z = (x_i, y_z), \quad 1 \le i \le n. \tag{16}$$

Here we use the standard one-dimensional hat function $\varphi_i(x)$ associated with the mesh $\{x_i\}$ (i.e. it has support on (x_{i-1}, x_{i+1}), equals 1 at $x = x_i$, and is linear on (x_{i-1}, x_i) and (x_i, x_{i+1})). Note that for $z = (x_i, 0)$, in view of $g = 0$ on Γ_D, the above definition (16) agrees with $\bar{g}_z = 0$, earlier prescribed on Γ_D.

Remark 4.2 An inspection of standard proofs for shape-regular meshes reveals that one obstacle in extending them to anisotropic meshes lies in the application of a scaled traced theorem when estimating the jump residual terms (this causes the mesh aspect ratios to appear in the estimator). This technical difficulty is addressed by choosing $\bar{g}_z$ as a certain one-dimensional average of g, as in (16), or in (17) below. To relate this to standard choices, for $x_z = x_i$, let $\bar{S}_z \subset \omega_z^*$ be the interval joining (x_{i-1}, y_z) and (x_{i+1}, y_z), $1 \le i \le n$. Then (16) is identical to $\int_{\bar{S}_z} (g - \bar{g}_z)\, \varphi_i = 0$. Also, for non-obtuse triangulations, it is equivalent to $\int_{\bar{S}_z} (g - \bar{g}_z)\, \phi_z = 0$. The reader may compare this with a more standard choice, denoted here by $\bar{g}_z'$: $\int_{\omega_z} (g - \bar{g}_z')\, \phi_z = 0$ (see, e.g., [15, Lecture 5]).

Remark 4.3 It is sometimes helpful to tweak the definition (16) of $\{\bar{g}_z\}_{z \in \mathcal{N}}$ and use instead $\{\bar{g}_z^*\}_{z \in \mathcal{N}}$ defined for $z \in \mathcal{N} \setminus \Gamma_D$ with $x_z = x_i$ by

$$\int_{\omega_z^*} [g(x, y) - \bar{g}_z^*]\, \varphi_i(x) = 0, \tag{17}$$

(where ω_z^* is from (15)), and $\bar g_z^* = 0$ for $z \in \Gamma_D$. Note that

$$h_z H_z |\bar g_z^*| \lesssim \|g\|_{1;\omega_z^*}, \quad H_z |\bar g_z - \bar g_z^*| \lesssim \|\nabla g\|_{1;\omega_z^*}, \quad |\omega_z^*| \simeq h_z H_z. \tag{18}$$

Here the first relation is obvious, while $\int_{\omega_z^*} [g(x,y_z) - g(x,y)]\,\varphi_i(x) \simeq h_z H_z (\bar g_z - \bar g_z^*)$ implies $H_z|\bar g_z - \bar g_z^*| \lesssim \|\partial_y g\|_{1;\omega_z^*}$ and so the second relation.

Theorem 4.1 *For the solution u of (1), (14), and the computed solution u_h of (2), let $g = G - G_h$ with G from (10) and any $G_h \in S_h$, and*

$$\Theta := \varepsilon^2 \|\nabla g\|^2_{2;\Omega} + \sum_{z\in\mathcal{N}} \big(1 + \varepsilon^2 H_z^{-2}\big) \|g\|^2_{2;\omega_z}. \tag{19}$$

Then $\|\!|u_h - u|\!\|_{\varepsilon;\Omega} \lesssim I + I\!I + \mathcal{E}_{\rm quad}$, where $\mathcal{E}_{\rm quad}$ is bounded by (12), and, under conditions A1–A3,

$$|I + I^\psi| \lesssim \Big\{\Theta \sum_{z\in\mathcal{N}} \lambda_z \big\|\varepsilon J_z\big\|^2_{\infty;\gamma_z}\Big\}^{1/2}, \qquad \lambda_z := h_z H_z \min\{1,\, \varepsilon H_z h_z^{-2}\}, \tag{20}$$

$$|I^\psi| \lesssim \Big\{\Theta \sum_{\substack{z\in\mathcal{N}:\\ |\gamma_z\cap\Gamma_N|\simeq H_z}} \lambda_z' \,\varepsilon H_z \big|\operatorname{osc}(\varepsilon\psi\,;\gamma_z\cap\Gamma_N)\big|^2\Big\}^{1/2}, \quad \lambda_z' := \min\{1,\, H_z\varepsilon^{-1}\}, \tag{21}$$

$$|I\!I| \lesssim \Big\{\Theta \sum_{z\in\mathcal{N}} \big\|\lambda_z' f_h^I\big\|^2_{2;\omega_z}\Big\}^{1/2}. \tag{22}$$

Additionally, one has an alternative bound

$$|I\!I| \lesssim \Big\{\Theta \sum_{z\in\mathcal{N}\setminus\mathcal{N}_D^*} \big\|\min\{1,\, h_z\varepsilon^{-1}\} f_h^I\big\|^2_{2;\omega_z} + \Theta \sum_{z\in\mathcal{N}\setminus\mathcal{N}_D^*} \big\|\lambda_z' \operatorname{osc}(f_h^I\,;\omega_z)\big\|^2_{2;\omega_z} + \Theta \sum_{z\in\mathcal{N}_D^*} \big\|\lambda_z' f_h^I\big\|^2_{2;\omega_z}\Big\}^{1/2}, \tag{23}$$

where $\mathcal{N}_D^* = \{z \in \mathcal{N} : x_z = 0\}$ (in agreement with (6)).

Corollary 4.2 (A Posteriori Error Estimator) *Under the conditions of Theorem 4.1, $\|\!|u_h - u|\!\|_{\varepsilon;\Omega} \lesssim I + I\!I + \mathcal{E}_{\rm quad}$, where $\mathcal{E}_{\rm quad}$ is bounded by (12), while for I and $I\!I$ one has bounds (20)–(23) with $\Theta := 1$.*

Proof Under more general conditions than A1–A3, there exists $G_h \in S_h$ such that $\Theta \lesssim \|\!|G|\!\|_{\varepsilon;\Omega} = 1$; see [9, Theorem 7.4]. □

Remark 4.4 An inspection of the proof shows that in the bound (21) for I^{ψ}, one can replace $\gamma_z \cap \Gamma_N$ by $[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma_N$, which gives a slightly sharper bound.

Proof of Theorem 4.1 We partially follow and invoke some auxiliary results from the proof of [9, Theorem 5.1]. The proof of (20) and (21) is given in Sect. 4.2 below. Note that we cannot simply focus on the new terms, denoted by I_z^N in (25), as one needs to look into a delicate interaction of a component of I_z^N and some other terms in I (see (26), (27), (29)).

For the remaining interior-residual bounds (22) and (23), an inspection of [9, Section 5.3] shows that the estimation of the interior-residual component $I\!I$ of the error (13) applies to our case, with the only change in that $I\!I$ involves $\sum_{i=1}^{n} I\!I_i$ (rather than $\sum_{i=1}^{n-1} I\!I_i$), where $I\!I_i := \sum_{z \in \mathscr{N}_i} \int_{\omega_z} f_h^I(x_i, y)\,(g - \bar{g}_z^*)\,\phi_z$ is defined in [9], with $\mathscr{N}_i := \{z \in \mathscr{N} : x_z = x_i\}$, while $\mathscr{N}_{\partial\Omega}^*$ of [9] is now denoted $\mathscr{N}_D^* = \mathscr{N}_0$. Note also that (17) and (18), as well as Remark 4.5 below, are crucial for (22) and (23). □

4.2 Jump Residual: Proof of (20) and (21)

Proof of (20) and (21) Split I of (13) as $I = \sum_{z \in \mathscr{N}} I_z$, where

$$I_z := \varepsilon^2 \int_{\gamma_z} J(g - \bar{g}_z)\phi_z. \tag{24}$$

When considering J on $\gamma_z \backslash \partial\Omega = \gamma_z \backslash \Gamma_N$, we adapt the notational convention that the unit normal ν to any edge in γ_z takes the clockwise direction about z, while $[\![w]\!]$, for any w, is the jump in w across any edge in γ_z evaluated in the anticlockwise direction about z. Then

$$J\big|_{\gamma_z \backslash \Gamma_N} = [\![\nabla u_h]\!] \cdot \nu = [\![\partial_x u_h]\!] \nu_x + [\![\partial_y u_h]\!] \nu_y\,.$$

So I_z can be split as

$$\begin{aligned} I_z = I_z' + I_z'' + I_z''' + I_z^N := {} & \varepsilon^2 \int_{\gamma_z \backslash \Gamma^N} (g - \bar{g}_z)\phi_z [\![\partial_x u_h]\!]\, \nu_x + \varepsilon^2 \int_{\mathring{\gamma}_z \cap \Gamma^N} J(g - \bar{g}_z)\phi_z \\ & + \varepsilon^2 \int_{\gamma_z \backslash \Gamma^N} [g - g(x, y_z)]\, \phi_z [\![\partial_y u_h]\!]\, \nu_y \\ & + \varepsilon^2 \int_{\gamma_z \backslash \Gamma^N} [g(x, y_z) - \bar{g}_z]\, \phi_z [\![\partial_y u_h]\!]\, \nu_y \\ & + \varepsilon^2 \int_{[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma^N} (\partial_\nu u_h - \psi)(g - \bar{g}_z)\phi_z\,. \end{aligned} \tag{25}$$

For the final term here one has, using $\partial_\nu u_h = \partial_y u_h$ on $[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma^N \subset \partial\Omega \cap \{y = 1\}$,

$$I_z^N = \varepsilon^2 \int_{[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma^N} \partial_y u_h (g - \bar{g}_z)\phi_z - \underbrace{\varepsilon^2 \int_{[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma^N} \psi (g - \bar{g}_z)\phi_z}_{=: I_z^\psi} . \tag{26}$$

We claim that to get the desired assertions (20) and (21), it suffices to show that

$$|I_z'| + |I_z''| \lesssim \varepsilon \,|||g|||_{1;\omega_z^*} \left\| \varepsilon J \right\|_{\infty;\gamma_z}, \qquad I_z''' + (I_z^N + I_z^\psi) = 0, \tag{27}$$

$$|I_z| \lesssim \varepsilon \frac{H_z}{h_z} \left\{ h_z \|g\|_{2;\omega_z^*} \, |||g|||_{2;\omega_z^*} \right\}^{1/2} \left\| \varepsilon J \right\|_{\infty;\gamma_z}, \tag{28}$$

and

$$|I_z^\psi| \lesssim \varepsilon \left\{ H_z \|g\|_{2;\omega_z^\psi} \, |||g|||_{2;\omega_z^\psi} \right\}^{1/2} \operatorname{osc}(\varepsilon\psi ; [\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma_N). \tag{29}$$

In (29),

$$\omega_z^\psi := (x_{i-1}, x_{i+1}) \times (1 - H_z, 1) \qquad \text{for any } \; z = (x_i, 1), \quad i = 1, \dots, n, \tag{30}$$

is an isotropic rectangle with the upper edge $[\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma_N$ (a similar triangle can be used instead).

To show that (20) and (21), indeed, follow from (27)–(30), let

$$\theta_z := \lambda_z^{-1} \varepsilon^2 \, \min \left\{ |||g|||_{1;\omega_z^*}^2 \, , \, H_z^2 \, h_z^{-1} \|g\|_{2;\omega_z^*} \, |||g|||_{2;\omega_z^*} \right\},$$

$$\theta_z^\psi := \lambda_z'^{-1} \varepsilon \|g\|_{2;\omega_z^\psi} \, |||g|||_{2;\omega_z^\psi} \, .$$

Note that an application of $\min(a, bc)/\min(1, c) \le a + b$ (for any a, b, $c > 0$) implies $\theta_z \lesssim \varepsilon^2 \, |||g|||_{2;\omega_z^*}^2 + \varepsilon \, \|g\|_{2;\omega_z^*} \, |||g|||_{2;\omega_z^*}$, while $\lambda_z'^{-1} \simeq 1 + \varepsilon H_z^{-1}$ yields $\theta_z^\psi \lesssim \varepsilon^2 |||g|||_{2;\omega_z^\psi}^2 + (1 + \varepsilon^2 H_z^{-2}) \|g\|_{2;\omega_z^\psi}^2$. Combining these two observations with (7), (19) and Remark 4.1 yields $\sum_{z \in \mathcal{N}} (\theta_z + \theta_z^\psi) \lesssim \Theta$.

Next, combining (27)–(30) with the above definitions of θ_z and θ_z^ψ, one gets

$$\min\big\{ |I_z + I_z^\psi|, \, |I_z| \big\} \lesssim (\theta_z \lambda_z)^{1/2} \, \|\varepsilon J\|_{\infty;\gamma_z} \, ,$$

$$|I_z^\psi| \lesssim (\theta_z^\psi \lambda_z' \varepsilon H_z)^{1/2} \operatorname{osc}(\varepsilon\psi ; [\gamma_z \backslash \mathring{\gamma}_z] \cap \Gamma_N).$$

Now, an application of Hölder's inequality shows that $\sum_{z\in\mathcal{N}} \min\{|I_z + I_z^\psi|, |I_z|\}$ is bounded by the right-hand side of (20), and $\sum_{z\in\mathcal{N}} |I_z^\psi|$ is bounded by the right-hand side of (21).

Finally, set $\widetilde{I}_z^\psi := 0$ if $\min\{|I_z + I_z^\psi|, |I_z|\} = |I_z|$, and $\widetilde{I}_z^\psi := I_z^\psi$ otherwise, so that one always has $|I_z + \widetilde{I}_z^\psi| = \min\{|I_z + I_z^\psi|, |I_z|\}$. The desired assertions (20) and (21) follow with $I^\psi := \sum_{z\in\mathcal{N}} \widetilde{I}_z^\psi$.

Hence, it remains to establish (27), (28) and (29). The bounds for I_z' and I_z'' in (27), as well as I_z in (28), can be found in [9, Section 5.2, see (5.12), (5.13)]; they are obtained from (25) and (24) using (8) and (9) respectively. It should be noted that, compared to [9], there is an additional term in I_z', which involves $\int_{\mathring{\gamma}_z\cap\Gamma^N}$ and can be easily estimated again using (8).

The proof of $I_z''' + (I_z^N + I_z^\psi) = 0$ in (27) is more delicate. It is convenient to adapt the convention that $u_h = 0$ in $\mathbb{R}^2 \setminus \bar{\Omega}$ when computing $[\![\partial_y u_h]\!]$ across the boundary edges. With this convention, one can show, for each $z = (x_i, y_z)$, that

$$I_z''' + (I_z^N + I_z^\psi) = \varepsilon^2 \Big(\sum_{S\in\gamma_z\setminus\mathring{\gamma}_z} [\![\partial_y u_h]\!] \Big) \int_{x_{i-1}}^{x_i} [g(x, y_z) - \bar{g}_z]\, \varphi_i(x)\, dx\,, \tag{31}$$

with $\int_{x_{i-1}}^{x_i}$, in the case of $i = 0$, replaced by $-\int_{x_i}^{x_{i+1}}$ (see [8, 9] for similar representations of I_z'''). First, consider the case $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N = \emptyset$. Then $I_z''' + (I_z^N + I_z^\psi) = I_z'''$, while in the definition of I_z''', one has $\nu_y = 0$ on $\mathring{\gamma}_z$ and $\phi_z = \varphi_i(x)$ on $\gamma_z\setminus\mathring{\gamma}_z$. In the latter case, we integrate the function $[g(x, y_z) - \bar{g}_z]\phi_z = [g(x, y_z) - \bar{g}_z]\varphi_i(x)$ of one variable x, which appears in the definition (16) of $\bar{g}_z$. Furthermore, $\nu_y ds = dx$ on any edge connecting z to the vertical line $\{x = x_{i-1}\}$ and $\nu_y ds = -dx$ on any edge connecting z to the vertical line $\{x = x_{i+1}\}$. Rewriting the integrals over such edges as integrals with respect to x over (x_{i-1}, x_i) and (x_i, x_{i+1}), respectively, and then employing (16) for the integrals over (x_i, x_{i+1}), one arrives at (31) for the case $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N = \emptyset$.

If $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N \neq \emptyset$, we additionally need to consider the integrals in $I_z^N + I_z^\psi$ (see (26)) of the same function $[g(x, y_z) - \bar{g}_z]\phi_z = [g(x, y_z) - \bar{g}_z]\varphi_i(x)$ over the edges in $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N$. Note that in these integrals, $ds = dx$, while $\partial_y u_h = [\![\partial_y u_h]\!]$ on any edge in $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N$ connecting z to the vertical line $\{x = x_{i-1}\}$, and $\partial_y u_h = -[\![\partial_y u_h]\!]$ on any edge in $[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N$ connecting z to the vertical line $\{x = x_{i+1}\}$. For the latter, we again employ (16) so that all integrals in $I_z^N + I_z^\psi$ are rewritten as integrals over (x_{i-1}, x_i) with respect to x. This again yields (31). So this relation is proved.

Whenever $i = n$ in (31), one immediately gets $I_z''' + (I_z^N + I_z^\psi) = 0$ from (16). Otherwise, if $0 \le i \le n-1$ and $y_z > 0$, noting that $[\![\partial_y u_h]\!] = 0$ on $\mathring{\gamma}_z$, as well as on any element edge lying on $\{x = 0\}$, one gets $\sum_{S\in\gamma_z\setminus\mathring{\gamma}_z} [\![\partial_y u_h]\!] = \sum_{S\in\mathscr{S}_z} [\![\partial_y u_h]\!] = 0$, so again $I_z''' + (I_z^N + I_z^\psi) = 0$ immediately follows from (16). Finally, if $y_z = 0$, one employs $g(x, y_z) = \bar{g}_z = 0$.

We now proceed to getting (29). Note that in the definition of I_z^ψ in (26) one has $\phi_z = \varphi_i(x)$ and $\int_{[\gamma_z\setminus\mathring{\gamma}_z]\cap\Gamma^N} = \int_{x_{i-1}}^{x_{i+1}} dx$. Now, if $z = (x_i, 1)$ for $1 \le i \le n$, recalling

(16), one can replace ψ in I_z^ψ by $\psi - \psi(z)$, so

$$|I_z^\psi| \lesssim \varepsilon \left\{ \int_{[\gamma_z \setminus \mathring{\gamma}_z] \cap \Gamma^N} |g| \right\} \operatorname{osc}(\varepsilon\psi\,; [\gamma_z \setminus \mathring{\gamma}_z] \cap \Gamma_N).$$

This yields (29) by an application of (9), in which ω_z is replaced by any isotropic domain ω_z^ψ of type (30), and hence h_z in (9) is also replaced by H_z. The remaining case of $z = (0,1)$ is considered similarly, only using $\bar{g}_z = 0$ and $|\psi| \le \operatorname{osc}(\varepsilon\psi\,; [\gamma_z \setminus \mathring{\gamma}_z] \cap \Gamma_N)$ (the latter follows from the final condition in (14)). This completes the proof of (27), (28) and (29), and hence of (20) and (21). □

Remark 4.5 The above proof remains valid if $\{\bar{g}_z\}_{z\in\mathcal{N}}$ defined by (16) are replaced by $\{\bar{g}_z^*\}_{z\in\mathcal{N}}$ from (17). Indeed, I_z will include an additional component $I_z^* := \varepsilon^2 \int_{\gamma_z} J(\bar{g}_z - \bar{g}_z^*)$, for which one easily gets $|I_z^*| \le \varepsilon H_z |\bar{g}_z - \bar{g}_z^*| \, \|\varepsilon J\|_{\infty;\gamma_z}$. For $|I_z^*|$, bounds of type (27) and (28) are then obtained using (18); see [9, Remark 5.6].

5 Numerical Results

We test the estimator of Theorem 4.1, using a simple version of (1) with $\Omega = (0,1)^2$, $\Gamma_N = \{(x,y) \in \partial\Omega : x = 0 \text{ or } y = 0\}$, and $f = u - F(x,y)$, where F is such that the unique exact solution $u = 4y(1-y)[\cos(\pi x/2) - (e^{-x/\varepsilon} - e^{-1/\varepsilon})/(1-e^{-x/\varepsilon})]$ (the latter exhibits a sharp boundary layer at $x = 0$). An example of anisotropic mesh refinement using similar estimators is given in [8, Section 7.7]. Here, we only consider one a-priori-chosen layer-adapted mesh, which is obtained by drawing diagonals from the tensor product of the Bakhvalov grid $\{\chi(\frac{i}{N})\}_{i=1}^N$ in the x-direction [2] and a uniform grid $\{\frac{j}{M}\}_{j=0}^M$ in the y-direction with $M = \frac{1}{2}N$ (see also [9, Fig. 3 (right)], and also [7]). The continuous mesh-generating function $\chi(t) = t$ if $\varepsilon > \frac{1}{6}$; otherwise, $\chi(t) = 3\varepsilon \ln \frac{1}{1-2t}$ for $t \in (0, \frac{1}{2} - 3\varepsilon)$ and is linear elsewhere subject to $\chi(1) = 1$.

Theorem 4.1 and Corollary 4.2 give the error estimator $\|\!|\!| u_h - u |\!|\!|_{\varepsilon;\Omega} \lesssim \mathscr{E}$, where $\mathscr{E} := \{\mathscr{E}_{(20)}^2 + \mathscr{E}_{(21)}^2 + \mathscr{E}_{(23)}^2 + \mathscr{E}_{(12)}^2\}^{1/2}$, with the notation $\mathscr{E}_{(\cdot)}$ for the right-hand side of $(\cdot)$ (e.g., $\mathscr{E}_{(12)} = \|f_h - f_h^I\|_{2;\Omega}$). By Corollary 4.2, all Θ-factors are set equal to 1. When computing the estimators, we replaced H_z from (4) by $\max_{T\subset\omega_z} H_T \simeq H_z$, and quantities of type $\min\{1, a\,\varepsilon^{-1}\}$ by their smoother analogues $\frac{a}{\varepsilon+a}$ (e.g., λ_z' was replaced by $\frac{H_z}{\varepsilon+H_z}$). We also replaced f_h and u by their quadratic Lagrange interpolants.

The effectivity indices in Table 1, computed as the ratio of the estimator $\mathscr{E}$ to the error $\|\!|\!| u_h - u |\!|\!|_{\varepsilon;\Omega}$, do not exceed 7.17. Table 1 also displays the ratios of the new component $\mathscr{E}_{(21)}$ in the jump residual estimator to its more standard part $\mathscr{E}_{(20)}$; they remain between 0.18 and 1.76. Note also that for the experiments of Table 1, the ratio of $\{\mathscr{E}_{(20)}^2 + \mathscr{E}_{(21)}^2 + \mathscr{E}_{(23)}^2\}^{1/2}$ to the error component $\{\varepsilon^2 \|\nabla u_h - (\nabla u)^I\|_{2;\Omega}^2 + \|u_h - u^I\|_{2;\Omega}^2\}^{1/2}$ does not exceed 7.76.

Table 1 Errors, estimators, their effectivity indices, and ratios of $\mathscr{E}_{(21)}$ to $\mathscr{E}_{(20)}$

N	$\varepsilon = 1$	$\varepsilon = 2^{-5}$	$\varepsilon = 2^{-10}$	$\varepsilon = 2^{-15}$	$\varepsilon = 2^{-20}$	$\varepsilon = 2^{-25}$	$\varepsilon = 2^{-30}$
	Errors $\|\|\|u_h - u\|\|\|_{\varepsilon;\Omega}$						
64	3.19e−2	4.99e−3	1.02e−3	6.71e−4	6.58e−4	6.57e−4	6.57e−4
128	1.60e−2	2.53e−3	4.26e−4	1.78e−4	1.64e−4	1.64e−4	1.64e−4
256	8.01e−3	1.28e−3	2.02e−4	5.36e−5	4.13e−5	4.08e−5	4.08e−5
512	4.01e−3	6.43e−4	9.96e−5	2.02e−5	1.06e−5	1.02e−5	1.02e−5
	Estimators $\mathscr{E}$						
64	1.12e−1	2.70e−2	5.95e−3	1.25e−3	6.83e−4	6.58e−4	6.57e−4
128	5.37e−2	1.26e−2	2.95e−3	5.60e−4	1.89e−4	1.64e−4	1.64e−4
256	2.62e−2	5.92e−3	1.45e−3	2.72e−4	6.27e−5	4.17e−5	4.08e−5
512	1.29e−2	2.83e−3	6.94e−4	1.35e−4	2.60e−5	1.10e−5	1.02e−5
	Effectivity Indices $\mathscr{E}/\|\|\|u_h - u\|\|\|_{\varepsilon;\Omega}$						
64	3.516	5.398	5.844	1.858	1.039	1.001	1.000
128	3.355	4.980	6.935	3.153	1.152	1.005	1.000
256	3.268	4.633	7.171	5.075	1.520	1.021	1.001
512	3.223	4.403	6.965	6.691	2.439	1.081	1.003
	Ratios $\mathscr{E}_{(21)}/\mathscr{E}_{(20)}$						
64	0.49	1.11	1.69	1.72	1.72	1.72	1.72
128	0.35	0.87	1.69	1.74	1.74	1.74	1.74
256	0.25	0.65	1.65	1.75	1.75	1.75	1.75
512	0.18	0.48	1.56	1.75	1.76	1.76	1.76

For the considered ranges of ε and N, the aspect ratios of the mesh elements take values between 2 and 3.6e+8. Considering these variations, the estimator $\mathscr{E}$ performs quite well and its effectivity indices stabilize as $\varepsilon \to 0$. A more comprehensive numerical study of the proposed estimators certainly needs to be conducted, and will be presented elsewhere.

References

1. Ainsworth, M., Oden, J.T.: A Posteriori Error Estimation in Finite Element Analysis. Wiley-Interscience, New York (2000)
2. Bakhvalov, N.S.: On the optimization of methods for solving boundary value problems with boundary layers. Zh. Vychisl. Mat. Mat. Fis. **9**, 841–859 (1969) (in Russian)
3. Chadha, N.M., Kopteva, N.: Maximum norm a posteriori error estimate for a 3d singularly perturbed semilinear reaction-diffusion problem. Adv. Comput. Math. **35**, 33–55 (2011)
4. Demlow, A., Kopteva, N.: Maximum-norm a posteriori error estimates for singularly perturbed elliptic reaction-diffusion problems. Numer. Math. **133**, 707–742 (2016)
5. Kopteva, N.: Maximum norm error analysis of a 2d singularly perturbed semilinear reaction-diffusion problem. Math. Comput. **76**, 631–646 (2007)
6. Kopteva, N.: Maximum norm a posteriori error estimate for a 2d singularly perturbed reaction-diffusion problem. SIAM J. Numer. Anal. **46**, 1602–1618 (2008)

7. Kopteva, N.: Linear finite elements may be only first-order pointwise accurate on anisotropic triangulations. Math. Comput. **83**, 2061–2070 (2014)
8. Kopteva, N.: Maximum-norm a posteriori error estimates for singularly perturbed reaction-diffusion problems on anisotropic meshes. SIAM J. Numer. Anal. **53**, 2519–2544 (2015)
9. Kopteva, N.: Energy-norm a posteriori error estimates for singularly perturbed reaction-diffusion problems on anisotropic meshes. Numer. Math. (2017). Published online 2 May 2017. doi:10.1007/s00211-017-0889-3
10. Kopteva, N.: Fully computable a posteriori error estimator using anisotropic flux equilibration on anisotropic meshes. (2017, submitted for publication). http://www.staff.ul.ie/natalia/pubs.html
11. Kopteva, N., O'Riordan, E.: Shishkin meshes in the numerical solution of singularly perturbed differential equations. Int. J. Numer. Anal. Model. **7**, 393–415 (2010)
12. Kunert, G.: An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes. Numer. Math. **86**, 471–490 (2000)
13. Kunert, G.: Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes. Adv. Comput. Math. **15**, 237–259 (2001)
14. Kunert, G., Verfürth, R.: Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes. Numer. Math. **86**, 283–303 (2000)
15. Nochetto, R.H.: Pointwise a posteriori error estimates for monotone semi-linear equations. Lecture Notes at 2006 CNA Summer School Probabilistic and Analytical Perspectives on Contemporary PDEs (2006). http://www.math.cmu.edu/cna/Summer06/lecturenotes/nochetto/
16. Roos, H.-G., Stynes, M., Tobiska, T.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Springer, Berlin (2008)
17. Siebert, K.G.: An a posteriori error estimator for anisotropic refinement. Numer. Math. **73**, 373–398 (1996)
18. Verfürth, R.: Robust a posteriori error estimators for a singularly perturbed reaction-diffusion equation. Numer. Math. **78**, 479–493 (1998)

A DG Least-Squares Finite Element Method for Nagumo's Nerve Equation with Fast Reaction: A Numerical Study

Runchang Lin

Abstract The Nagumo equation is a simple nonlinear reaction-diffusion equation, which has important applications in neuroscience and biological electricity. If the equation is reaction-dominated, numerical oscillations may appear near the traveling wave front, which makes it challenging to find stable solutions. In the present study, a new method is developed on uniform meshes to solve the Nagumo equation. Numerical results are given to demonstrate the performance of the algorithm. Convergence rates with respect to spatial and temporal discretization are obtained experimentally. Some properties of the nerve model are confirmed numerically.

1 Introduction

All cells maintain an electrical potential difference across the cell membrane, which is used to assist or control their metabolic processes. Some cells make specialized use of bioelectric potentials and currents for distinctive physiological functions. For example, information is carried by *action potentials* passing along nerve cells, which is many orders of magnitude faster than molecular communications via mechanical transport. With the ingenious application of *voltage clamp* method to the giant axon of the Atlantic squid, Hodgkin and Huxley developed the first quantitative description of action potential propagation, which eventually lead them to the 1963 Nobel Prize in Physiology or Medicine together with Eccles. The Hodgkin-Huxley (HH) model is a system of one nonlinear differential equation (DE) for the membrane potential coupled with three ordinary DEs, which shows how cells can produce propagating pulses in multicellular organisms [13]. The HH model exhibits a principal *all-or-none* feature of the nerve: if an external applied current is below some threshold, the membrane potential returns quickly to the rest; if the current is above some threshold, there is an action potential; if the applied current

R. Lin (✉)
Department of Mathematics and Physics, Texas A&M International University (TAMIU), Laredo, TX 78041, USA
e-mail: rlin@tamiu.edu

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_12

is sufficiently large and held for a sufficiently long time, then the model generates a periodic response which propagates down the line as a traveling wave.

The FitzHugh-Nagumo (FHN) model is a simplification of the HH equations, which preserves major qualitative properties of the HH model [11, 21]. The general form of the FHN equations for the voltage variable v and the recovery variable w is

$$\begin{aligned} v_t &= Dv_{xx} + \rho v(1-v)(v-\alpha) - w + I_{app}, \\ w_t &= \epsilon(\beta_1 v - \beta_2 w), \end{aligned} \tag{1}$$

for $x \in \mathbb{R}$ and $t > 0$, where D and ρ are positive constants, α is the excitation threshold parameter, I_{app} is an external current applied to the cell, ϵ is the ratio of time scales, and β_1 and β_2 are positive constants for the rest state.

In most applications, ϵ is very small, so that the recovery variable is much slower than the voltage variable. In an initial time period we may assume $w = 0$ [6]. The resulting model from the FHN system (1) is the Nagumo equation

$$v_t = Dv_{xx} + \rho v(1-v)(v-\alpha), \quad x \in \mathbb{R}, \quad t > 0, \tag{2}$$

which is a widely used model of excitation and propagation of impulse in nerve membranes. The Nagumo equation is also used in studies of circuit theory and population genetics. Moreover, its counterpart in higher spacial dimensions is the famous Allen-Cahn equation [1], which was introduced to describe the process of phase separation in multi-component alloy systems.

In this study, we focus on approximation of the *traveling wave solution* (TWS) to Eq. (2), which is guaranteed by the following result [6, Theorem 4.35].

Theorem 1 *Consider the bounded solution for the generalized Nagumo equation*

$$v_t = v_{xx} + f(v), \quad x \in \mathbb{R}, \quad t > 0 \tag{3}$$

with initial conditions $v(x,0) = v_0(x) \geq \max\{v_p(x), 0\}$ *satisfies* $v(x,t) \to 1$ *as* $t \to +\infty$ *for each* $x \in \mathbb{R}$, *where* $v_p(x)$ *is a particular stationary solution of* (3). *Here* f *satisfies* $f(0) = f(\alpha) = f(1) = 0$, $f < 0$ *in* $(0, \alpha)$, $f > 0$ *in* $(\alpha, 1)$, $f'(0) < 0$, $f'(1) < 0$, *and* $\int_0^1 f(z)dz > 0$. *Then there exists a wave front from* $v = 0$ *to* $v = 1$ *with a unique wave speed* $c > 0$.

Solving nonlinear DEs is in general challenging. Numerical investigation has been playing a more and more important role in providing insight for the analytical study of nonlinear DEs. A variety of numerical approaches, such as finite difference (FD) method [25], finite element (FE) method [20, 26], finite volume method [9], and pseudospectral method [23], have been applied to solve the FHN equations.

In this article, we develop a discontinuous Galerkin (DG) least-squares (LS) FE method for solving Nagumo's equation, where the temporal discretization is by a FD scheme. For details about the LSFE and the DG methods, the reader is referred to [3] and [8, 10], respectively, and the references therein. The DG and LSFE techniques are combined in order to retain their advantageous features and to obtain a superior FE method for problems with sharp changes in solutions. This idea has been adopted by many researchers in numerical investigations of different problems; see, e.g., [2–5, 7, 15]. In particular, a DG LSFE method was introduced to solve linear second order elliptic reaction-diffusion equations with singular perturbation in [17, 18]. A DG LSFE method has been proposed for the Fisher-KPP equation [19], which is a more challenging problem. However, for reaction dominated problems (i.e. $\rho/D \gg 1$), numerical oscillations will occur near the wave front. Moveover, solutions to the problem of higher spatial dimensions (i.e. the Allen-Cahn equation) may develop into patterns with complicated interfaces. Therefore, finding accurate and stable numerical solutions to the Nagumo equation with fast reaction remains a challenging problem [25].

The paper is organized as follows. In Sect. 2, a DG LSFE scheme is developed for (2). In Sect. 3, we demonstrate the performance of the numerical scheme via experiments. In Sect. 4, the article concludes with some comments and remarks.

2 A DG LSFE Scheme

We consider the equation

$$v_t = Dv_{xx} + \rho v(1-v)(v-\alpha), \tag{4a}$$

$$v(x,0) = v_0(x), \quad \text{and} \quad \lim_{x\to\pm\infty} v_x = 0, \tag{4b}$$

where $x \in \mathbb{R}$, $t > 0$, and $0 < \alpha < 1/2$. We restrict the computational domain to bounded spatial interval $\Omega = (x_L, x_R)$ for some $x_L, x_R \in \mathbb{R}$ and temporal interval $(0, T]$. The corresponding truncated initial-boundary value (IBV) problem of (4) is

$$\begin{aligned} &v_t = Dv_{xx} + \rho v(1-v)(v-\alpha), \quad x \in \Omega,\ t \in (0,T], \\ &v(x,0) = v_0(x), \quad x \in \Omega, \\ &v_x(x_L,t) = 0,\ v_x(x_R,t) = 0, \quad t \in (0,T]. \end{aligned}$$

When Ω is sufficiently large, the difference between this truncated problem and the problem (4) in unbounded domain is negligible. Including a new variable $q = \sqrt{D}v_x$, the second order reaction-diffusion equation is recast into a system of first

order DEs

$$\begin{aligned} &q - \sqrt{D} v_x = 0, \quad x \in \Omega,\, t \in (0, T], \\ &v_t - \sqrt{D} q_x - \rho v(1-v)(v-\alpha) = 0, \quad x \in \Omega,\, t \in (0, T], \\ &v(x, 0) = v_0(x), \quad x \in \Omega, \\ &v_x(x_L, t) = 0,\ v_x(x_R, t) = 0, \quad t \in (0, T]. \end{aligned} \tag{5}$$

We first discretize the time derivative. The pseudo Crank-Nicolson approximation of the governing equations in (5) is

$$\begin{aligned} &q^{n+1} - \sqrt{D} v_x^{n+1} = 0, \\ &\frac{v^{n+1} - v^n}{\tau} - \sqrt{D} \frac{q_x^{n+1} + q_x^n}{2} - \rho \frac{v^{n+1} + v^n}{2} (1 - v^n)(v^n - \alpha) = 0, \end{aligned} \tag{6}$$

for $n \geq 0$, where τ is the time step size, $v^n = v(x, t_n)$ is the solution at time level $t_n = n\tau$, and q^n is defined similarly. Let $\mathbf{v}^n = (q^n\ v^n) \in H_0^1(\Omega) \times H^1(\Omega)$. Define

$$A_{\mathbf{v}^n} \mathbf{v}^{n+1} = \begin{pmatrix} q^{n+1} - \sqrt{D} v_x^{n+1} \\ -\tau \sqrt{D}\, q_x^{n+1} + \big(2 - \tau\rho(1 - v^n)(v^n - \alpha)\big) v^{n+1} \end{pmatrix},$$

$$\mathbf{f}_{\mathbf{v}^n} = \begin{pmatrix} 0 \\ \tau \sqrt{D}\, q_x^n + \big(2 + \tau\rho(1 - v^n)(v^n - \alpha)\big) v^n \end{pmatrix}.$$

Then Eq. (6) read

$$A_{\mathbf{v}^n} \mathbf{v}^{n+1} = \mathbf{f}_{\mathbf{v}^n} \tag{7}$$

for $n \geq 0$. At each time level, define the LS functional $\mathscr{J}(\mathbf{u}; \mathbf{v}^n)$ by

$$\mathscr{J}(\mathbf{u}; \mathbf{v}^n) = ||A_{\mathbf{v}^n} \mathbf{u} - \mathbf{f}_{\mathbf{v}^n}||^2_{L^2(\Omega)}. \tag{8}$$

Minimizing (8), the residual of (7), we obtain an LS variational formulation: find $\mathbf{v}^{n+1} = (q^{n+1}\ v^{n+1}) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$\big(A_{\mathbf{v}^n} \mathbf{v}^{n+1}, A_{\mathbf{v}^n} \mathbf{u}\big) = \big(\mathbf{f}_{\mathbf{v}^n}, A_{\mathbf{v}^n} \mathbf{u}\big) \quad \forall \mathbf{u} \in H_0^1(\Omega) \times H^1(\Omega), \quad n \geq 0, \tag{9}$$

where $(\cdot, \cdot)$ denotes the inner product in $(L^2(\Omega))^2$.

Let $\mathscr{T} : x_L = x_0 < \cdots < x_M = x_R$ be a partition of Ω. In this paper, we use uniform mesh. Thus the mesh size is $h = (x_R - x_L)/M$ and $x_i = x_L + ih$ for $0 \leq i \leq M$. The standard broken Sobolev spaces $H^1(\Omega, \mathscr{T})$ and $H_0^1(\Omega, \mathscr{T})$ can be defined. The FE space $\mathbf{V}_h \subset H_0^1(\Omega, \mathscr{T}) \times H^1(\Omega, \mathscr{T})$ consists of vector-valued

functions with each component a piecewise polynomial of degree k, which may not be continuous. For $g \in \mathbf{V}_h$, define its *jump* and *average* at a node x_i by

$$[\![f]\!]_i = \begin{cases} f(x_i^-) - f(x_i^+) & 0 < i < M, \\ 0 & i = 0, M, \end{cases} \qquad \{f\}_i = \begin{cases} \frac{f(x_i^-)+f(x_i^+)}{2} & 0 < i < M, \\ f(x_i) & i = 0, M, \end{cases}$$

respectively. Given the numerical solution $\mathbf{v}_h^n$, the DG finite element approximation of (9) at time level $n+1$ is: find $\mathbf{v}_h^{n+1} = (q_h^{n+1}\ v_h^{n+1}) \in \mathbf{V}_h$ such that

$$\begin{aligned} &B_h^n(\mathbf{v}_h^{n+1}, \mathbf{u}) = L_h^n(\mathbf{u}) \quad \forall\, \mathbf{u} = (p\ u) \in \mathbf{V}_h, \\ &v_h^0(x) = v_0(x), \quad x \in \Omega, \\ &q_h^{n+1}(x_L, t) = 0, \quad q_h^{n+1}(x_R, t) = 0, \quad t \in (0, T], \end{aligned} \tag{10}$$

where

$$L_h^n(\mathbf{u}) = \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} \left((4 - \delta_h^n)\bar{v}_h^n + \tau\sqrt{D}\bar{q}_{h,x}^n\right)(\delta_h^n u - \tau\sqrt{D}p_x)dx,$$

$$\begin{aligned} B_h^n(\mathbf{v}_h^{n+1}, \mathbf{u}) = &\sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} \Big[(p - \sqrt{D}u_x)q_h^{n+1} - \sqrt{D}(1 - \tau\delta_h^n)pv_{h,x}^{n+1} + \sqrt{D}\tau(\delta_h^n p)_x v_h^{n+1} \\ &+ Du_x v_{h,x}^{n+1} + \delta_h^n\left(\delta_h^n u - \tau\sqrt{D}p_x\right)v_h^{n+1} + \tau\sqrt{D}(\delta_h^n u)_x q_h^{n+1} + \tau^2 Dp_x q_{h,x}^{n+1}\Big]dx \\ &- \sum_{i=1}^{M} \tau\delta_h^n\sqrt{D}\Big[p\widehat{v}_h^{n+1} + u\widehat{q}_h^{n+1}\Big]_{x_{i-1}}^{x_i} + \sum_{i=1}^{M-1} \tau\delta_h^n\sqrt{D}\Big([\![p]\!]_i\{v_h^{n+1}\}_i + [\![u]\!]_i\{q_h^{n+1}\}_i\Big), \end{aligned}$$

and

$$\delta_h^n = 2 - \tau\rho(1 - \bar{v}_h^n)(\bar{v}_h^n - \alpha). \tag{11}$$

Here $\bar{v}_h^n$ and $\bar{p}_h^n$ are continuous piecewise such that $\bar{v}_h^n(x_i) = \frac{v_h^n(x_i^-)+v_h^n(x_i^+)}{2}$ and $\bar{p}_h^n(x_i) = \frac{p_h^n(x_i^-)+p_h^n(x_i^+)}{2}$, $\widehat{v}_h^{n+1}$ and $\widehat{q}_h^{n+1}$ are numerical fluxes defined by

$$\begin{aligned} \widehat{v}_h^{n+1}(x_i) &= \begin{cases} \xi v_h^{n+1}(x_i^-) + (1-\xi)v_h^{n+1}(x_i^+) & i = 1, \ldots, M-1, \\ v_h^{n+1}(x_i) & i = 0, M, \end{cases} \\ \widehat{q}_h^{n+1}(x_i) &= \begin{cases} (1-\xi)q_h^{n+1}(x_i^-) + \xi q_h^{n+1}(x_i^+) & i = 1, \ldots, M-1, \\ 0 & i = 0, M, \end{cases} \end{aligned} \tag{12}$$

where $0 \le \xi \le 1$ is a parameter. We shall remark that the bilinear form of the DG formulation is obtained from the conventional integration by parts process, where

the parameter ξ is included to allow the convenience of implementing different numerical fluxes. A natural *energy norm* $|||\cdot|||$ in $\mathbf{V}_h$ is thus defined by

$$|||\mathbf{u}|||^2 = \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} \left(p_x^2 + p^2 + u_x^2 + u^2\right) dx.$$

From (12), it is straightforward to verify that problems (9) and (10) are *consistent*. Using the Cauchy-Schwarz inequality, it is plain to show that $L_h^n(\cdot)$ is continuous in $\mathbf{V}_h$. The following results are needed for well-posedness of problem (10) in $\mathbf{V}_h$.

Theorem 2 *Let $B_h^n(\cdot,\cdot)$ be the bilinear form defined in* (10)*. Assume that τ is sufficiently small.*

(*i*) *There exists a constant $C_1 > 0$ such that*

$$B_h^n(\mathbf{u},\mathbf{u}) \geq C_1 |||\mathbf{u}|||^2 \quad \forall \mathbf{u} \in \mathbf{V}_h. \tag{13}$$

(*ii*) *There exists a constant $C_2 > 0$ such that*

$$|B_h^n(\mathbf{v},\mathbf{u})| \leq C_2 |||\mathbf{v}||| \, |||\mathbf{u}||| \quad \forall \mathbf{u},\mathbf{v} \in \mathbf{V}_h. \tag{14}$$

Proof

(*i*) It is evident from (11) that $\delta_h^n \geq \delta_0 > 1$ when τ is small. Note that

$$\begin{aligned} B_h^n(\mathbf{u},\mathbf{u}) = \sum_{i=1}^{M} \int_{x_{i-1}}^{x_i} \Big[& (p - \sqrt{D}u_x)p - \sqrt{D}(1 - \tau\delta_h^n)pu_x + \tau\sqrt{D}(\delta_h^n p)_x u \\ & + Du_x^2 + (\delta_h^n)^2 u^2 - \tau\sqrt{D}\delta_h^n p_x u + \tau\sqrt{D}(\delta_h^n u)_x p + \tau^2 D p_x^2 \Big] dx \\ & - \sum_{i=1}^{M} \tau\sqrt{D}\delta_h^n \Big[p\widehat{u} + u\widehat{p} \Big]_{x_{i-1}}^{x_i} + \sum_{i=1}^{M-1} \tau\delta_h^n \sqrt{D} \Big(\{p\}_i [\![u]\!]_i + \{u\}_i [\![p]\!]_i \Big). \end{aligned} \tag{15}$$

Using integration by parts, one gets

$$\begin{aligned} \int_{x_{i-1}}^{x_i} (\delta_h^n p)_x u \, dx &= -\int_{x_{i-1}}^{x_i} \delta_h^n p u_x \, dx + \Big[\delta_h^n pu \Big]_{x_{i-1}}^{x_i}, \\ -\int_{x_{i-1}}^{x_i} \delta_h^n p_x u \, dx &= \int_{x_{i-1}}^{x_i} (\delta_h^n u)_x p \, dx - \Big[\delta_h^n pu \Big]_{x_{i-1}}^{x_i}. \end{aligned}$$

Hence, by Young's inequality, the first term in the right hand side of (15) is

$$\sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}\Big[p^2+Du_x^2+(\delta_h^n)^2u^2+\tau^2Dp_x^2-2\sqrt{D}(1-\tau\delta_h^n)pu_x+2\tau\sqrt{D}(\delta_h^n)_xup\Big]dx$$

$$\geq\sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}\Big[\tau\delta_h^np^2+\tau\delta_h^nDu_x^2+(\delta_h^n)^2u^2+\tau^2Dp_x^2+2\tau\sqrt{D}(\delta_h^n)_xup\Big]dx$$

$$\geq\sum_{i=1}^{M}\int_{x_{i-1}}^{x_i}\Big[\big(\tau\delta_h^n-\tau^2(\delta_h^n)_x^2D\big)p^2+\tau\delta_h^nDu_x^2+\big((\delta_h^n)^2-1\big)u^2+\tau^2Dp_x^2\Big]dx$$

$$\geq\min\left\{\tau\delta_0-\tau^2\max_{1\leq i\leq M}(\delta_h^n|_{I_i})_x^2D,\delta_0^2-1,\tau^2D\right\}|||\mathbf{u}|||^2.$$

On the other hand, the second term in the right hand side of (15) is

$$\begin{aligned}&\sum_{i=1}^{M}\tau\delta_h^n\sqrt{D}\Big[p(x_i^-)\big(\xi u(x_i^-)+(1-\xi)u(x_i^+)\big)-p(x_{i-1}^+)\big(\xi u(x_{i-1}^-)+(1-\xi)u(x_{i-1}^+)\big)\\&\quad+u(x_i^-)\big((1-\xi)p(x_i^-)+\xi p(x_i^+)\big)-u(x_{i-1}^+)\big((1-\xi)p(x_{i-1}^-)+\xi p(x_{i-1}^+)\big)\Big]\\&=\sum_{i=1}^{M}\tau\delta_h^n\sqrt{D}\Big[p(x_i^-)u(x_i^-)+\xi p(x_i^+)u(x_i^-)+(1-\xi)p(x_i^-)u(x_i^+)\\&\quad-p(x_{i-1}^+)u(x_{i-1}^+)-\xi p(x_{i-1}^+)u(x_{i-1}^-)-(1-\xi)p(x_{i-1}^-)u(x_{i-1}^+)\Big]\\&=\sum_{i=1}^{M-1}\tau\delta_h^n\sqrt{D}\Big[p(x_i^-)u(x_i^-)-p(x_i^+)u(x_i^+)\Big]=\sum_{i=1}^{M-1}\tau\delta_h^n\sqrt{D}\Big(\{p\}_i[\![u]\!]_i+\{u\}_i[\![p]\!]_i\Big),\end{aligned}$$

where homogeneous boundary conditions for p and the continuity of δ_h^n have been used to obtain the last two equalities. The desired result (13) follows from the above inequalities with $C_1=\min\left\{\tau\delta_0-\tau^2\max_{1\leq i\leq M}(\delta_h^n|_{I_i})_x^2D,\delta_0^2-1,\tau^2D\right\}$, which is positive when τ is sufficiently small.

(*ii*) The continuity result (14) is proved by integration by parts, triangle inequality, Cauchy-Schwarz inequality, and Young's inequality. □

Remark 1 The well-posedness of problem (10) depends on the choice of τ. In particular, it is evident from the proof of Theorem 2 that the stability constants C_1 and C_2 can be achieved only if τ, i.e. $\tau\rho$, is sufficiently small.

The iteration (10) can now be initiated with the initial condition, from which $q_{h,x}^0$ can be obtained. The evolution of the solution over time can thus be approximated provided that the wave front is away from the boundaries of Ω.

3 Numerical Simulations and Discussions

In this section, numerical results are presented to demonstrate the effectiveness of the DG LSFE scheme. Simulations with different configurations have been performed. For all examples, $D = 1$, $\alpha = 0.25$, and $\xi = 1/2$. Linear finite elements have been used for spatial discretization.

Example 1 **Example with exact solution**. Some exact solutions to the Nagumo equation (4) are available in the literature, which will be used to test accuracy of the numerical scheme. In particular, Kawahara and Tanaka [14] (cf. also [22]) found the following exact solution to (4)

$$v(x,t) = \frac{Ae^{\Gamma_1} + \alpha Be^{\Gamma_2}}{Ae^{\Gamma_1} + Be^{\Gamma_2} + C}, \tag{16}$$

where $\Gamma_1 = \left(\pm\sqrt{2\rho/D}x + (1-2\alpha)\rho t\right)/2$, $\Gamma_2 = \left(\pm\sqrt{2\rho/D}\alpha x + \alpha(\alpha-2)\rho t\right)/2$, and A, B, and C are arbitrary constants. Three other exact solutions are also reported in [14]. The plus and minus signs are due to the symmetry of the Nagumo equation in x, which are corresponding to wave fronts traveling to the left and to the right, respectively. On the other hand, the following exact solution to (4) is found in [24]:

$$v(x,t) = \left(1 + \exp\left(\frac{\pm\sqrt{2\rho/D}x + (2\alpha-1)\rho t}{2}\right)\right)^{-1}. \tag{17}$$

More exact solutions to (4) are also found in [16, 22].

Consider Eq. (5) with an initial condition in accordance with the exact solutions (17), where positive sign is used. The solution satisfies $\lim_{t\to+\infty} v(x,t) = 1$ and $\lim_{x\to\pm\infty} v_x(x,t) = 0$. The numerical errors are measured in discrete maximum norms; e.g., $||v^n - v_h^n||_\infty = \max_{0\le i\le M} \left|v(x_i,t_n) - v_h(x_i,t_n)\right|$.

To test the order of accuracy of the DG LSFE scheme with respect to spatial discretization, we fix the temporal step τ, and half the spatial step h for several times; and vice versa. First consider the case when $\rho = 1$ with an initial condition from (17). Let $\Omega = [-25, 25]$. Table 1 demonstrates numerical errors and convergence rates in spatial and temporal variables at time levels $t = 2.0, 4.0$, and 8.0. Data for various h with fixed $\tau = 2.5e{-}4$ and for various τ with fixed $h = 0.05$ are provided. The convergence rate r is computed as of order $O(h^r)$ or $O(\tau^r)$, respectively. The numerical results indicate that $O(h^2)$ order is obtained for both v_h^n and q_h^n in spatial variable, and the convergence rates in temporal variable are both $O(\tau)$. Recall that linear finite elements are used for spatial discretization, and that the linearized FD scheme in time is a *pseudo* Crank-Nicolson scheme. The spatial and temporal convergence rates in Table 1 are hence both optimal. For the case when $\rho = 1000$, the solution has steeper wave front (cf. [12]), which propagates at a much higher velocity. In this case, τ has to be sufficiently small by Theorem 2; cf. also (11). As a consequence, h has to be small as well to meet the standard stability requirement.

Table 1 Example 1—Numerical errors and convergence rates when $\rho = 1$ with $\Omega = [-25, 25]$

Convergence in spatial discretization with $\tau = 2.5e-4$						
	$\|\|v^n - v_h^n\|\|_\infty$			$\|\|q^n - q_h^n\|\|_\infty$		
h	$t = 2.0$	$t = 4.0$	$t = 8.0$	$t = 2.0$	$t = 4.0$	$t = 8.0$
0.8	3.2329e−3	3.9163e−3	4.0093e−3	2.0134e−3	1.3582e−3	1.0958e−3
0.4	8.5642e−4	1.0343e−3	1.2272e−3	4.0292e−4	2.4636e−4	2.6035e−4
0.2	2.1738e−4	2.6104e−4	3.2546e−4	9.3501e−5	5.9948e−5	6.9888e−5
0.1	5.5091e−5	6.5322e−5	8.9589e−5	2.2221e−5	1.5494e−5	1.9840e−5
0.8	–	–	–	–	–	–
0.4	1.92	1.92	1.71	2.32	2.46	2.07
0.2	1.98	1.99	1.91	2.11	2.04	1.90
0.1	1.98	2.00	1.86	2.07	1.95	1.82
Convergence in temporal discretization with $h = 0.05$						
	$\|\|v^n - v_h^n\|\|_\infty$			$\|\|q^n - q_h^n\|\|_\infty$		
τ	$t = 2.0$	$t = 4.0$	$t = 8.0$	$t = 2.0$	$t = 4.0$	$t = 8.0$
0.08	1.0616e−3	1.8959e−3	3.5205e−3	6.9533e−4	9.8226e−4	1.3983e−3
0.04	5.2850e−4	9.4375e−4	1.7521e−3	3.4647e−4	4.8935e−4	6.9628e−4
0.02	2.6304e−4	4.6976e−4	8.7213e−4	1.7284e−4	2.4404e−4	3.4704e−4
0.01	1.3142e−4	2.3452e−4	4.3512e−4	8.6013e−5	1.2167e−4	1.7312e−4
0.08	–	–	–	–	–	–
0.04	1.01	1.01	1.01	1.01	1.01	1.01
0.02	1.01	1.01	1.01	1.00	1.00	1.00
0.01	1.00	1.00	1.00	1.01	1.00	1.00

To focus on the investigation of accuracy orders, we use $\Omega = [-3, 3]$ for this case. Errors and convergence rates at time levels $t = 0.01, 0.02$, and 0.04 are collected in Table 2. Data for various h with fixed $\tau = 1.0e-6$ and data for various τ with fixed $h = 1.0e-3$ are provided. Similar optimal convergence results are observed.

Example 2 **Example without exact solution**. In this example, we verify the conclusion of Theorem 1 by solving Eq. (5) with $\rho = 1$, the computational domain $[-80, 80] \times (0, 20]$, and $v_0(x) = 0$ if $x < 0$, $v_0(x) = \alpha$ if $0 \le x < 30$, and $v_0(x) = 1$ if $x \ge 30$. Here $h = 0.1$ and $\tau = 0.05$. The exact solution is not available. Figure 1 shows how the initial step function evolves into a monotone unique speed traveling wave front that joins the two stable states.

Example 3 **Impact of initial stimulus**. The physiological background of the Nagumo model is that if the stimuli are below threshold of conscious awareness then no information will be conveyed; but if a stimulus is above threshold, it may convert into a train of pulses of fixed shape and travel down the nerve with little distortion.

Table 2 Example 1—Numerical errors and convergence rates when $\rho = 1000$ with $\Omega = [-3, 3]$

Convergence in spatial discretization with $\tau = 1.0e-6$						
	$\|\|v^n - v_h^n\|\|_\infty$			$\|\|q^n - q_h^n\|\|_\infty$		
h	$t = 0.01$	$t = 0.02$	$t = 0.04$	$t = 0.01$	$t = 0.02$	$t = 0.04$
0.08	8.7936e−2	1.1596e−1	2.0171e−1	3.0588e−1	6.4459e−1	1.4226e−0
0.04	1.6309e−2	2.4982e−2	4.5830e−2	1.0620e−1	1.9568e−1	3.4558e−1
0.02	3.7582e−3	6.0187e−3	1.0729e−2	2.6090e−2	4.8068e−2	9.4322e−2
0.01	9.5568e−4	1.5447e−3	2.8110e−3	6.7946e−3	1.2641e−2	2.4722e−2
0.64	–	–	–	–	–	–
0.32	2.43	2.21	2.14	1.53	1.72	2.04
0.16	2.12	2.05	2.09	2.03	2.03	1.87
0.08	1.98	1.96	1.93	1.94	1.93	1.93
Convergence in temporal discretization with $h = 1.0e-3$						
	$\|\|v^n - v_h^n\|\|_\infty$			$\|\|q^n - q_h^n\|\|_\infty$		
τ	$t = 0.01$	$t = 0.02$	$t = 0.04$	$t = 0.01$	$t = 0.02$	$t = 0.04$
8.0*e*-5	4.3286e−3	8.5319e−3	1.7064e−2	5.0506e−2	8.4813e−2	1.5720e−1
4.0*e*-5	2.1705e−3	4.2774e−3	8.5552e−3	2.5245e−2	4.2422*e*−2	7.8659e−2
2.0*e*-5	1.0893e−3	2.1459e−3	4.2921e−3	1.2627e−2	2.1242*e*−2	3.9414e−2
1.0*e*-5	5.4822e−4	1.0792e−3	2.1587e−3	6.3210e−3	1.0656*e*−2	1.9795e−2
8.0*e*-5	–	–	–	–	–	–
4.0*e*-5	1.00	1.00	1.00	1.00	1.00	1.00
2.0*e*-5	0.99	1.00	1.00	1.00	1.00	1.00
1.0*e*-5	0.99	0.99	0.99	1.00	1.00	0.99

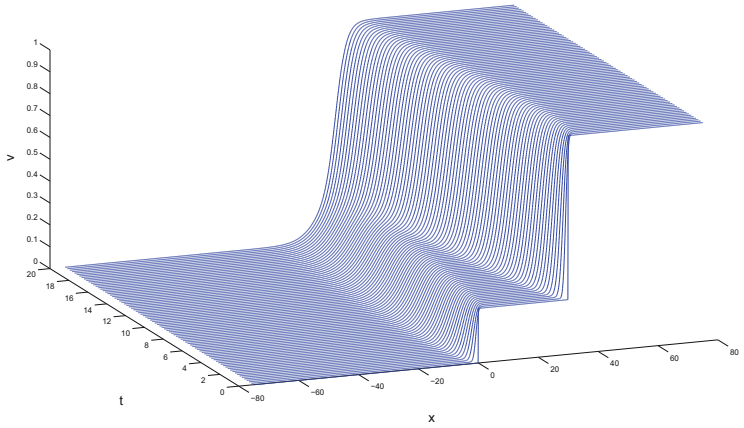

Fig. 1 Example 2—The evolution of the wavefront during time range $0 \le t \le 20$, where the time increment between profiles is 0.25

We solve the Nagumo model (2) with $\rho = 1$. Set $\Omega = [-50, 50]$. We first test the situation with a subthreshold stimulus, e.g. $v_0(x) = 0.249$ for $x \in [-25, 25]$, and $v_0(x) = 0$ elsewhere. It is observed in Fig. 2 (left pair) that the initial wave

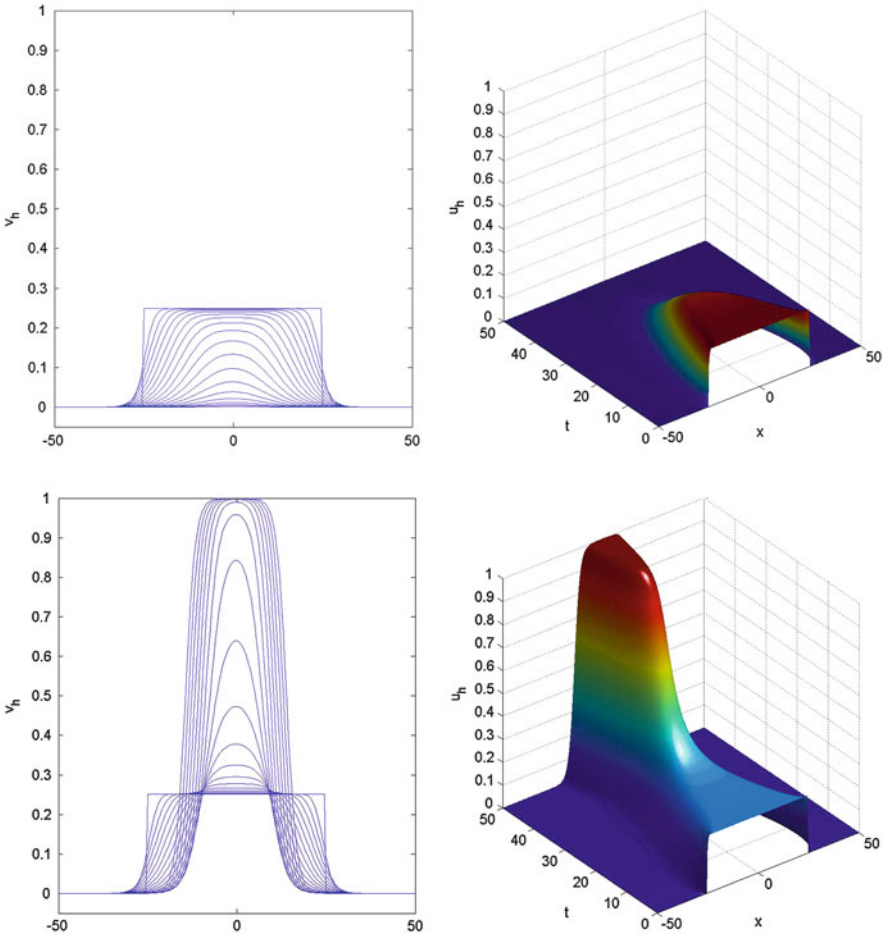

Fig. 2 Example 3—Evolutions of waves with subthreshold (initial stimulus position 0.249) and suprathreshold (initial stimulus position 0.251) stimuli, which are presented in the *left* and *right* pairs of plots, respectively. Both initial stimuli are supported over $[-25, 25]$. Snapshots and histories over time range $0 \leq t \leq 50$ with time increment 2.5 are presented

damps to zero eventually. On the other hand, if given a suprathreshold stimulus, e.g. $v_0(x) = 0.251$ for $x \in [-25, 25]$, and $v_0(x) = 0$ elsewhere, then it will rapidly develop into a traveling wave, as depicted in Fig. 2 (right pair).

The development of wave depends also on the support of the initial stimulus, which determines the energy of the stimulant. Figure 3 shows the histories of waves with nonzero initial value $v_0(x) = 0.5$ for $x \in [-3, 3]$ and $x \in [-2, 2]$, respectively. The initial pulse of the latter case fails to propagate into a traveling wave due to low energy, though it surpasses the threshold.

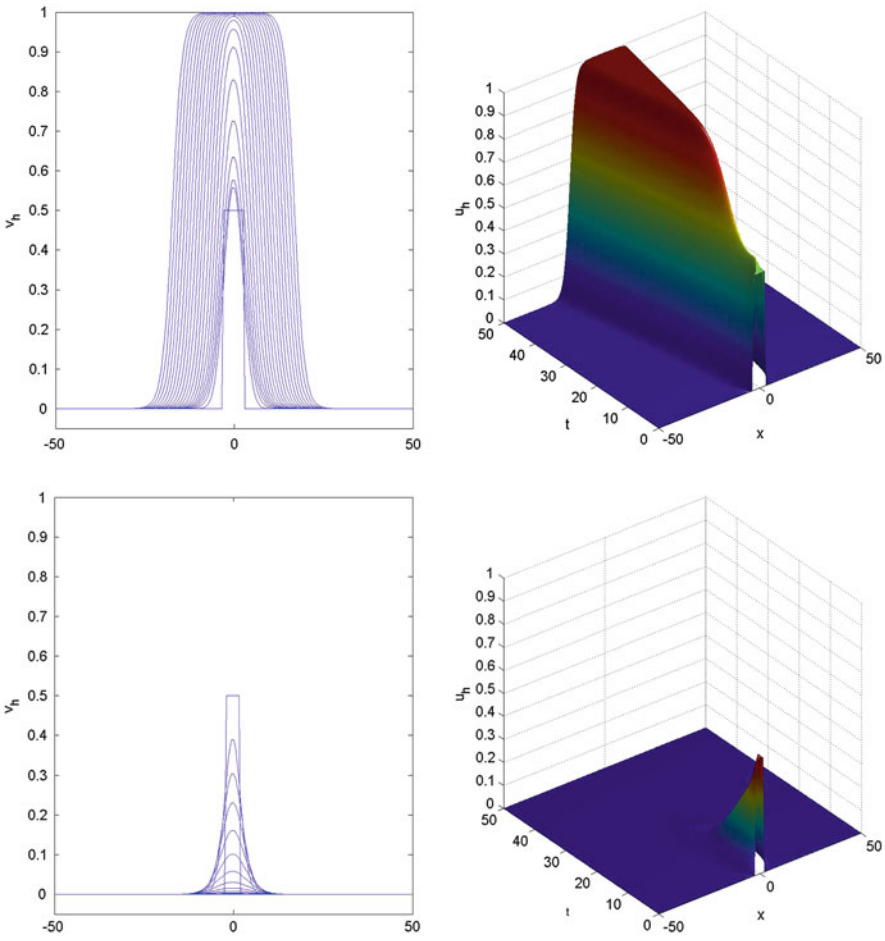

Fig. 3 Example 3—Evolutions of waves with initial stimuli supported by $[-3, 3]$ (*left pair*) and $[-2, 2]$ (*right pair*), respectively. Both initial stimulus positions are 0.5. Snapshots and histories over time range $0 \leq t \leq 50$ with time increment 2.5 are presented

Example 4 **Metastability**. For a multi-stable system, the stable state with lesser potential value is called the metastable state. Under random perturbations of proper intensity, a TWS connecting the stable states will be moving in the direction of the metastable state. The transitions between different stable states can last very long, which is a characteristic feature of metastable systems. In this example, we exam the metastability of solutions to the Nagumo equation (2) with $\rho = 1$ on $[-60, 60]$ by using an initial condition $v_0(x) = 0.01$ for $x \in [-35, -20] \cup [5, 35]$, and $v_0(x) = 1$ elsewhere. In Fig. 4, it is observed that the two wells are absorbed by the excited state $v \equiv 1$ after a long time. Moreover, the absorption time for the well with larger width is much longer.

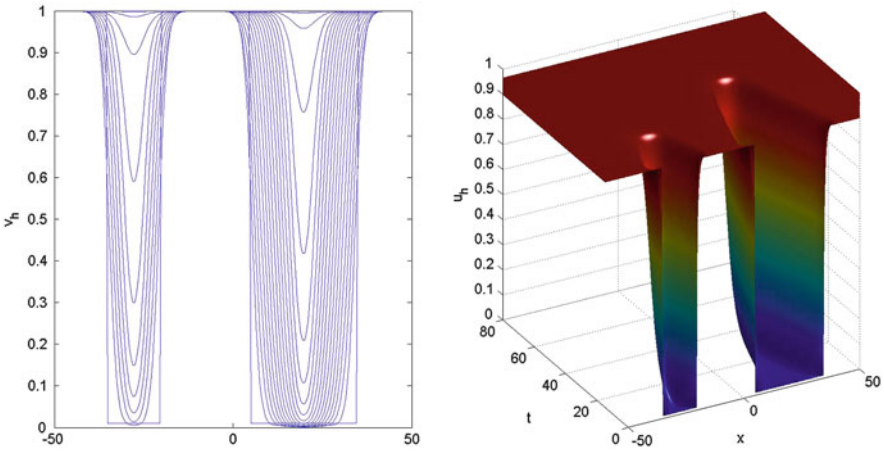

Fig. 4 Example 4—Metastability of solutions to Nagumo equations. *Left*: snapshots for $0 \leq t \leq 75$ with time increment 2.5 between profiles; *Right*: wave evolution history for $0 \leq t \leq 75$

4 Conclusions

In this article an efficient and accurate DG LSFE scheme is developed for solving the Nagumo equation with fast reaction. The second order problem is cast into a mixed system of first order DEs. A pseudo Crank-Nicolson scheme is used to discretize the time derivative. At each time level, the variational formulation is obtained in least-squares sense, which is approximated by using discontinuous Galerkin FE method. This method is stable and efficient. The convergence rates of the method are $O(h^2)$ and $O(\tau)$ in spatial and temporal discretizations, respectively. Numerical simulations for the Nagumo equation with more generic conditions can be conducted using the DG LSFE scheme. Several important properties of the Nagumo equation have been studied numerically. This method is ready to be extend to solve the Allen-Cahn equation, which is an undergoing project.

Acknowledgements This research is partially supported by the US NSF grant DMS 1217268, the NNSF of China under grant 11428103, and a University Research Grant of TAMIU.

References

1. Allen, S.M., Cahn, J.W.: Ground state structures in ordered binary alloys with second neighbor interactions. Acta Metall. **20**(3), 423–433 (1972)
2. Bensow, R.E., Larson, M.G.: Discontinuous/continuous least-squares finite element methods for elliptic problems. Math. Models Methods Appl. Sci. **15**, 825–842 (2005)
3. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Applied Mathematical Sciences, vol. 166. Springer, New York (2009)
4. Bochev, P.B., Lai, J., Olson, L.: A locally conservative, discontinuous least-squares finite element method for the Stokes equations. Int. J. Numer. Methods Fluids **68**, 782–804 (2012)

5. Bochev, P.B., Lai, J., Olson, L.: A non-conforming least-squares finite element method for incompressible fluid flow problems. Int. J. Numer. Methods Fluids **72**, 375–402 (2013)
6. Britton, N.F.: Reaction-Diffusion Equations and Their Applications to Biology. Academic, London (1986)
7. Cao, Y., Gunzburger, M.D.: Least-squares finite element approximations to solutions of interface problems. SIAM J. Numer. Anal. **35**, 393–405 (1998)
8. Cockburn, B., Karniadakis, G.E., Shu, C.-W.: Discontinuous Galerkin Methods. Theory, Computation and Applications. Lecture Notes in Computational Science and Engineering, vol. 11. Springer, New York (2000)
9. Coudière, Y., Pierre, C.: Stability and convergence of a finite volume method for two systems of reaction-diffusion equations in electro-cardiology. Nonlinear Anal. Real World Appl. **7**(4), 916–935 (2006)
10. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Mathématiques & Applications, vol. 69. Springer, Heidelberg (2012)
11. FitzHugh, R.A.: Impulses and physiological states in theoretical models of nerve membrane. Biophys. J. **1**, 445–466 (1961)
12. Griffiths, G.W., Schiesser, W.E.: Traveling Wave Analysis of Partial Differential Equations. Numerical and analytical Methods with MATLAB® and Maple^{TM}. Elsevier/Academic, Amsterdam (2012)
13. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerves. J. Physiol. **117**, 500–544 (1952)
14. Kawahara, T., Tanaka, M.: Interactions of traveling fronts: an exact solution of a nonlinear diffusion equation. Phys. Lett. A **97**(8), 311–314 (1983)
15. Lai, J., Bochev, P.B., Olson, L., Peterson, K., Ridzal, D., Siefert, C.: A discontinuous velocity least squares finite element method for the stokes equations with improved mass conservation. In: CSRI Summer Proceedings 2010, Sandia National Laboratory (2010)
16. Li, H., Guo, Y.: New exact solutions to the Fitzhugh-Nagumo equation. Appl. Math. Comput. **180**(2), 524–528 (2006)
17. Lin, R.: Discontinuous discretization for least-squares formulation of singularly perturbed reaction-diffusion problems in one and two dimensions. SIAM J. Numer. Anal. **47**(1), 89–108 (2008/09)
18. Lin, R.: Discontinuous Galerkin least-squares finite element methods for singularly perturbed reaction-diffusion problems with discontinuous coefficients and boundary singularities. Numer. Math. **112**(2), 295–318 (2009)
19. Lin, R., Zhu, H.: A discontinuous Galerkin least-squares finite element method for solving Fishers equation. Discrete Contin. Dyn. Syst. Ser. A Suppl. 489–497 (2013)
20. Mitchell, A.R., Griffiths, D.F., Meiring, A.: Finite element Galerkin methods for convection-diffusion and reaction-diffusion. Analytical and numerical approaches to asymptotic problems in analysis. In: Proceedings of the Conference, University of Nijmegen, Nijmegen, pp. 157–177 (1980)
21. Nagumo, J., Arimoto, S., Yoshizawa, S.: An active pulse transmission line simulating nerve axon. Proc. IREE **50**, 2061–2070 (1962)
22. Nucci, M.C., Clarkson, P.A.: The nonclassical method is more general than the direct method for symmetry reductions. An example of the Fitzhugh-Nagumo equation. Phys. Lett. A **164**, 49–56 (1992)
23. Olmos, D., Shizgal, B.D.: Pseudospectral method of solution of the Fitzhugh-Nagumo equation. Math. Comput. Simul. **79**(7), 2258–2278 (2009)
24. Polyanin, A.D., Zaitsev, V.F.: Handbook of Nonlinear Partial Differential Equations. Chapman & Hall/CRC, Boca Raton, FL (2004)
25. Ruiz-Ramírez, J., Macías-Díaz, J.E.: A finite-difference scheme to approximate non-negative and bounded solutions of a FitzHugh-Nagumo equation. Int. J. Comput. Math. **88**(15), 3186–3201 (2011)
26. Sanfelici, S.: Convergence of the Galerkin approximation of a degenerate evolution problem in electrocardiology. Numer. Methods Partial Differ. Equ. **18**(2), 218–240 (2002)

Local Projection Stabilization for Convection-Diffusion-Reaction Equations on Surfaces

Kristin Simon and Lutz Tobiska

Abstract The numerical solution of convection-diffusion-reaction equations in two and three dimensional domains Ω is thoroughly studied and well understood. Stabilized finite element methods have been developed to handle boundary or interior layers and to localize and suppress unphysical oscillations. Much less is known about convection-diffusion-reaction equations on surfaces $\Gamma = \partial\Omega$. We propose a Local Projection Stabilization (LPS) for convection-diffusion-reaction equations on surfaces based on a linear surface approximation and first order finite elements. Unique solvability of the continuous and discrete problem are established. Numerical test examples show the potential of the proposed method.

1 Introduction

Surface partial differential equations appear in the modeling of several phenomena, e.g. in fluid mechanics, cell biology and material science. A prominent example in fluid mechanics is the presence of surface active agents (surfactants), which modify the surface tension at fluidic interfaces locally. Transport and diffusion of surfactants on an interface can be described by convection-diffusion-reaction equations.

There is a lot of literature treating these type of equations in domains $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, see e.g. [13]. However, much less is known on convection-diffusion-reaction equations on surfaces $\Gamma = \partial\Omega$. The introduction of finite elements for elliptic equations on surfaces can be traced back to Dziuk [6]. Higher order finite elements for the Poisson equation on surfaces have been studied by Demlow [5]. This was extended to coupled diffusion-reaction equations in the bulk and on the surface by Ranner [12] and to parabolic equations on moving surfaces by Dziuk and Elliott [7].

It is well known that boundary and internal layers in the solution of differential equations can lead to unphysical oscillations unless the mesh is fine enough.

K. Simon • L. Tobiska (✉)
Institute for Analysis and Computational Mathematics, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: Kristin.Simon@ovgu.de; Lutz.Tobiska@ovgu.de

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_13

Therefore different stabilization techniques have been developed for convection and/or reaction dominated problems in the bulk. In [11] an unfitted finite element method applied to surface partial differential equations has been analyzed. For stabilization the residual based Streamline Upwind Petrov-Galerkin approach has been used. In this work we consider fitted finite elements and extend the Local Projection Stabilization (LPS) developed in [2, 9, 10] to surface equations.

In Sect. 2 we introduce the weak formulation of the problem and discuss its solvability. The surface approximation, the extension of data, the discrete problem and its solvability are the topic of Sect. 3. Section 4 is devoted to the LPS and the solvability of the stabilized discrete problem. Finally, the potential of the proposed approach is illustrated by three numerical examples in Sect. 5.

In the following, $(\cdot,\cdot)_G$ denotes the $L^2(G)$ inner product on the surface G. We also use the notation $\|.\|_{k,p,G}$ and $|.|_{k,p,G}$ for the standard norm and semi norm on the Sobolev space $W^{k,p}(G)$, respectively. We write shortly $\|.\|_{k,G}$ for $\|.\|_{k,2,G}$ and $|.|_{k,G}$ for $|.|_{k,2,G}$ ($k \in \mathbb{N}$). The norms and semi norms are defined for vectors and matrices in an component wise manner.

2 Problem Formulation

We start with some notations in differential geometry. Let us consider a closed, oriented and non-selfintersecting C^2-surface $\Gamma = \partial\Omega$ of a two or three dimensional domain $\Omega \subset \mathbb{R}^n$, $n = 2, 3$. Using the distance function $\mathrm{dist}(\mathbf{x}, \Gamma) := \inf_{\mathbf{y}\in\Gamma}\{|\mathbf{x} - \mathbf{y}|\}$ the surface is given as zero level of the signed distance function

$$d(\mathbf{x}) := \begin{cases} \mathrm{dist}(\Gamma, \mathbf{x}), & \text{if } \mathbf{x} \notin \Omega, \\ -\mathrm{dist}(\Gamma, \mathbf{x}), & \text{if } \mathbf{x} \in \Omega. \end{cases}$$

The unit outward normal $\mathbf{n}$ on Γ can be expressed as derivative of the signed distance function, i.e. $\mathbf{n} = \nabla d$ on Γ. For a sufficiently smooth function $g : \Gamma \to \mathbb{R}$ we define the surface gradient by

$$\nabla_\Gamma g := \nabla\tilde{g} - (\mathbf{n} \cdot \nabla\tilde{g})\mathbf{n} \quad \text{on } \Gamma,$$

where $\tilde{g}$ is an arbitrary smooth extension of g. It can be shown that the surface gradient on Γ is independent of the extension, see [12], and therefore it is well defined. The Laplace-Beltrami operator is given by $\Delta_\Gamma := \nabla_\Gamma \cdot \nabla_\Gamma$.

We pose the stationary convection-diffusion-reaction equation

$$-\varepsilon\Delta_\Gamma u + \nabla_\Gamma \cdot (\mathbf{w}u) + cu = f \quad \text{on } \Gamma, \tag{1}$$

where $\varepsilon > 0$ is a given constant diffusion coefficient, $\mathbf{w} \in W^{1,\infty}(\Gamma)^n$ is the velocity field, and the reaction coefficient c and the source term f belong to $L^\infty(\Gamma)$ and

$L^2(\Gamma)$, respectively. Having in mind that the surface is not moving we have, that the surface velocity is equal to zero, i.e. $\mathbf{w} \cdot \mathbf{n} = 0$ on Γ. Equation (1) can be written as

$$-\varepsilon \Delta_\Gamma u + \mathbf{w} \cdot \nabla_\Gamma u + \sigma u = f \quad \text{on } \Gamma, \tag{2}$$

where $\sigma := \nabla_\Gamma \cdot \mathbf{w} + c \in L^\infty(\Gamma)$. We will assume that there is a positive constant σ_0 such that

$$\sigma - \frac{1}{2}\nabla_\Gamma \cdot \mathbf{w} \geq \sigma_0 > 0 \quad \text{on } \Gamma. \tag{3}$$

This assumption is an adaption of the standard assumption used in the analysis of convection-diffusion-reaction equations posed in a bounded domain $\Omega \subset \mathbb{R}^n$ [13].

The weak formulation is obtained as usual by multiplying (2) with a test function v, integrating over Γ, and using the integration by parts formula to transfer the highest order derivative terms. For a vector-valued function $\mathbf{f} : \Gamma \to \mathbb{R}^n$ the divergence theorem on surfaces reads [7, Theorem 2.10]

$$\int_\Gamma \nabla_\Gamma \cdot \mathbf{f}\, dS = \int_\Gamma H\mathbf{f} \cdot \mathbf{n}\, dS + \int_{\partial\Gamma} \mathbf{f} \cdot \boldsymbol{\mu}\, dL$$

where H is the sum of principal curvatures and $\boldsymbol{\mu}$ is the conormal vector, which is normal to $\partial\Gamma$ and tangent to Γ. Since Γ is assumed to be closed, we have $\partial\Gamma = \emptyset$ and the second term on the right hand side vanishes. Setting $\mathbf{f} = v\nabla_\Gamma u$ and using $\mathbf{n} \cdot \nabla_\Gamma u = 0$ we get

$$(\nabla_\Gamma \cdot \nabla_\Gamma u, v)_\Gamma + (\nabla_\Gamma u, \nabla_\Gamma v)_\Gamma = (H\mathbf{n} \cdot \nabla_\Gamma u, v)_\Gamma = 0.$$

Then, the weak formulation of (2) reads:

Problem 1 Find $u \in H^1(\Gamma)$ such that $a(u, v) = (f, v)_\Gamma$ for all $v \in H^1(\Gamma)$.

Here, the continuous bilinear form $a : H^1(\Gamma) \times H^1(\Gamma) \to \mathbb{R}$ is given by

$$a(u, v) := \varepsilon(\nabla_\Gamma u, \nabla_\Gamma v)_\Gamma + (\mathbf{w} \cdot \nabla_\Gamma u, v)_\Gamma + (\sigma u, v)_\Gamma.$$

Taking the identity

$$(\mathbf{w} \cdot \nabla_\Gamma v, v)_\Gamma = \frac{1}{2}(\mathbf{w} \cdot \nabla_\Gamma v^2, 1)_\Gamma = -\frac{1}{2}(\nabla_\Gamma \cdot \mathbf{w}, v^2)_\Gamma + (H\,\mathbf{w} \cdot \mathbf{n}, v^2)_\Gamma$$

into consideration and using $\mathbf{w} \cdot \mathbf{n} = 0$ on Γ we obtain under condition (3)

$$a(v, v) = \varepsilon \|\nabla_\Gamma v\|^2_{0,\Gamma} + \left(\sigma - \frac{1}{2}\nabla_\Gamma \cdot w, v^2\right)_\Gamma \geq \varepsilon \|\nabla_\Gamma v\|^2_{0,\Gamma} + \sigma_0 \|v\|^2_{0,\Gamma}$$

for all $v \in H^1(\Gamma)$, i.e. the coercivity of the bilinear form. Unique solvability of Problem 1 follows from the Lax-Milgram theorem.

3 Discretization

3.1 Surface Approximation

In order to formulate the discrete problem we need an approximation Γ_h of the given surface Γ. We use a polyhedral mesh consisting of elements K, lines in 2d or triangles in 3d, where the vertices of all elements are located at the surface Γ. The diameter of an element K is denoted by h_K, the mesh size h is set to $h = \max_K h_K$ and the mesh is assumed to be shape regular. A mesh constructed in this way is a natural linear interpolation of the given surface.

Recognize that $\Gamma_h \not\subseteq \Gamma$. Thus, all data given on Γ, i.e. the velocity field $\mathbf{w}$, the reaction coefficient σ and the source term f, are not defined on Γ_h. Similarly, the surface operators differ on the given and the approximated surface.

3.2 Extension of Data

To transfer the given data defined on Γ to Γ_h we introduce a projection $\mathbf{p} : U \to \Gamma$ from a neighbourhood U of Γ onto the nearest point on Γ via

$$\mathbf{p}(\mathbf{x}) = \mathbf{x} - d(\mathbf{x})\mathbf{n}(\mathbf{p}(\mathbf{x})) = \mathbf{x} - d(\mathbf{x})\nabla d(\mathbf{x}) \quad \text{for all } \mathbf{x} \in U.$$

We choose U small enough such that the projection is well defined. Then, an extension of functions $g : \Gamma \to \mathbb{R}$ into the neighbourhood U can be given by

$$g^e(\mathbf{x}) := g(\mathbf{p}(\mathbf{x})) \quad \text{for } \mathbf{x} \in U.$$

Further, we assume the element size h to be small enough such that $\Gamma_h \subset U$. Then, the projection $\mathbf{p} : \Gamma_h \to \Gamma$ is bijectiv.

3.3 Discrete Problem

We introduce a finite dimensional, continuous finite element space $V_h \subset H^1(\Gamma_h)$. Then, the standard Galerkin discretization of Problem 1 reads:

Problem 2 Find $u_h \in V_h$ such that $a_h(u_h, v_h) = (f^e, v_h)_{\Gamma_h}$ for all $v_h \in V_h$.

Here, the discrete bilinear form $a_h : V_h \times V_h \to \mathbb{R}$ is given by

$$a_h(u_h, v_h) = \varepsilon(\nabla_{\Gamma_h} u_h, \nabla_{\Gamma_h} v_h)_{\Gamma_h} + (\mathbf{w}^e \cdot \nabla_{\Gamma_h} u_h, v_h)_{\Gamma_h} + (\sigma^e u_h, v_h)_{\Gamma_h}.$$

Lemma 1 *Assuming h to be sufficiently small, then for all* $v_h \in V_h$

$$a_h(v_h, v_h) \geq \varepsilon \|\nabla_{\Gamma_h} v_h\|^2_{0,\Gamma_h} + \frac{\sigma_0}{2} \|v_h\|^2_{0,\Gamma_h}.$$

Consequently, Problem 2 has a unique solution.

Proof We try to repeat the ideas used for the continuous bilinear form and get

$$a_h(v_h, v_h) = \varepsilon \|\nabla_{\Gamma_h} v_h\|^2_{0,\Gamma_h} + \frac{1}{2} \left(\mathbf{w}^e \cdot \nabla_{\Gamma_h} v_h^2, 1\right)_{\Gamma_h} + \left(\sigma^e, v_h^2\right)_{\Gamma_h}.$$

Now, we apply the integration by parts formula on the convective term in an elementwise manner to obtain

$$\begin{aligned}\left(\mathbf{w}^e \cdot \nabla_{\Gamma_h} v_h^2, 1\right)_{\Gamma_h} &= -\sum_K \left(\nabla_{\Gamma_h} \cdot \mathbf{w}^e, v_h^2\right)_K + \sum_K \left(H\, \mathbf{w}^e \cdot \mathbf{n}_K, v_h^2\right)_K + \sum_K \sum_{E \subset \partial K} \left\langle \mathbf{w}^e \cdot \boldsymbol{\mu}_K, v_h^2 \right\rangle_E \\ &= -\left(\nabla_{\Gamma_h} \cdot \mathbf{w}^e, v_h^2\right)_{\Gamma_h} + \sum_E \left\langle [\mathbf{w}^e \cdot \boldsymbol{\mu}]_E, v_h^2 \right\rangle_E\end{aligned}$$

Here, we used $H = 0$ on each K and wrote $[\cdot]_E$ for the jump across the face $E \subset \partial K$. Summarizing we obtain

$$\begin{aligned}a_h(v_h, v_h) = \varepsilon \|\nabla_{\Gamma_h} v_h\|^2_{0,\Gamma_h} + \left(\sigma^e - \frac{1}{2}(\nabla_\Gamma \cdot \mathbf{w})^e, v_h^2\right)_{\Gamma_h} + \frac{1}{2} \sum_E \left\langle [\mathbf{w}^e \cdot \boldsymbol{\mu}]_E, v_h^2 \right\rangle_E \\ + \frac{1}{2} \left((\nabla_\Gamma \cdot \mathbf{w})^e - \nabla_{\Gamma_h} \cdot \mathbf{w}^e, v_h^2\right)_{\Gamma_h}.\end{aligned}$$

The jump of the conormal vectors across a face $E = \partial K \cap \partial K'$ behaves like $\mathcal{O}(h_E)$, however its projection into the tangential plane like $\mathcal{O}(h_E^2)$ [11]. Then, taking $\mathbf{w}^e = (Id - \mathbf{n} \otimes \mathbf{n})\mathbf{w}^e$ and a discrete trace inequality into consideration we conclude

$$\left| \sum_E \left\langle [\mathbf{w}^e \cdot \boldsymbol{\mu}]_E, v_h^2 \right\rangle_E \right| \leq C \|\mathbf{w}^e\|_{0,\infty,\Gamma_h} \sum_E h_E^2 \|v_h\|^2_{0,E} \leq Ch \|\mathbf{w}^e\|_{0,\infty,\Gamma_h} \|v_h\|^2_{0,\Gamma_h}. \tag{4}$$

A detailed study shows that

$$\|(\nabla_\Gamma \cdot \mathbf{w})^e - \nabla_{\Gamma_h} \cdot \mathbf{w}^e\|_{0,\infty,\Gamma_h} \leq Ch \|\nabla_\Gamma \mathbf{w}\|_{0,\infty,\Gamma}. \tag{5}$$

Applying (3), (4), and (5) we get

$$a_h(v_h, v_h) \geq \varepsilon \|\nabla_{\Gamma_h} v_h\|^2_{0,\Gamma_h} + \sigma_0 \|v_h\|^2_{0,\Gamma_h} - Ch\|\mathbf{w}\|_{1,\infty,\Gamma} \|v_h\|^2_{0,\Gamma_h},$$

from which the first statement of the lemma follows. The unique solvability of Problem 2 follows from the Lax-Milgram theorem.

4 Local Projection Stabilization

Now, we consider the one level Local Projection Stabilization in the case that convection and/or reaction dominates diffusion. Beside the ansatz space $V_h \subset H^1(\Gamma_h)$ a discontinuous projection space $D_h = \oplus_K D_h(K)$ is introduced on the same mesh. Let $\pi_{h,K} : L^2(K) \to D_h(K)$ denote the local L^2-projection and $\kappa_{h,K} : L^2(K) \to L^2(K)$ the fluctuation operator given by $\kappa_{h,K} := id - \pi_{h,K}$.

Now, we can introduce the stabilization term

$$S_h(u_h, v_h) = \sum_K \alpha_K (\kappa_{h,K} \nabla_{\Gamma_h} u_h, \kappa_{h,K} \nabla_{\Gamma_h} v_h)_K$$

with $\alpha_K > 0$ being user chosen stabilization parameters. The stabilized formulation of the Problem 2 reads

Problem 3 Find $u_h \in V_h$ such that $a_h(u_h, v_h) + S_h(u_h, v_h) = (f^e, v_h)_{\Gamma_h}$ for all $v_h \in V_h$.

The associated mesh-dependent norm is given by

$$|||u||| := \left(\varepsilon \|\nabla_{\Gamma_h} u\|^2_{0,\Gamma_h} + \sigma_0 \|u\|^2_{0,\Gamma_h} + \sum_K \alpha_K \|\kappa_h \nabla_{\Gamma_h} u\|^2_{0,K} \right)^{1/2}.$$

Lemma 2 *Assuming h to be sufficiently small, then for all $v_h \in V_h$*

$$a_h(v_h, v_h) + S_h(v_h, v_h) \geq \frac{1}{2} |||v_h|||^2$$

Consequently, Problem 3 has a unique solution.

Proof Using Lemma 1 and the definition of the stabilizing term S_h we get

$$a_h(v_h, v_h) + S_h(v_h, v_h) \geq \varepsilon \|\nabla_\Gamma v_h\|^2_{0,\Gamma_h} + \frac{\sigma_0}{2} \|v_h\|^2_{0,\Gamma_h} + S_h(v_h, v_h) \geq \frac{1}{2} |||v_h|||^2.$$

The unique solvability of Problem 3 follows from the Lax-Milgram theorem.

The convergence properties of the solution of Problem 3 depend on the choice of ansatz and projection space, on the approximation order of the discrete surface Γ_h,

and on the size of the stabilization parameters α_K. Let V_h be the space of continuous, piecewise linear functions enriched with piecewise quadratic ($n = 2$) or piecewise cubic ($n = 3$) bubble functions, D_h the space of discontinuous, piecewise constant functions, Γ_h be the continuous, piecewise linear approximation of Γ, and $\alpha_K \sim h_K$. Then, the solution u of Problem 1 and the solution u_h of Problem 3 satisfy the following error estimate [14]

$$|||u^e - u_h||| \leq C \left(\varepsilon^{1/2} + h^{1/2}\right) h \left(\|\nabla_\Gamma^2 u\|_{0,\Gamma} + \|f\|_{0,\Gamma}\right).$$

5 Numerical Results

The numerical tests reported in this section have been performed with an in-house code written in Julia [3]. The first two test examples concern diffusion-convection-reaction equations on a 1d hypersurface embedded in $\mathbb{R}^2$ whereas in the last example a diffusion-reaction equation on a 2d hypersurface embedded in $\mathbb{R}^3$ is considered.

5.1 Testcase 1: Layer in the First Derivative

We consider the unit circle Γ embedded in $\mathbb{R}^2$ and Eq. (2) with the velocity field $\mathbf{w}(\mathbf{x}) = 2(x_2, -x_1)$ describing a clockwise rotation and the reaction coefficient $\sigma = 1$. The right hand side f is set to the arc length $s(\mathbf{x})$ measured along Γ clockwise starting at $(1, 0)$, i.e.

$$f(\mathbf{x}) = s(\mathbf{x}) = \begin{cases} \arccos(x_1) & \text{for } x_2 \leq 0, \\ \pi + \arccos(-x_1) & \text{for } x_2 > 0. \end{cases} \tag{6}$$

The exact solution $u : [0, 2\pi] \to \mathbb{R}$ of this problem is given as a function of s

$$u(s) = 2\pi \left[\frac{\lambda_2}{\lambda_2 - \lambda_1} \frac{\exp(\lambda_1 s)}{1 - \exp(2\pi\lambda_1)} - \frac{\lambda_1}{\lambda_2 - \lambda_1} \frac{\exp(\lambda_2 s)}{1 - \exp(2\pi\lambda_2)} \right] + s - 2 \tag{7}$$

where

$$\lambda_1 = \frac{1 + \sqrt{1 + \varepsilon}}{\varepsilon}, \qquad \lambda_2 = -\frac{1}{1 + \sqrt{1 + \varepsilon}}$$

and extended 2π-periodic. The solution u does not develop a layer for $\varepsilon \to 0$, however a layer appears in the first derivative u' at $s = 0$, as shown in Fig. 1. Thus, the solution u has no strong but a weak layer.

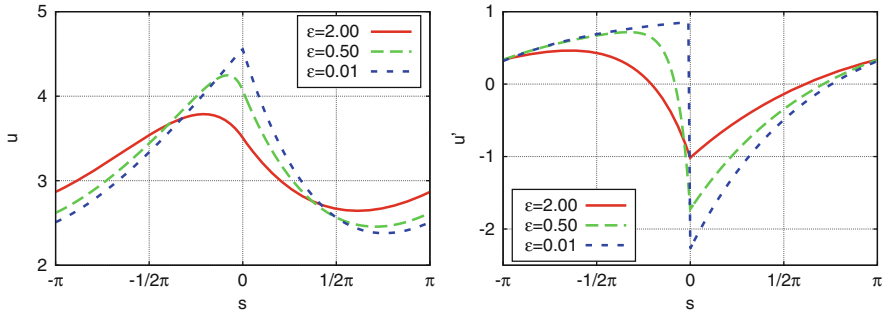

Fig. 1 The function u (*left*) and its first derivative u' (*right*) plotted over the arc length s

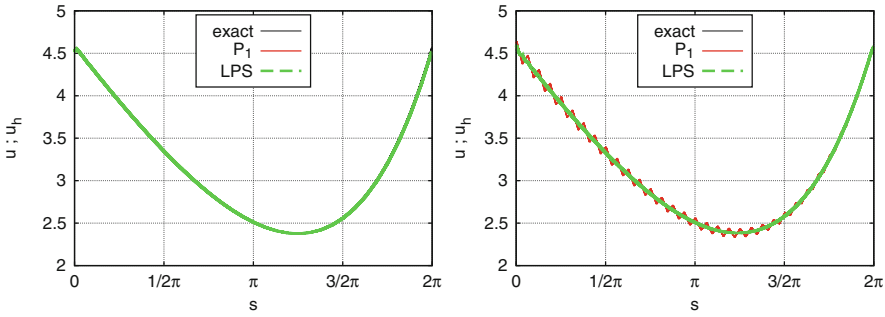

Fig. 2 Comparison of the numerical solutions u_h with the exact solution u plotted over the arc length s on a fitted (*left*) and unfitted (*right*) grid

We solve the problem for $\varepsilon = 10^{-8}$ numerically by standard piecewise linear finite elements and the LPS as described in Sect. 4 with the stabilization parameters $\alpha_K = \alpha h_K$ and $\alpha = 0.1$. Figure 2 shows the results for two different meshes. If the point of discontinuity of f is a grid point even standard piecewise linear finite elements work well without spurious oscillations (Fig. 2 left). If it is not a grid point, oscillations appear which are damped out by LPS (Fig. 2 right).

5.2 Testcase 2: Exponential Layer in the Solution

On the unit circle Γ we look at the concentration distribution of a substance transported from $(0, 1)$ to $(0, -1)$ by the flow field $\mathbf{w}(\mathbf{x}) = x_1\,(x_2, -x_1)$. The reaction coefficient σ is set to $\sigma(\mathbf{x}) = x_2 + |x_2|$ and the right hand side f is chosen such that the exact solution is given by

$$u\,(\mathbf{x}) = 1 - \exp\left(-\frac{1 + x_2}{\varepsilon}\right).$$

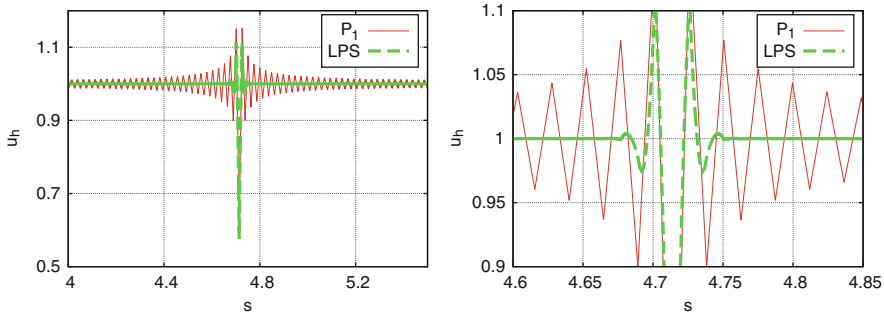

Fig. 3 Standard linear (P1) and stabilized (LPS) finite element solution plotted over the arc length. Overview (*left*) and zoom into the layer region (*right*)

The exact solution has a point layer at $(0, -1)$, which leads to strong numerical oscillations for standard piecewise linear finite elements in the convection dominated case ($\varepsilon = 10^{-6}$, $h_K = 0.01227$), see the P1 graph in Fig. 3. The LPS graph shows the numerical solution of the stabilized problem using a stabilization parameter $\alpha_K = \alpha h_K$ with $\alpha = 0.0045$. As shown in Fig. 3, LPS suppresses and localize the oscillations to a few elements around the layer.

Next an array of computations has been performed to study the stabilization effect depending on the user chosen constant α. For convection-diffusion equations in a bulk it is known that starting with the standard Galerkin piecewise linear finite element method on simplices enriched by bubble functions and eliminating the bubble part yields the streamline diffusion method [1, 4]. Unfortunately, the symmetric version of the bubble generates the streamline diffusion method with a parameter suitable for the diffusion dominated but not for the convection dominated case. It has been shown in [8] that starting with the LPS approach and eliminating the bubble part yields also the streamline-diffusion method, however, with a parameter appropriate for the convection dominated case.

In Fig. 4 we present the stabilized finite element solutions and their bubble eliminated linear parts for decreasing stabilization parameter α. For larger values of α the linear part tends to oscillate whereas for smaller values the eliminated LPS tends to smear out the layer. This behaviour is similar to that known from two-point boundary value problems reported in [15].

In Fig. 5 the results for different values of ε (left) and different mesh refinement levels (right) are shown. They indicate that the 'optimal' value of α is nearly independent of the parameter ε and the mesh size h. However, as a comparison with testcases 1 and 3 demonstrates, the choice of an 'optimal' value of α depends primarily on the problem, in particular on the type of the layer.

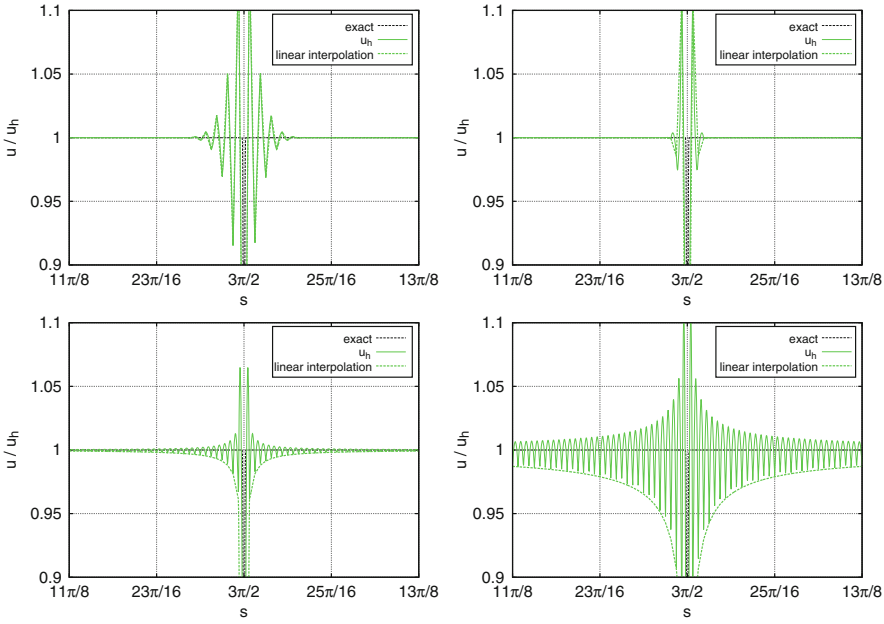

Fig. 4 Stabilized (LPS) finite element solutions u_h and their linear parts (linear interpolation) for $\alpha = 0.045$, $\alpha = 0.0045$, $\alpha = 0.00045$, $\alpha = 0.000045$ (*top left to bottom right*)

5.3 Testcase 3: Layer on a Sphere

We consider a diffusion-reaction equation with a discontinuous right hand side on a unit sphere Γ embedded in $\mathbb{R}^3$. We set the diffusion parameter $\varepsilon = 10^{-8}$, the reaction parameter $\sigma = 1$, and the right hand side $f(\mathbf{x}) = \text{sign}(x_3)$. For $\varepsilon \to 0$ the exact solution u tends to f away from the equatorial line $x_3 = 0$.

We solve this problem using standard linear finite elements and the LPS stabilized method with $\alpha_K = \alpha h_K$ and $\alpha = 4.5 \cdot 10^{-5}$. In Fig. 6 (left) a layer along the equator of the sphere for the unstabilized method indicated by the dark red and dark blue regions occur. For the stabilized solution, compare Fig. 6 (right), these regions are less pronounced.The stabilizing effect of LPS is clearly visible by cutting the sphere along a meridian, see Fig. 7.

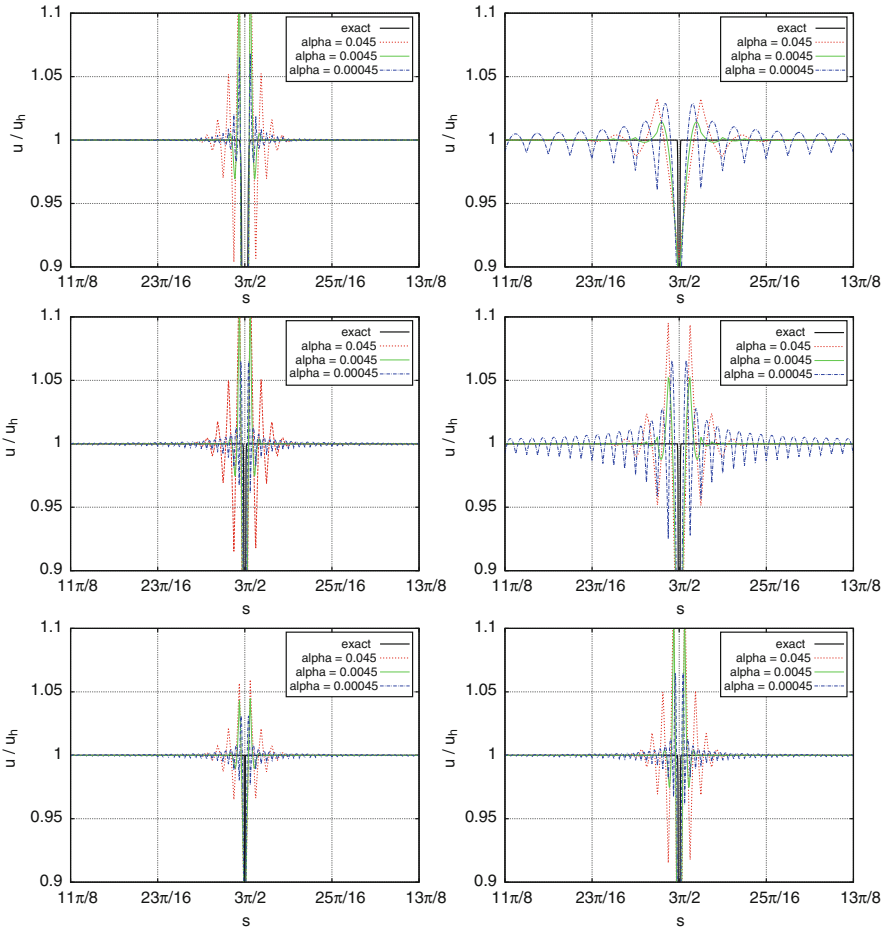

Fig. 5 Stabilized (LPS) finite element solutions for refinement level $l = 4$ and $\varepsilon = 10^{-5}$, $\varepsilon = 10^{-6}$, $\varepsilon = 10^{-7}$ (*top to bottom left*) as well as $\varepsilon = 10^{-6}$ and $l = 2$, $l = 3$, $l = 4$ (*top to bottom right*)

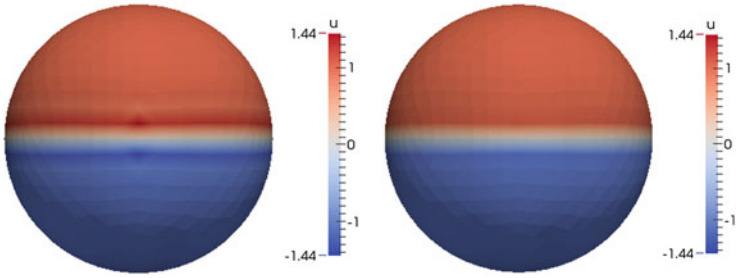

Fig. 6 Unstabilized (*left*) and stabilized (*right*) solution

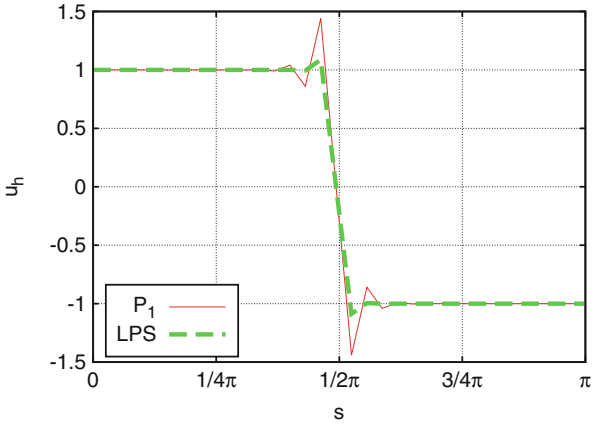

Fig. 7 Standard linear (P1) and stabilized (LPS) finite element solution plotted over the arc length s along a meridian of the sphere starting at (0,0,1)

Acknowledgements The authors wish to thank the German Research Foundation (DFG) for financial support within the Priority Programm SPP 1506 "Transport Processes at Fluidic Interfaces" with the project To143/11-2 and within the graduate program Micro-Macro-Interactions in Structured Media and Particle Systems (GK 1554).

References

1. Baiocchi, C., Brezzi, F., Franca, L.P.: Virtual bubbles and Galerkin-least-squares type methods (Ga.L.S.). Comput. Methods Appl. Mech. Eng. **105**(1), 125–141 (1993). doi:10.1016/0045-7825(93)90119-I, http://dx.doi.org/10.1016/0045-7825(93)90119-I
2. Becker, R., Braack, M.: A finite element pressure gradient stabilization for the Stokes equations based on local projections. Calcolo **38**(4), 173–199 (2001)
3. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. CoRR **abs/1411.1607** (2014). http://arxiv.org/abs/1411.1607
4. Brezzi, F., Russo, A.: Choosing bubbles for advection-diffusion problems. Math. Models Methods Appl. Sci. **4**(4), 571–587 (1994). doi:10.1142/S0218202594000327, http://dx.doi.org/10.1142/S0218202594000327
5. Demlow, A.: Higher-order finite element methods and pointwise error estimates for elliptic problems on surfaces. SIAM J. Numer. Anal. **47**(2), 805–827 (2009). doi:10.1137/070708135, http://dx.doi.org/10.1137/070708135
6. Dziuk, G.: Finite elements for the Beltrami operator on arbitrary surfaces. In: Partial Differential Equations and Calculus of Variations. Lecture Notes in Mathematics, vol. 1357, pp. 142–155. Springer, Berlin (1988)
7. Dziuk, G., Elliott, C.M.: Finite element methods for surface PDEs. Acta Numer. **22**, 289–396 (2013)
8. Ganesan, S., Tobiska, L.: Stabilization by local projection for convection-diffusion and incompressible flow problems. J. Sci. Comput. **43**(3), 326–342 (2010)
9. Matthies, G., Skrzypacz, P., Tobiska, L.: A unified convergence analysis for local projection stabilisations applied to the Oseen problem. M2AN Math. Model. Numer. Anal. **41**(4), 713–742 (2007)

10. Matthies, G., Skrzypacz, P., Tobiska, L.: Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. Electron. Trans. Numer. Anal. **32**, 90–105 (2008)
11. Olshanskii, M.A., Reusken, A., Xu, X.: A stabilized finite element method for advection-diffusion equations on surfaces. IMA J. Numer. Anal. **34**(2), 732–758 (2014)
12. Ranner, T.: Computational surface partial differential equations. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of Warwick (United Kingdom) (2013)
13. Roos, H.G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems. Springer Series in Computational Mathematics, vol. 24, 2nd edn. Springer, Berlin (2008)
14. Simon, K., Tobiska, L.: Local projection stabilization for convection-diffusion-reaction equations on linear approximated surfaces. Technical Report 2017–04, Department of Mathematics, Otto von Guericke University (2017)
15. Tobiska, L.: On the relationship of local projection stabilization to other stabilized methods for one-dimensional advection-diffusion equations. Comput. Methods Appl. Mech. Eng. **198**(5–8), 831–837 (2009)

A Comparison Study of Parabolic Monge-Ampère Equations Adaptive Grid Methods

Mohamed H.M. Sulman

Abstract We consider two recently developed adaptive grid methods for solving time dependent partial differential equations (PDEs) in higher dimensions. These methods compute the adaptive grid based on solving an optimal mass transport problem also known as Monge-Kantorovich problem (MKP). The optimal solution of the MKP is reduced to solving Monge-Ampère equation and is known to have some nice theoretical properties that are desirable for the mesh adaptation. However, these two adaptive grid methods solve the Monge-Ampère equation differently and they are distinctly different in their approaches for computing the adaptive mesh over time. A comparison study to address these various distinctions between the two methods is presented. Several numerical experiments are conducted to illustrate the main differences between the two methods in terms of their mesh quality and performances.

1 Introduction

Adaptive grid methods are efficient numerical techniques for computing the solutions of partial differential equations (PDEs) with higher accuracy, in particular when the solutions of the PDEs have large variations or shocks in some regions of their physical domains. For the adaptive methods discussed here, the adaptive mesh is obtained by defining a coordinate transformation $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\xi}, t)$ from a computational domain Ω_c to a physical domain Ω. The grid nodes are redistributed continuously in time by the mapping $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\xi}, t)$ so that they are concentrated in regions of interest. For one spatial dimensional PDEs, a class of adaptive grid methods known as moving mesh partial differential equations methods (MMPDE) [10] have been developed based on the so-called equidistribution principle, a

M.H.M. Sulman (✉)
Department of Mathematics and Statistics, Wright State University, 3640 Col. Glenn Highway, Dayton, OH 45435, USA
e-mail: mohamed.sulman@wright.edu

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_14

measure of the solution variation is equally distributed over each sub interval [5],

$$\rho(\boldsymbol{x})\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\xi}} = c, \quad \boldsymbol{\xi} \in \Omega_c,\ \boldsymbol{x} \in \Omega, \tag{1}$$

where $\rho(\boldsymbol{x})$ is the density (or monitor) function defined to measure the variations in the solution of the given physical problem. In one spatial dimension, one can use the equidistribution principle to compute the adaptive mesh. However, in higher dimensions the condition (1) alone is insufficient to uniquely determine the adaptive mesh.Therefore, some additional conditions are required for the mesh adaptation in higher dimensions. In [6, 8, 9], the mapping $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\xi}, t)$ in higher dimensions is obtained as the solution of the gradient flow equation of a quadratic functional. Huang [7] develops a variational approach for mesh adaptation by minimizing a functional that combines two functionals associated with the mesh isotropy and equidistribution properties. More recently, two adaptive mesh methods [14] and [4] have been introduced by extending the equidistribution principle (1) for higher dimensional problems. These methods are derived based on solving the L^2 Monge-Kantorovich problem (MKP) [11, 13].

In this work we present a comparison study of the parabolic Monge-Ampère equations adaptive grid methods of [14] and [4]. The two methods solve the L^2 MKP based on solving two different types of parabolic Monge-Ampère equations, and they use two different strategies for computing the adaptive mesh at each time step. In this paper we investigate the main differences between these two parabolic Monge-Ampère (PMA) methods and assess the mesh quality and overall performance of each method. For the purpose of the comparison, here we label the method of [14] as PMA-Log and the method of [4] as PMA-Sqrt.

The rest of the paper is organized as follows. In Sect. 2, we briefly introduce the L^2 Monge-Kantorovich problem. In Sect. 3, we describe the parabolic Monge-Ampère adaptive grid methods PMA-Log and PMA-Sqrt. In Sect. 4, we present several numerical experiments to illustrate the differences between PMA-Log method and PMA-Sqrt method. In Sect. 5, we give a discussion on the results and some concluding remarks.

2 The L^2 Monge–Kantorovich Problem

The L^2 Monge–Kantorovich problem is stated as follows: Given two positive and bounded density functions ρ_0 and ρ_1 of equal masses defined on $\Omega_c \subset \mathbb{R}^d$ and $\Omega \subset \mathbb{R}^d$ respectively. Find a mapping $\boldsymbol{x} = \phi(\boldsymbol{\xi})$, $\boldsymbol{\xi} \subset \Omega_c, \boldsymbol{x} \in \Omega$, that transfers ρ_0 to ρ_1 and minimizes the transport cost

$$C(\phi) = \int_{\Omega_c} |\phi(\boldsymbol{\xi}) - \boldsymbol{\xi}|^2 \rho_0(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \tag{2}$$

where $|\cdot|$ is the Euclidean norm defined on the space $\mathbb{R}^d$. The map ϕ realizes the transfer of ρ_0 to ρ_1 if $\int_{\Omega_c} \rho_0(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\Omega} \rho_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and this leads to

$$\rho_1(\boldsymbol{x})\det\left(\partial \boldsymbol{x}/\partial \boldsymbol{\xi}\right) = \rho_0(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \Omega_c,\ \boldsymbol{x} \in \Omega, \tag{3}$$

where $\det(\partial \boldsymbol{x}/\partial \boldsymbol{\xi})$ is the determinant of the Jacobian matrix. Suppose that Ω_c and Ω are bounded, connected and convex sets in $\mathbb{R}^d$, $d \geq 2$, and let ρ_0 and ρ_1 be strictly positive and bounded. Then there is a unique mapping ϕ that minimizes (2) with the constraint (3) and it is given as (for more details see [2, 3])

$$\boldsymbol{x} = \phi(\boldsymbol{\xi}) = \nabla\Psi(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \Omega_c, \tag{4}$$

for some convex potential Ψ.

From (3) and (4) we obtain the Monge-Ampère equation

$$\rho_1\left(\nabla\Psi(\boldsymbol{\xi})\right)\det D^2\Psi(\boldsymbol{\xi}) = \rho_0(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \Omega_c, \tag{5}$$

where $D^2\Psi$ is the Hessian matrix of Ψ, in two dimensions $D^2\Psi = \Psi_{\xi\xi}\Psi_{\eta\eta} - \Psi^2_{\xi\eta}$.

3 The Parabolic Monge-Ampére Adaptive Grid Methods

The minimizer of the cost functional (2) is the closest mapping to the identity. Moreover, if we treat ρ_1 as the monitor function ρ and set ρ_0 as a positive constant, then the constraint (3) is equivalent to the equidistribution principle (1). The parabolic Monge-Ampère adaptive grid methods are derived based on solving the Monge-Ampère equation (5) which is obtained from the equidistribution condition (3). A popular choice of the mesh density function is the arc-length function

$$\rho = \sqrt{1 + \alpha|\nabla u|^2} \tag{6}$$

where u is the solution of the physical problem and α is a user defined parameter, in our computations we set $\alpha = 5$. Notice that the equidistribution of (6) enforces the clustering of the grid points in regions of large solution variations.

3.1 PMA-Log Adaptive Grid Method

The PMA-Log [14] method computes the solution of Monge-Ampère equation (5) as a steady-state solution of the parabolic Monge-Ampère equation

$$\frac{\partial \Psi}{\partial \tau} = \log\left(\rho_1(\nabla\Psi)\det D^2\Psi\right) \tag{7}$$

starting from Ψ_0 given by the initial condition

$$\Psi(\boldsymbol{\xi},0) = \Psi_0(\boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\xi}\cdot\boldsymbol{\xi}^T \tag{8}$$

and with the Neumann boundary conditions

$$\nabla\Psi\cdot\mathbf{n} = \boldsymbol{\xi}\cdot\mathbf{n}, \quad \text{for } \boldsymbol{\xi}\in\partial\Omega_c, \tag{9}$$

where $\partial\Omega_c$ is the boundary of Ω_c and $\mathbf{n}$ is the unit normal to $\partial\Omega_c$. The results of the existence and uniqueness of the steady state solution of (7) can be found in [15].

We use the central finite difference scheme for the spatial discretization of (7) and then apply Euler method for the time integration. Given the mesh $\boldsymbol{x}^n(\boldsymbol{\xi})$ and the physical solution u^n at time level n, then $\boldsymbol{x}^{n+1}$ and u^{n+1} are computed as follows:

1. Set $\rho_1 = \sqrt{1+\alpha\nabla u^n|^2}$, and integrate (7) for the steady state solution Ψ^∞
2. Set $\boldsymbol{x}^{n+1}(\boldsymbol{\xi}) = \nabla\Psi^\infty(\boldsymbol{\xi}),\ \boldsymbol{\xi}\in\Omega_c$
3. Integrate the physical PDE for one time level using $\boldsymbol{x}^{n+1}$ to obtain u^{n+1}
4. Set $n = n+1$ and go to step 1

where $\nabla\Psi^\infty(\boldsymbol{\xi})$ is computed using central finite difference approximation. The iteration is stopped using the criterion

$$\left\|\nabla\Psi^{n+1}-\nabla\Psi^n\right\|_2 = \left(\int_\Omega \left|\nabla\Psi^{n+1}(\boldsymbol{\xi})-\nabla\Psi^n(\boldsymbol{\xi})\right|^2 d\boldsymbol{\xi}\right)^{1/2} \le \text{TOL}, \tag{10}$$

where TOL is some specified tolerance. Notice that the choice of the tolerance affects the computational time mainly in computing the initial mesh since after then the previously computed mesh is used as an initial mesh.

3.2 PMA-Sqrt Adaptive Grid Method

The PMA-Sqrt method [4] solves (5) by introducing a parabolic Monge-Ampère equation given in two spatial dimensions as

$$\hat{\epsilon}(\mathrm{I}-\nu\triangle)\Psi_\tau = \left(\rho(\nabla\Psi,\tau)\det D^2\Psi\right)^{1/2} \tag{11}$$

where $\hat{\epsilon}$ is a time smoothing parameter and ν is a spatial smoothing parameter. In the computation, (11) is solved subject to the initial and boundary conditions (8) and (9). We use the central finite difference scheme for the spatial discretization of (11), and then apply Euler method for the time integration. Given the initial solution of the physical problem u^0, the computation of the adaptive mesh and the solution of the physical problem proceeds as follows:

1. Set n=0 and $\rho = \sqrt{1 + \alpha \nabla u^n}$, then integrate (11) for the pseudo time $0 < \tau \leq T$ where T is a fixed time, and set $\boldsymbol{x}^n = \nabla \Psi(\boldsymbol{\xi}, T)$.
2. Integrate the physical problem for one time step to obtain the solution u^{n+1}
3. Integrate (11) for one real time step to obtain Ψ^{n+1}, set $\boldsymbol{x}^{n+1} = \nabla \Psi^{n+1}$
4. Set $n = n + 1$ and go to step 2.

Notice that PMA-Sqrt method requires solving a Poisson problem at each time step. Moreover, the PMA-Sqrt method computes an approximate equidistributed mesh at the initial time and the subsequent meshes are obtained by integrating (11) for one time step. While the PMA-Log method computes the adaptive mesh at each time step by solving (7) to the steady state so that the computed mesh is equidistributed at all time levels.

4 Numerical Experiments

We present several numerical experiments to illustrate the differences in the performance and mesh quality of the PMA-Log and PMA-Sqrt methods. For all the tested problems, the physical PDE is discretized in the computational domain Ω_C using central finite difference scheme for the spatial derivatives, and Matlab ode113 solver is used for the time integration. We set $\nu = 1$ in all the computations. The choice of $\hat{\epsilon}$ affects the accuracy of the MKP-Sqrt method, thus we choose $\hat{\epsilon}$ that works best for each problem, precisely we use $\hat{\epsilon} = .01$ in Example 1 and Example 2, and $\hat{\epsilon} = 1$ in Example 3. The tolerance for (10) is chosen so that the mesh is sufficiently equidistributed, in the computation Tol $= 10^{-5}$. All the computations are run on a machine with 2.3 GHz intel Core i7 processor and 16 GB memory.

4.1 Example 1

In this example, we consider the parabolic equation

$$\frac{\partial u}{\partial t} = \epsilon \Delta u - \frac{\partial u}{\partial y} + x(1-x) + 2\epsilon \left(y - \frac{1 - \exp(y/\epsilon)}{1 - \exp(1/\epsilon)} \right), \ (x, y) \in \Omega = (0, 1)^2 \tag{12}$$

with homogeneous initial and Dirichlet boundary conditions. The steady-state solution of this problem is given as

$$u^{\infty}(x, y) = x(1-x) \left(y - \frac{1 - \exp(y/\epsilon)}{1 - \exp(1/\epsilon)} \right). \tag{13}$$

The steady-state solution (13) has a boundary layer along $y = 1$. We solve (12) over a time interval $[0, 5]$ that is sufficiently large to obtain a good approximation of the

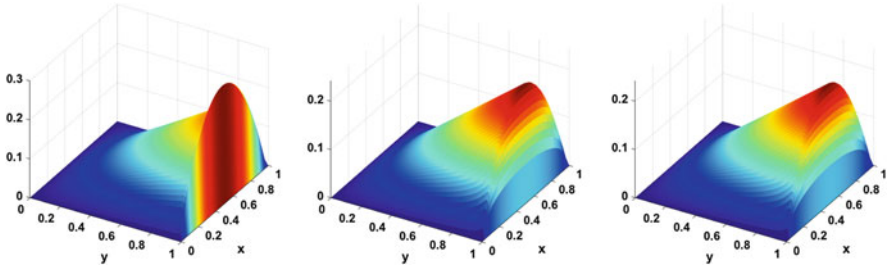

Fig. 1 Example 1. Solution of (12) at $t = 5$ computed using uniform fixed grid (UFG), PMA-Log method and PMA-Sqrt method

steady-state solution at $t = 5$. Figure 1 shows the solution at time $t = 5$ computed using uniform fixed grid (UFG) of size 65×65, and PMA methods for $\epsilon = 5 \times 10^{-3}$. Notice that the uniform grid method doesn't resolve the boundary layer accurately. Figure 2 shows the adaptive mesh computed by the PMA methods. Table 1 reports the results of the errors and cpu time for the whole computation from $t = 0$ to $t = 5$. To compare the convergence rate of the solutions, in Fig. 3 we plot the L^2 error vs number of the grid nodes. Notice that the numerical solution of (12) converges faster on meshes produced by PMA-Log method than those with PMA-Sqrt method which indicates a better accuracy when using PMA-Log.

4.2 Example 2

In this example, we consider the two dimensional Burgers' equation

$$\frac{\partial u}{\partial t} = \epsilon \Delta u - u\frac{\partial u}{\partial x} - u\frac{\partial u}{\partial y}, \quad (x, y) \in (0, 1)^2, \quad 0.25 \le t \le 1.25. \tag{14}$$

The initial and Dirichlet boundary conditions are defined from the exact solution

$$u(x, y, t) = 1/(1 + \exp((x + y - t)/(2\epsilon)).$$

We take $\epsilon = 5 \times 10^{-3}$ and compute the solution of (14) with the PMA methods on a grid of size 65×65. Figure 4 presents the solution and adaptive mesh at time $t = .75$. Figure 5 shows the L^2 error plotted as a function of the number of the grid points. In Table 2 we report the results of the errors and cpu time for the whole computation from $t = 0$ to $t = 1.25$.

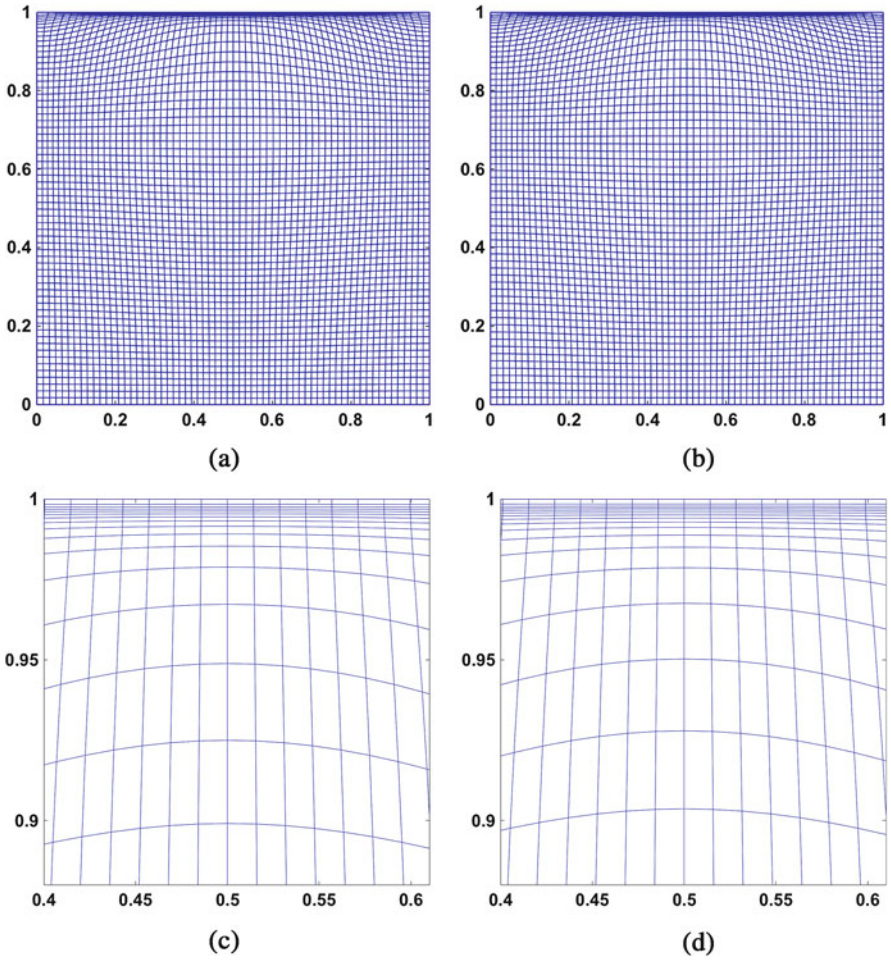

Fig. 2 Example 1. Adaptive mesh of (**a**) PMA-Sqrt method and (**b**) PMA-Log method. Plots (**c**) and (**d**) are cutaway of the mesh near the boundary layer region

Table 1 Example 1. Results of the errors and the cpu time for solving (12)

Method	$\|u_e - u_c\|_\infty$	$\|u_e - u_c\|_2$	Cpu time
PMA-Sqrt	0.0150	9.6597e-04	1 min and 41 s
PMA-Log	0.0148	8.6603e-04	2 min and 5 s

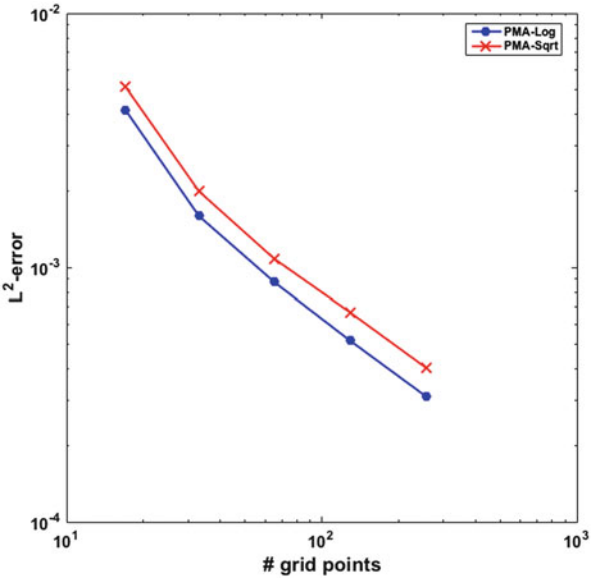

Fig. 3 Example 1. The L^2 error of the PMA methods for solving (12)

4.3 Example 3

We consider a phase-field model of a mixture of two incompressible fluid flows separated by a thin free moving interface of thickness $\hat{\eta}$. The interface is defined by a phase-field function $\phi(\boldsymbol{x}, t)$, on the interface $\phi(\boldsymbol{x}, t) = 0$, on one side of the interface $\phi(\boldsymbol{x}, t) = 1$, and on the other side $\phi(\boldsymbol{x}, t) = -1$. The time evolution of the interface is described by the Allen-Cahn phase field model [1, 12]

$$\phi_t + \boldsymbol{u} \cdot \nabla \phi = \gamma \left(\Delta \phi - \frac{1}{\hat{\eta}^2} \phi(\phi^2 - 1) + \hat{\xi}(t) \right), \tag{15a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \phi \mathrm{d}\boldsymbol{x} = 0, \tag{15b}$$

where $\boldsymbol{u}$ is the velocity of the fluids, γ is a time relaxation parameter, $\hat{\xi}(t)$ is the Lagrange multiplier introduced so that the volume fraction, defined as $\int_{\Omega} \phi d\boldsymbol{x}$, is conserved by the constraint (15b). We solve (15) for $\boldsymbol{u} = \boldsymbol{0}$, $\hat{\eta} = 0.0078$, $\gamma = 6.10351 \times 10^{-5}$ and initial phase function

$$\phi_0(x, y) = -\tanh\left((\sqrt{x^2 + y^2} - r_0)/\hat{\eta} \right), \quad (x, y) \in [-1, 1] \times [-1, 1] \tag{16}$$

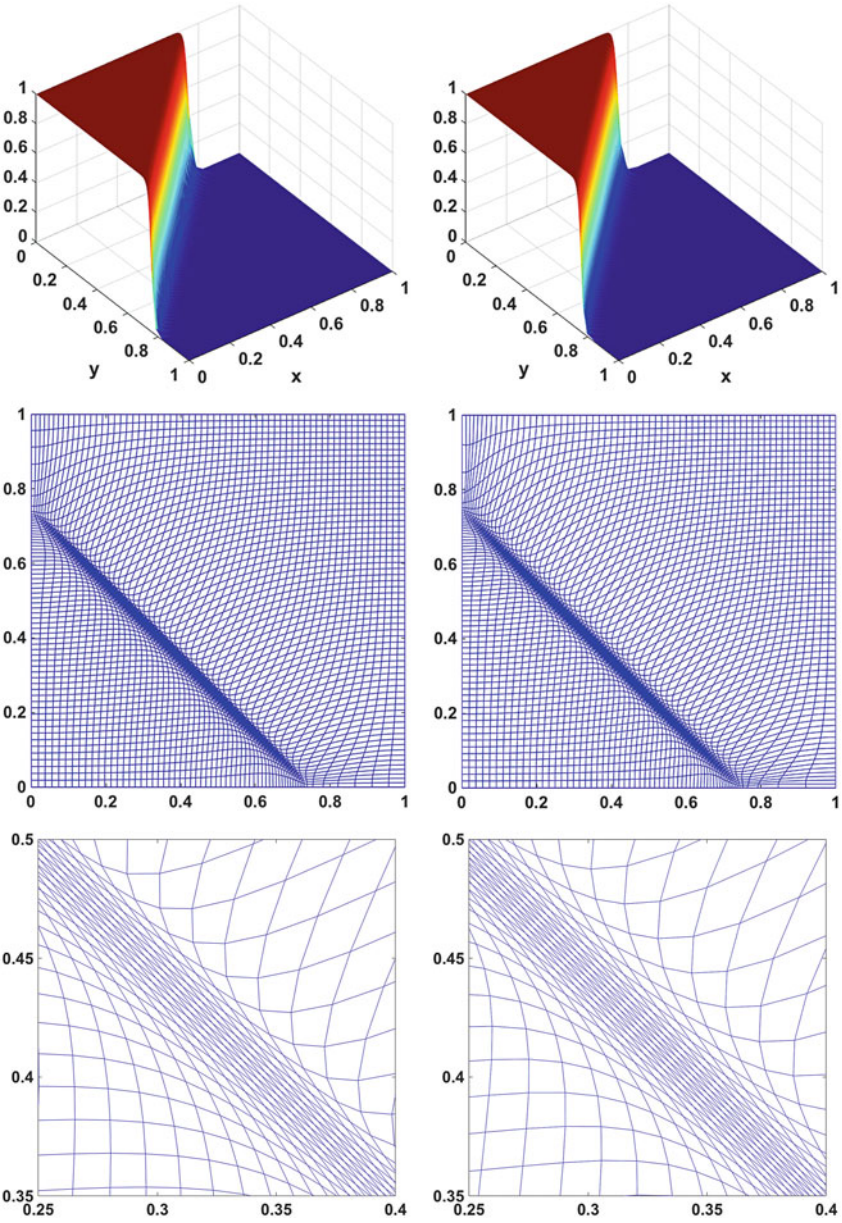

Fig. 4 Example 2. Solution surfaces, meshes, and cutaway of the meshes at $t = .75$ of Burgers' equation (14) computed with PMA-Sqrt (*left*) and PMA-Log (*right*)

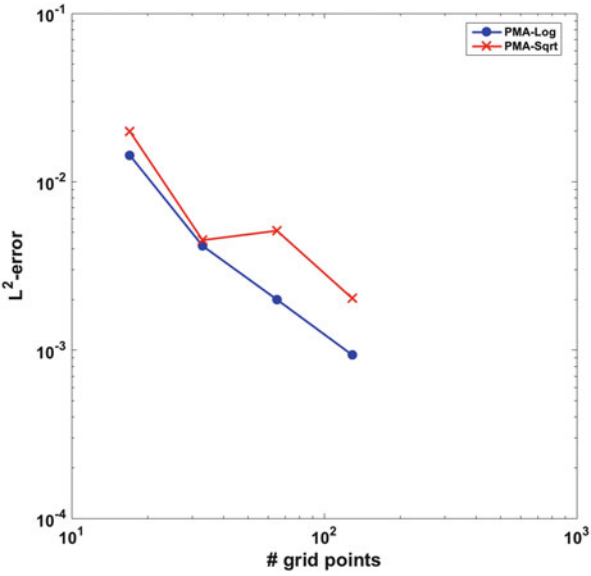

Fig. 5 Example 2. The L^2 error of the PMA methods for solving (14)

Table 2 Example 2. Results of the errors and cpu time for solving (14)

Method	$\|u_e - u_c\|_\infty$	$\|u_e - u_c\|_2$	Cpu time (in seconds)
PMA-Sqrt	0.0698	5.5×10^{-3}	40 s
PMA-Log	0.0296	1.9×10^{-3}	55 s

where $r_0 = 100/128$. In this case, the phase model describes an interface of a shrinking circular domain which moves towards its center under the mean curvature, and its radius is given by Liu and Shen [12]

$$R(t)^2 = R_0^2 - 2t, \tag{17}$$

and for $R_0 = 100$, the radius shrinks to zero at time $t = 5000$. Figure 6 shows the contour plots of the solution and Fig. 7 shows the adaptive mesh computed at times $t = 1000, 2000, 3000,$ and 5000. To assess the accuracy of the solutions, in Fig. 8 we plot $R(t)^2$ given by (17) and by the approximate solutions of the PMA methods. Notice that the results of PMA-Sqrt method are inaccurate specially for large times t while PMA-Log method maintains the same level of accuracy and has the mesh clustered in the interface region at all times up to $t = 5000$.

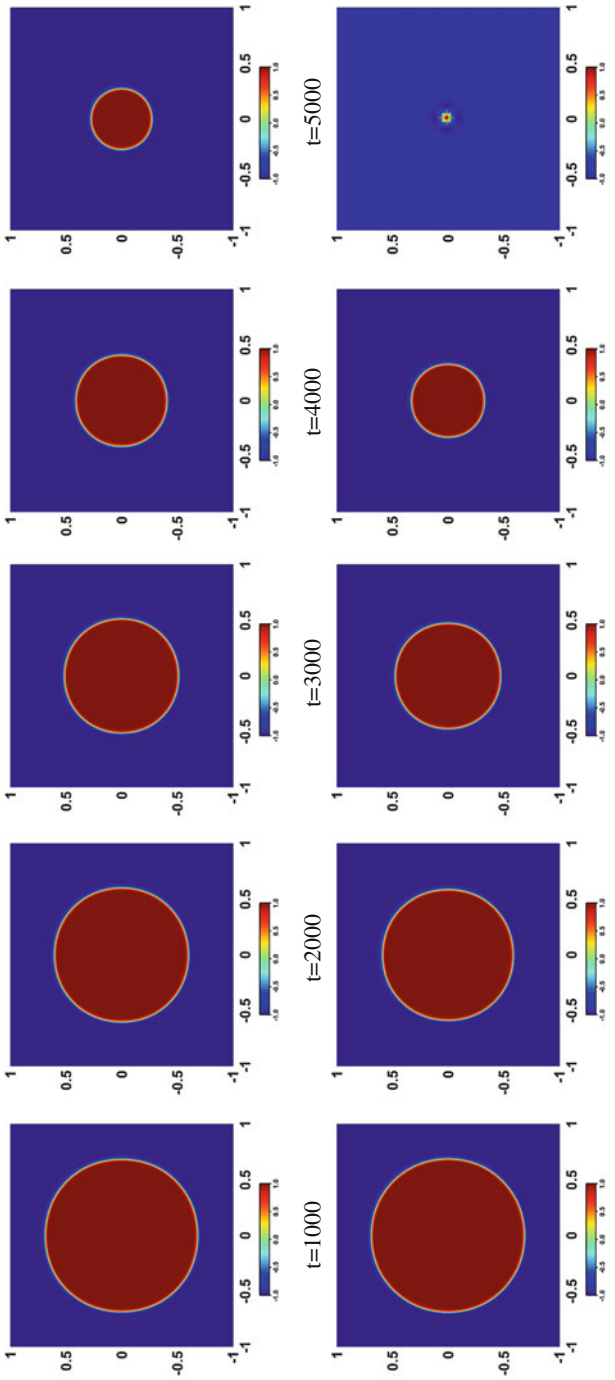

Fig. 6 Example 3. Solution contour plots: PMA-Sqrt (*top*) and PMA-Log (*bottom*)

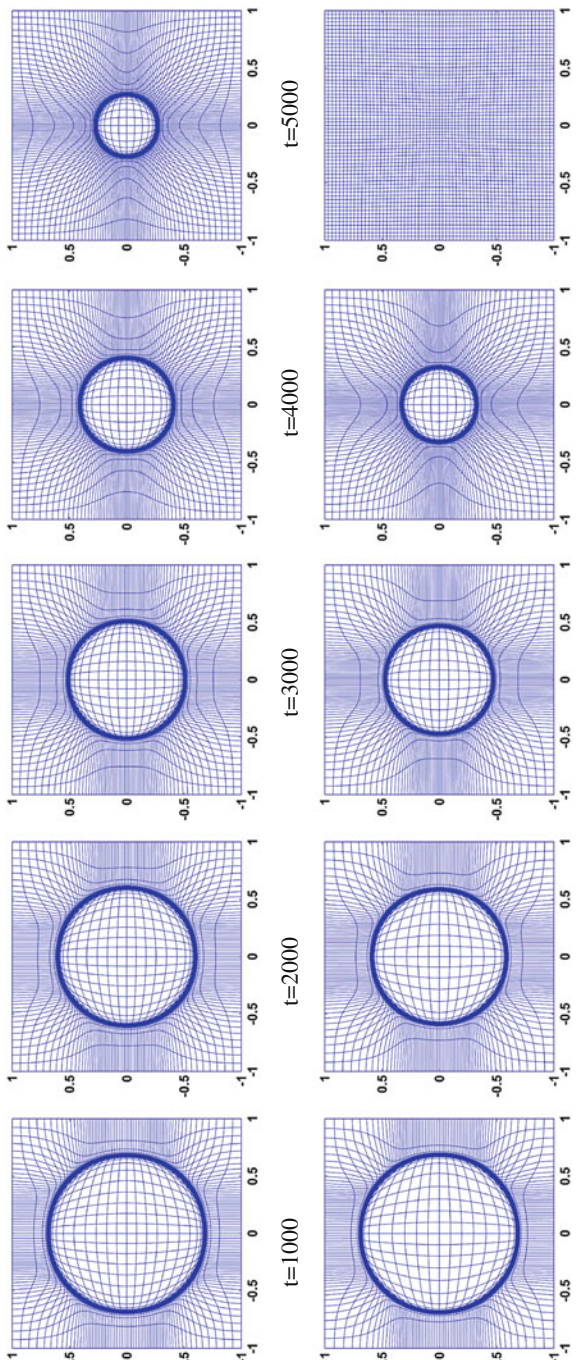

Fig. 7 Example 3. Adaptive meshes: PMA-Sqrt (*top*) and PMA-Log (*bottom*)

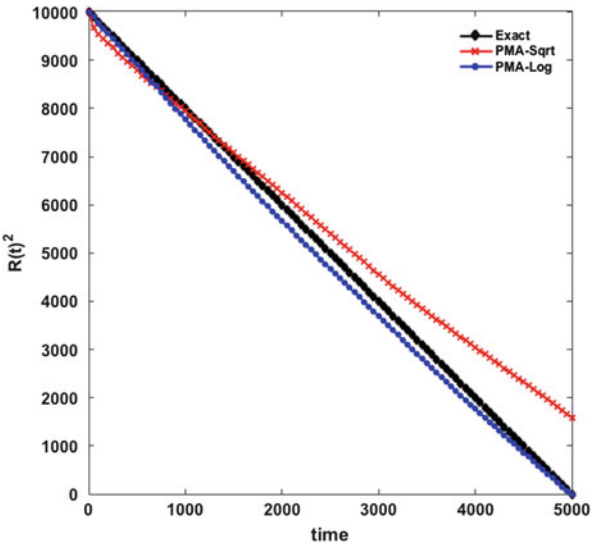

Fig. 8 Example 3. Plots of $R(t)^2$ computed by (17), PMA-Sqrt and PMA-Log

5 Conclusion

In this paper we have presented a comparison study of two parabolic Monge-Ampère equations adaptive grid methods, namely, PMA-Log method [14] and PMA-Sqrt [4]. The numerical results illustrate that the PMA-Log has faster convergence rate and is more accurate than the PMA-Sqrt. The PMA-Log method ensures exact equidistribution of the mesh at each time step which allows taking larger time step for integrating the physical problem while maintaining the same level of mesh quality throughout the physical time. This can offset the extra time associated with integrating (7) for the steady state solution at each time step. The results of Example 3 show that the mesh obtained by PMA-Sqrt method looses the equidistribution property when integrating the physical problem over a long period of time which leads to inaccurate solution of the physical problem.

We would like to point out that the parameter $\hat{\epsilon}$ (for temporal smoothing) of the PMA-Sqrt method has significant effect on how well the computed adaptive mesh are concentrated in the regions of large physical solution variations in the physical domain. Therefore, the choice of $\hat{\epsilon}$ can affect the accuracy of the PMA-Sqrt method. In this study, we have tested several values of $\hat{\epsilon}$ and then make use of the value that works best for each problem we have studied.

References

1. Allen, S.M., Cahn, J.W.: A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. Acta Metall. **27**(6), 1085–1095 (1979)
2. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
3. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**(4), 375–417 (1991)
4. Budd, C.J., Williams, J.F.: Moving mesh generation using the parabolic Monge-Ampère equation. SIAM J. Sci. Comput. **31**(5), 3438–3465 (2009)
5. de Boor, C.: Good approximation by splines with variable knots. II. In: Conference on the Numerical Solution of Differential Equations (University of Dundee, Dundee, 1973). Lecture Notes in Mathematics, vol. 363, pp. 12–20. Springer, Berlin (1974)
6. Huang, W.: Practical aspects of formulation and solution of moving mesh partial differential equations. J. Comput. Phys. **171**(2), 753–775 (2001)
7. Huang, W.: Variational mesh adaptation: isotropy and equidistribution. J. Comput. Phys. **174**(2), 903–924 (2001)
8. Huang, W., Russell, R.D.: Moving mesh strategy based on a gradient flow equation for two-dimensional problems. SIAM J. Sci. Comput. **20**(3), 998–1015 (1999) (electronic)
9. Huang, W., Russel, R.D.: Adaptive Moving Mesh Methods. Springer, New York (2011)
10. Huang, W., Ren, Y., Russell, R.D.: Moving mesh partial differential equations (MMPDES) based on the equidistribution principle. SIAM J. Numer. Anal. **31**(3), 709–730 (1994)
11. Kantorovich, L.V.: On a problem of Monge. Uspehki Mat. Nauk **3**, 225–226 (1948)
12. Liu, C., Shen, J.: A phase field model for the mixture of two incompressible fluids and its approximation by a fourier-spectral method. Physica D Nonlinear Phenom. **179**(3–4), 211–228 (2003)
13. Monge, G.: Mémoire sur la théorie des déblais et des remblais. In: Histoire de l'Académie Royale des Sciences de Paris, pp. 666–704. De l'Imprimerie Royale, Paris (1781)
14. Sulman, M., Williams, J.F., Russell, R.D.: Optimal mass transport for higher dimensional adaptive grid generation. J. Comput. Phys. **230**(9), 3302–3330 (2011)
15. Sulman, M., Williams, J., Russell, R.D.: An efficient approach for the numerical solution of the Monge-Ampère equation. Appl. Numer. Math. **61**(3), 298–307 (2011)

Approximate Solutions to Poisson Equation Using Least Squares Support Vector Machines

Ziku Wu, Zhenbin Liu, Fule Li, and Jiaju Yu

Abstract This article deals with Poisson Equations with Dirichlet boundary conditions. A new approach based on least squares support vector machines (LS-SVM) is proposed for obtaining their approximate solutions. The approximate solution is presented in closed form by means of LS-SVM, whose parameters are adjusted to minimize an appropriate error function. The approximate solutions consist of two parts. The first part is a known function that satisfies boundary conditions. The other is two terms product. One term is known function which is zero on boundary, another term is unknown which is related to kernel functions. This method has been successfully tested on rectangle and disc domain and has yielded higher accuracy solutions.

1 Introduction

There are a variety of physical phenomena and engineering problems described by Poisson Equation [1]. For instance, it is used to describe the potential energy field caused by a given charge or mass density distribution. However, we can't obtain its analytic solutions in most conditions. It needs to resort numerical methods. The most commonly used classical numerical methods for the solution of Poisson Equations are the finite difference method, the finite element method and the finite volume method [2]. Over the past decades, several kinds of fast methods for solving Poisson Equations have been proposed. Multigrid methods and the fast multipole methods are the most valid fast methods [3, 4]. While these methods are generally useful in

Z. Wu • Z. Liu • F. Li
Science and Information College, Qingdao Agricultural University, Qingdao 266109, China
e-mail: zkwu1968@126.com; lzbzj@163.com

J. Yu (✉)
Science and Information College, Qingdao Agricultural University, Qingdao 266109, China

School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
e-mail: yujiajuccm@163.com; 157770733@qq.com

Z. Huang et al. (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, Lecture Notes in Computational Science and Engineering 120, https://doi.org/10.1007/978-3-319-67202-1_15

many problems, one obvious limitation is that the obtained solutions are discrete or have limited differentiability.

In order to avoid this defect, some researchers employ artificial neural networks (ANN) to solve Poisson Equations. Lagaris et al. [5, 6] employed ANN for solving ODE and PDE. Baymani et al. [7] used ANN method to solve Stokes equation. Alli et al. [8] solved the vibration control problems using artificial neural networks. Support vector machines (SVM) was presented by Vapnik et al. [9], which has been successfully applied in many aspects for its high generalization ability and global optimization property.The simplicity of LS-SVM promotes the applications of SVM [10]. Over the last decade, many pattern recognition and function approximation problems have successfully been tackled with LS-SVM method. Recently Mehrkanoon et al. [11–14] proposed a new approach based on LS-SVM to solve ODEs and PDEs.

In this work, we employ a method based on LS-SVM to solve Poisson Equations. This paper is organized as follows. In Sect. 2 the LS-SVM regression is given briefly. In Sect. 3, we formulate the LS-SVM method for Poisson Equations. Section 4 describes the numerical experiments, including two examples, which is followed by the conclusion in Sect. 5.

2 Brief Introduction to the Modified LS-SVM Regression

Let us consider a given training set $\{X_i, Y_i\}_{i=1}^N$ with input data $X_i \in R^k$ and output data $Y_i \in R$. Our goal is to estimate a model of the following form using the regression:

$$Y = g(X) = \sum_{j=1}^{N} \alpha_j G(V, V_j) + b \tag{1}$$

where $G(X, X_j)$=exp$(-\frac{1}{2\sigma^2}\|X - X_j\|_2^2)$ is the Gaussian kernel function with kernel width σ, α_j and b are the regression parameters which need to be estimated. The values of the parameters can be obtained by solving the following quadratic programming problem:

$$\min_{\alpha.b,e} \frac{1}{2}\alpha^T\alpha + \frac{\gamma}{2}e^T e \tag{2}$$

$$s.t.\ Y = \sum_{j=1}^{N} \alpha_j G(V_i, V_j) + b + e_i, i = 1, \cdots, N \tag{3}$$

where $\gamma \in R^+$ is regularization parameter and $e_i \in R$ is the bias term. We denote $\hat{Y} = [Y_1, Y_2, \cdots, Y_N]^T \in R^N$ $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_N]^T \in R^N$, $1_N = [1, 1, \cdots, 1]^T \in R^N$ and $0_N = [0, 0, \cdots, 0]^T \in R^N$. The solutions is given by

$$\begin{bmatrix} I_N & -\Omega_N^T & 0_N \\ \Omega_N & \gamma^{-1} I_N & 1_N \\ 0_N & 1_N & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \lambda \\ b \end{bmatrix} = \begin{bmatrix} 0_N \\ \hat{Y} \\ 0 \end{bmatrix} \tag{4}$$

where $\Omega = G(X_i, X_j)_N$ is the kernel matrix and I_N is the identity matrix of order N, and $\lambda = (\lambda_1, \lambda_2, \cdots \lambda_N)^T$ is the Lagrangian multiplier.

3 Formulation of the Method for Poisson Equation

Consider the following Poisson equation with Dirichlet boundary conditions:

$$\begin{cases} \Delta^2 u = f(x, y) \in \Omega \\ u|_{\partial\Omega} = g(x, y) \end{cases} \tag{5}$$

The problem is defined on the domain $\Omega \subset R^2$. In order to get approximate function, we consider a mesh of M interior points of the domain. We reshape these points as a vector $V = [V_1, V_2, \cdots, V_M]^T$, $V_i = (x_i, y_i)$ is i-th points. Assuming the solution of equation (5) has the following expression:

$$u(V) = A(V) + B(V)(\sum_{j=1}^{M} \alpha_j G(V, V_j) + b) \tag{6}$$

where $V = (x, y)$,$V_j = (x_j, y_j)$. in this expression, $A(V)$ is a known function that satisfies the boundary conditions. In the other hand, the function $B(V)$ takes zero value on the boundary. α and b are the regression parameters which have to be determined. Inserting (6) into (5), we obtain the following equation:

$$\sum_{j=1}^{M} \alpha_j \Phi(V, V_j) + D(V) + Q(V)b - f(V) = 0 \tag{7}$$

where $D(V) = A_{xx}(V) + A_{yy}(V)$, $Q(V) = B_{xx}(V) + B_{yy}(V)$,$\Phi(V, V_j) = G_1(V, V_j) + G_2(V, V_j)$, $G_1(V, V_j) = B_{xx}(V)G(V, V_j) + 2B_x(V)G_x(V, V_j) + B(V)G_{xx}(V, V_j)$, $G_2(V, V_j) = B_{yy}(V)G(V, V_j) + 2B_y(V)G_y(V, V_j) + B(V)G_{yy}(V, V_j)$, To obtain the

optimal values of and , we can solve the following optimization problem,

$$\min_{\alpha.b,e} \frac{1}{2}\alpha^T\alpha + \frac{\gamma}{2}e^T e \tag{8}$$

$$s.t.\ Y = \sum_{j=1}^{M} \alpha_j \Phi(V_i, V_j) + Q(V_i)b + D(V_i) - f(V_i) + e_i, i = 1, \cdots, M \tag{9}$$

where $\gamma \in R^+$is a regularization constant and $e_i, i = 1, 2, \cdots, M$ are bias terms. The Lagrangian function of the constrained optimization problem (8) and (9) becomes

$$\begin{aligned} L(\alpha, e, b, \eta) =& \frac{1}{2}(\alpha^T\alpha + \bar{\alpha}^T\bar{\alpha}) + \frac{\gamma}{2}(e^T e + \bar{e}^T\bar{e}) \\ & - \sum_{i=1}^{M} \eta_i (\sum_{j=1}^{M} \alpha_j \Phi_j(V_i, V_j) + Q(V_i)b + D(V_i) - f(V_i) + e_i) \end{aligned} \tag{10}$$

Then, the Karush-Kuhn-Tucker (KKT) optimality conditions as follows:

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_i} = \alpha_i - \displaystyle\sum_{j=1}^{M} \eta_j \Phi(V_j, V_i) = 0 \\ \dfrac{\partial L}{\partial e_i} = \gamma_i e_i - \eta_i = 0 \\ \dfrac{\partial L}{\partial \eta_i} = -(\displaystyle\sum_{j=1}^{M} \alpha_j \Phi(V_i, V_j) + Q(V_i)b + D(V_i) - f(V_i) + e_i) = 0 \\ \dfrac{\partial L}{\partial b} = -\displaystyle\sum_{j=1}^{M} \eta_j Q(V_j) = 0 \end{cases} \tag{11}$$

After elimination of the primal variable e_i , the solution is given by

$$\begin{bmatrix} I_N & -KM^T & Z \\ KM & \frac{I_M}{\gamma} & LQ \\ Z^T & LQ^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \eta \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ f\alpha \\ 0 \end{bmatrix} \tag{12}$$

where $f\alpha = [f(V_1) - D(V_1), f(V_2) - D(V_2), \cdots, f(V_M) - D(V_M)]$, $Z = [0, 0, \cdots, 0]$ is an M dimension vector, I_N is a unit matrix of order M, $LQ = [Q(V_1), Q(V_2), \cdots, Q(V_M)]$ and $KM = [\Phi(V_i, V_j)]$ is the kernel matrix of order. Expression (12) is a linear equation, which can be solved easily. Finally, we can obtain the approximate solutions of problem (5).

Remark Of course you can choose different kernel functions, such as Multiquadrics kernel function. The approximate solution we used is similar to ANN's, but the method for solving the parameters is different. There are two distinct differences between ours and references [11–14]. The first is the form of the approximate solutions, and the other is our method provide approximate solution directly without need dual form.

4 Numerical Experiments

In this section, we test the performance of the method on two problems, one with rectangle domain and the other with disc domain. In order to show the approximation capability of the method, we compare the computed approximate solution with the analytic solution and numerical solution.

Problem 1 Consider the following equation

$$\begin{cases} \Delta u = -2, (x, y) \in (0, 1)^2 \\ u|_{\partial\Omega} = 0 \end{cases} \tag{13}$$

And its analytic solution is

$$u(x, y) = x(1 - x) - \frac{8}{\pi^3} \sum_{n=1}^{\infty} [\frac{sinh[(2n-1)(1-y)\pi] + sinh[(2n-1)y\pi]}{sinh(2n-1)\pi}] \frac{sin(2n-1)\pi x}{(2n-1)^3}$$

For this example, we take $A(x, y) = 0$, $B(x, y) = (x-x^2)(y-y^2)$. We partitioned the domain by equal step. For instance, $h = 0.05$, then $M = 361$. The approximate solutions obtained by the method are compared with the numerical solution and results are depicted in Fig. 1. The obtained absolute errors for interior points are tabulated in Table 1. It is clear that the approximate solutions are quite acceptable, despite the fact that fewer training points are employed.

Problem 2 Consider the following equation

$$\begin{cases} \Delta u = x^2 - y^2, x^2 + y^2 < 1 \\ u|_{\partial\Omega} = 0 \end{cases} \tag{14}$$

Its analytic solution is $u(x, y) = \frac{1}{12}(x^2 + y^2)(x^2 + y^2 - 1)$. For this problem, we take $A(x, y) = 0$ and $B(x, y) = x^2 + y^2 - 1$. In this example, we divide the interval $[-1, 1]$ on the X axis and the Y axis, respectively. The interior points are employed as training points. The corresponding training point number is 305. The results of the approximate solutions compared with the analytical solutions are listed in Fig. 2 and Table 1.

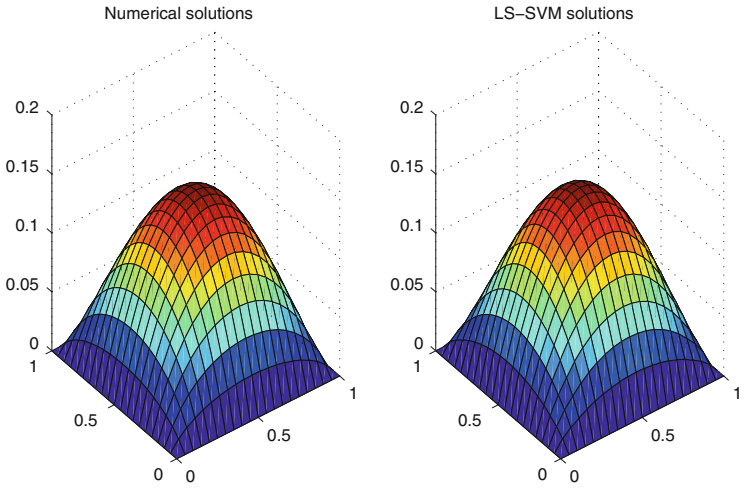

Fig. 1 Numerical results of Problem 1

Table 1 Numerical results errors

	Training points	$\|U^H - U^L\|_\infty$	Mean error
Problem 1	361	$4.0e-4 = o(h^2)$	$1.8e-4$
Problem 2	305	$3.7e-4 = o(h^2)$	$1.6e-4$

In this table, U^H stands for numerical solutions or analytic solutions, U^L denotes LS-SVM solutions and h is space step

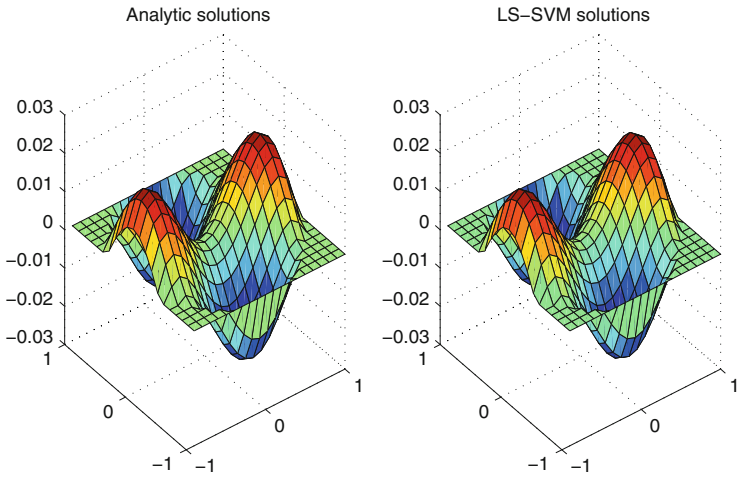

Fig. 2 Numerical results of Problem 2

5 Conclusions

The method based on LS-SVM can solve successfully ODEs and PDEs. Although the method based on ANN can solve ordinary differential equations and partial differential equations with higher accuracy, it has some obvious drawbacks. Theoretically, PDEs can be solved by ANN that must be solved by LS-SVM. Because of complex boundary conditions and nonlinearity, the method based on LS-SVM may have some trouble to solve PDEs. That is why we focus on Poisson Equations in this paper. Taking into account complexity, we assume approximate solutions directly which does not need dual form. This is different from previous ones. On the tested problems, the method proposed in this paper is successful with higher accuracy. Consequently, this method can be used for Poisson Equations with complex boundary conditions. We believe this method can be extended to solve other partial differential equations.

Acknowledgements This work is supported by the international cooperation for excellent lectures of 2013, Shandong provincial education department, and the NNSF of China (61403233).

References

1. Strauss, W.A.: Partial Differential Equations. Joans Wiley & Sons, Ltd, Hoboken (2007)
2. Tadmor, E.: A review of numerical methods for nonlinear partial differential equations. Bull. Am. Math. Soc. **49**, 507–554 (2012)
3. Trottenberg, U., Oosterlee, C.W., Schller, A.: Multigrid. Academic Press, Cambridge (2001)
4. Ma, Z., Chew, W.C., Jiang, L.: A novel faster solver for Poisson's equation with Neumann boundary condition. Prog. Electromagn. Res. **136**, 195–209 (2013)
5. Lagaris, I.,Likas, A., Fotiadis, D.: Artificial neural networks for solving ordinary and partial differential equations. IEEE Trans. Neural Netw. **9**, 987–1000 (1998)
6. Lagaris, I., Likas, A., Papageorgio, D.: Neural-network methods for boundary value problems with irregular boundaries. IEEE Trans. Neural Netw. **11**, 1041–1049 (2000)
7. Baymani, M., Kerayechian, A., Effati, S.: Artificial neural networks approach for solving Stokes problem. Appl. Math. **1**, 288–292 (2010)
8. Alli, H., Ucar, A., Demir, Y.: The solutions of vibration control problems using artificial neural networks. J. Frankl. Inst. **340**, 307–325 (2003)
9. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
10. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., et al.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
11. Mehrkanoon, S., Suykens, J.: LS-SVM approximate solution to linear time varying descriptor systems. Automatica **48**, 2502–2511 (2012)
12. Mehrkanoon, S., Falck, T., Suykens, J.: Approximate solution to ordinary differential equations using least squares support vector machines. IEEE Trans. Neural Netw. Learn. Syst. **23**, 1356–1367 (2012)
13. Mehrkanoon, S., Suykens, J.: Parameter estimation of delay differential equations: an integration-free LS-SVM approach. Commun. Nonlinear Sci. Numer. Simul. **19**, 830–841 (2014)
14. Mehrkanoon, S., Suykens, J.A.K.: Learning solutions to partial differential equations using LS-SVM. Neurocomputing **159**, 105–116 (2015)

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Tutorials
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). Tutorials are lecture notes typically arising via summer schools or similar events, which are used to teach graduate students. These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgement on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

– at least 100 pages of text;
– a table of contents;
– an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
– a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact the Editor for CSE at Springer at the planning stage, see *Addresses* below.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Only works in English will be considered. For evaluation purposes, manuscripts may be submitted in print or electronic form, in the latter case, preferably as pdf- or zipped ps-files. Authors are requested to use the LaTeX style files available from Springer at http://www.springer.com/gp/authors-editors/book-authors-editors/manuscript-preparation/5636 (Click on LaTeX Template → monographs or contributed books).

For categories ii) and iii) we strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. Electronic material can be included if appropriate. Please contact the publisher.

Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online.

5. The following terms and conditions hold. Categories i), ii) and iii):

Authors receive 50 free copies of their book. No royalty is paid.
Volume editors receive a total of 50 free copies of their volume to be shared with authors, but no royalties.

Authors and volume editors are entitled to a discount of 33.3 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Springer secures the copyright for each volume.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
griebel@ins.uni-bonn.de

David E. Keyes
Mathematical and Computer Sciences
and Engineering
King Abdullah University of Science
and Technology
P.O. Box 55455
Jeddah 21534, Saudi Arabia
david.keyes@kaust.edu.sa

and

Department of Applied Physics
and Applied Mathematics
Columbia University
500 W. 120 th Street
New York, NY 10027, USA
kd2112@columbia.edu

Risto M. Nieminen
Department of Applied Physics
Aalto University School of Science
and Technology
00076 Aalto, Finland
risto.nieminen@aalto.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
dirk.roose@cs.kuleuven.be

Tamar Schlick
Department of Chemistry
and Courant Institute
of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA
schlick@nyu.edu

Editor for Computational Science
and Engineering at Springer:
Martin Peters
Springer-Verlag
Mathematics Editorial IV
Tiergartenstrasse 17
69121 Heidelberg, Germany
martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*
2. H.P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.
3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*
4. P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*
5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*
6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.
7. R. von Schwerin, ***M**ulti **B**ody **S**ystem **SIM**ulation.* Numerical Methods, Algorithms, and Software.
8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
9. T.J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*
10. H.P. Langtangen, A.M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing*.
11. B. Cockburn, G.E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.
12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.
13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*
14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*
15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*
16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.
17. B.I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*
18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*
19. I. Babuška, P.G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*
20. T.J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.
21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.
23. L.F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H.H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T.J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M.A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H.P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B.N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III*.

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV*.

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science.*

67. I.H. Tuncer, Ü. Gülcat, D.R. Emerson, K. Matsuno (eds.), *Parallel Computational Fluid Dynamics 2007.*

68. S. Yip, T. Diaz de la Rubia (eds.), *Scientific Modeling and Simulations*.

69. A. Hegarty, N. Kopteva, E. O'Riordan, M. Stynes (eds.), *BAIL* 2008 – *Boundary and Interior Layers.*

70. M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XVIII.*

71. B. Koren, C. Vuik (eds.), *Advanced Computational Methods in Science and Engineering.*

72. M. Peters (ed.), *Computational Fluid Dynamics for Sport Simulation.*

73. H.-J. Bungartz, M. Mehl, M. Schäfer (eds.), *Fluid Structure Interaction II - Modelling, Simulation, Optimization.*

74. D. Tromeur-Dervout, G. Brenner, D.R. Emerson, J. Erhel (eds.), *Parallel Computational Fluid Dynamics 2008.*

75. A.N. Gorban, D. Roose (eds.), *Coping with Complexity: Model Reduction and Data Analysis.*

76. J.S. Hesthaven, E.M. Rønquist (eds.), *Spectral and High Order Methods for Partial Differential Equations.*

77. M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*.

78. Y. Huang, R. Kornhuber, O.Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XIX*.

79. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations V*.

80. P.H. Lauritzen, C. Jablonowski, M.A. Taylor, R.D. Nair (eds.), *Numerical Techniques for Global Atmospheric Models*.

81. C. Clavero, J.L. Gracia, F.J. Lisbona (eds.), *BAIL 2010 – Boundary and Interior Layers, Computational and Asymptotic Methods*.

82. B. Engquist, O. Runborg, Y.R. Tsai (eds.), *Numerical Analysis and Multiscale Computations*.

83. I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (eds.), *Numerical Analysis of Multiscale Problems*.

84. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*.

85. J. Blowey, M. Jensen (eds.), *Frontiers in Numerical Analysis - Durham 2010*.

86. O. Kolditz, U.-J. Gorke, H. Shao, W. Wang (eds.), *Thermo-Hydro-Mechanical-Chemical Processes in Fractured Porous Media - Benchmarks and Examples*.

87. S. Forth, P. Hovland, E. Phipps, J. Utke, A. Walther (eds.), *Recent Advances in Algorithmic Differentiation*.

88. J. Garcke, M. Griebel (eds.), *Sparse Grids and Applications*.

89. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VI*.

90. C. Pechstein, *Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems*.

91. R. Bank, M. Holst, O. Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XX*.

92. H. Bijl, D. Lucor, S. Mishra, C. Schwab (eds.), *Uncertainty Quantification in Computational Fluid Dynamics*.

93. M. Bader, H.-J. Bungartz, T. Weinzierl (eds.), *Advanced Computing*.

94. M. Ehrhardt, T. Koprucki (eds.), *Advanced Mathematical Models and Numerical Techniques for Multi-Band Effective Mass Approximations*.

95. M. Azaïez, H. El Fekih, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2012*.

96. F. Graziani, M.P. Desjarlais, R. Redmer, S.B. Trickey (eds.), *Frontiers and Challenges in Warm Dense Matter*.

97. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Munich 2012*.

98. J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXI*.

99. R. Abgrall, H. Beaugendre, P.M. Congedo, C. Dobrzynski, V. Perrier, M. Ricchiuto (eds.), *High Order Nonlinear Numerical Methods for Evolutionary PDEs - HONOM 2013*.

100. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VII*.

101. R. Hoppe (ed.), *Optimization with PDE Constraints - OPTPDE 2014*.

102. S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab, H. Yserentant (eds.), *Extraction of Quantifiable Information from Complex Systems*.

103. A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso (eds.), *Numerical Mathematics and Advanced Applications - ENUMATH 2013*.

104. T. Dickopf, M.J. Gander, L. Halpern, R. Krause, L.F. Pavarino (eds.), *Domain Decomposition Methods in Science and Engineering XXII*.

105. M. Mehl, M. Bischoff, M. Schäfer (eds.), *Recent Trends in Computational Engineering - CE2014*. Optimization, Uncertainty, Parallel Algorithms, Coupled and Complex Problems.

106. R.M. Kirby, M. Berzins, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM'14*.

107. B. Jüttler, B. Simeon (eds.), *Isogeometric Analysis and Applications 2014*.

108. P. Knobloch (ed.), *Boundary and Interior Layers, Computational and Asymptotic Methods – BAIL 2014*.

109. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Stuttgart 2014*.

110. H. P. Langtangen, *Finite Difference Computing with Exponential Decay Models*.

111. A. Tveito, G.T. Lines, *Computing Characterizations of Drugs for Ion Channels and Receptors Using Markov Models*.

112. B. Karazösen, M. Manguoğlu, M. Tezer-Sezgin, S. Göktepe, Ö. Uğur (eds.), *Numerical Mathematics and Advanced Applications - ENUMATH 2015*.

113. H.-J. Bungartz, P. Neumann, W.E. Nagel (eds.), *Software for Exascale Computing - SPPEXA 2013-2015*.

114. G.R. Barrenechea, F. Brezzi, A. Cangiani, E.H. Georgoulis (eds.), *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*.

115. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VIII*.

116. C.-O. Lee, X.-C. Cai, D.E. Keyes, H.H. Kim, A. Klawonn, E.-J. Park, O.B. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXIII*.

117. T. Sakurai, S. Zhang, T. Imamura, Y. Yusaku, K. Yoshinobu, H. Takeo (eds.), *Eigenvalue Problems: Algorithms, Software and Applications, in Petascale Computing*. EPASA 2015, Tsukuba, Japan, September 2015.

118. T. Richter (ed.), *Fluid-structure Interactions*. Models, Analysis and Finite Elements.

119. M.L. Bittencourt, N.A. Dumont, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2016*.

120. Z. Huang, M. Stynes, Z. Zhang (eds.), *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/3527

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: www.springer.com/series/7417

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition
2. A. Quarteroni, F. Saleri, P. Gervasio, *Scientific Computing with MATLAB and Octave.* 4th Edition
3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition
4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.
5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.
6. H. P. Langtangen, *A Primer on Scientific Programming with Python.* 5th Edition
7. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai, *Elements of Scientific Computing.*
8. B. Gustafsson, *Fundamentals of Scientific Computing.*
9. M. Bader, *Space-Filling Curves.*
10. M. Larson, F. Bengzon, *The Finite Element Method: Theory, Implementation and Applications.*
11. W. Gander, M. Gander, F. Kwok, *Scientific Computing: An Introduction using Maple and MATLAB.*
12. P. Deuflhard, S. Röblitz, *A Guide to Numerical Modelling in Systems Biology*.
13. M. H. Holmes, *Introduction to Scientific Computing and Data Analysis*.
14. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with MATLAB/Octave.
15. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with Python.
16. H.P. Langtangen, S. Linge, *Finite Difference Computing with PDEs* - A Modern Software Approach.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/5151