Springer Proceedings in Mathematics & Statistics

Ali R. Ansari *Editor*

Advances in Applied Mathematics

Springer

Springer Proceedings in Mathematics & Statistics

Volume 87

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Ali R. Ansari
Editor

Advances in Applied Mathematics

Springer

Editor
Ali R. Ansari
Gulf University for Science
and Technology
Mishref Campus, Kuwait

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-06922-7 ISBN 978-3-319-06923-4 (eBook)
DOI 10.1007/978-3-319-06923-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014945204

Mathematics Subject Classification (2010): 92Bxx, 74S05, 91Gxx, 35Q35, 65L11

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book contains the best papers presented at the Gulf International Conference on Applied Mathematics 2013 (GICAM'13). The conference is the first, in a series we hope, of conferences on Applied Mathematics held at the Gulf University for Science and Technology in Kuwait in cooperation with the Society for Industrial and Applied Mathematics (SIAM). Our intention at this conference, as any gathering of academics, was to bring together people in the region with world-renowned experts to help set in motion a significant drive to focus on certain research areas. In line with this, three major themes were chosen,mathematical biology, computational science and applications of mathematics in industry/business. The latter two themes form the bulk of research pursued in applied mathematics in the Gulf region; however, mathematical biology is a relatively new area of research.

One of the world's leading centres on the subject is housed in the Mathematical Institute at Oxford University in the form of the Wolfson Centre for Mathematical Biology (WCMB) (http://www.maths.ox.ac.uk/groups/mathematical-biology). Thus, our keynote address for this theme was given by professor Philip Maini, the director of the WCMB. The keynote address for the second theme on computational science was given by professor Grigorii Shishkin, well known in the field for his non-uniform mesh that has helped in solving many singularly perturbed differential equations and problems with thin boundary layers. The third theme was on applications of mathematics in general used for solving industrial problems; the keynote address for this theme was given by professor Ali Nayfeh, well known for his book on perturbation methods and his significant contributions to engineering.

We received over 100 abstracts and accepted 77 for presentation; these conference proceedings represent 25 of the best papers presented at the conference.

Of course any conference of this magnitude comprising of participants from over 25 countries cannot be successful without financial support. In this regard we would like to thank and acknowledge the support of the Kuwait Foundation for the Advancement of Science (KFAS); the Gulf University for Science and Technology (GUST); Business Development and Corporate Relations, GUST; Growmore; Springer; McGrawhill Education; Naseej, Arabian Advanced Systems; Institute of Numerical Computation and Analysis (INCA), Ireland.

Finally, we acknowledge the support of the Department of Mathematics and Natural Sciences in organising the conference. All of the faculty were most supportive, particularly Dr. Helmi Temimi and Dr. Wasim Daher for their effort in making sure the conference was successful. In addition, we are most grateful to all our participants and our invited speakers without whom this conference would not have been possible.

Mishref Campus, Kuwait Ali R. Ansari
March 2014

Acknowledgements

The following colleagues are acknowledged for their help and advice in reviewing the papers published in this book:

Harun Aydilek	Gulf University for Science and Technology, Kuwait
Amin Boumenir	Kuwait University, Kuwait
Bruno Buonomo	University of Naples Federico II via Cintia, Italy
Abdul Hanan Choudhury	Srikishan Sarda College, India
Wasim Daher	Gulf University for Science and Technology, Kuwait
Sadeq Damrah	Australian College of Kuwait, Kuwait
Ahmet Duran	Istanbul Technical University, Turkey
Abdurazzak Ghezal	Universite Kasdi Merbah, Algeria
Mudassar Imran	Lahore University for Management Sciences, Pakistan
Razidah Ismail	Universiti Teknologi MARA, Malaysia
Rihab Khalaf	Public Authority for Education and Training, Kuwait
Tadeusz Kosztolowicz	Jan Kochanowski University, Poland
Munir Mehmood	Gulf University for Science and Technology, Kuwait
Mohammed Ben Romadhane	Gulf University for Science and Technology, Kuwait
Loris Serafino	Australian College of Kuwait, Kuwait
Helmi Temimi	Gulf University for Science and Technology, Kuwait
Harun Krukcu	Gulf University for Science and Technology, Kuwait

Contents

Modelling Collective Cell Motion in Biology

P.K. Maini and R.E. Baker

Abstract This paper reviews three mathematical modelling approaches that have recently been used to understand three different modes of collective cell motion in biology. Firstly, a cell-based model is presented for the study of cell motion in epithelial sheets, then a hybrid discrete cell-based model is described for neural crest cell invasion and, finally, a traditional partial differential equation model is described for tumour cell invasion. It is shown that the behaviour of all of these models can, in limiting cases, be recapitulated by nonlinear diffusion equations where the particular nonlinearity of the diffusion coefficient captures, on the global scale, the inherent interactions on the local scale.

Keywords Cell-based model • Vertex-based model • Travelling waves • Volume exclusion

1 Introduction

The collective movement of individuals is a common feature in nature, ranging from the migration of herds of wildebeest to the dramatic aerial displays of bird flocks. The body also plays host to collective motion, but of cells. For example, in early development, cells move within sheets by exchanging neighbours or crawl long distances by repeated expansion and retraction of the front and back of the cell, respectively. In cases such as wound healing, cell motion is essential for correct

This is a brief summary of a presentation given by the first author at the Gulf International Conference on Applied Mathematics, Kuwait 2013.

P.K. Maini (✉) • R.E. Baker
Wolfson Centre for Mathematical Biology, University of Oxford, Oxford, UK
e-mail: maini@maths.ox.ac.uk; baker@maths.ox.ac.uk
webpage: http://people.maths.ox.ac.uk/maini/; http://people.maths.ox.ac.uk/baker/

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_1,

restoration of the tissue, while in cancer it can lead to fatal metastases. In this paper we briefly review three recent studies on collective cell motion. In Sect. 2, we present a model for the movement of cells within an epithelial sheet of the mouse embryo; in Sect. 3 we consider the movement of cranial neural crest cells and in Sect. 4 we consider a model for acid-mediated cell invasion in cancer. We show that, at some level, the migration of cells during all of these seemingly very different processes can be modelled by a nonlinear diffusion equation.

2 Epithelial Cell Migration

Epithelial tissues line the surfaces and cavities of structures throughout the embryo. During embryonic development, epithelial tissues undergo a complex series of deformations (see, for example, [1]) as the body shape is sculpted. The cells within the tissue can be visualised as polygons that share edges and vertices [2] and typically three or four cells will meet at any one vertex (see the reviews [3, 4]). However, there are cases where five or more cells meet at a vertex and the resulting structure has been given the name rosette [5]. In the mouse embryo, rosettes form in the visceral endoderm (VE, the epithelium that forms the outer layer of the egg-cylinder stage mouse embryo) due to the movement of a specialised subset of these cells known as the anterior visceral endoderm (AVE). For normal development, it is vital that these cells move as a coherent group. A recent study [6], on which this section is based, analysed the question of whether the formation of rosettes played an important role in this process.

2.1 Model

There are a number of ways to model cell movement in epithelial sheets. In [6], the so-called vertex-based model is chosen, in which each vertex in the system moves due to the forces acting upon it and the system is considered as overdamped (see also [4, 7]). The forces within this model framework are, to a large extent phenomenological, capturing the numerous forces acting on cells in a simple way. In particular, a tension force is considered to act along the edges of the cell, while a pressure force is considered to act outwards.

The egg-shaped embryo is considered to be a growing three-dimensional ellipsoid and each cell has a volume that changes over time, with a certain probability of division, depending on its size. The assumption is made that when a number of cell vertices come within a certain distance of each other, there is a certain probability that the cells will coalesce to form a rosette. The full details of the model are presented in [6].

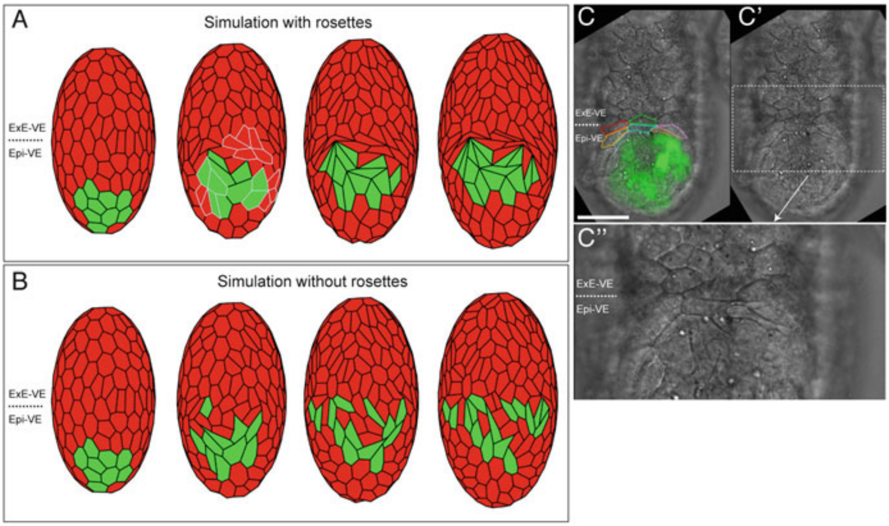


Fig. 1 Images taken at regular intervals of AVE cell migration (*light grey*) with (**A**) and without (**B**) rosettes. It can be seen that rosettes appear to facilitate the orderly and collective migration of the AVE while inhibition of rosette formation leads to a break-up of the AVE cell cluster (see [6] for full details). Note that the upper part of the ExE-VE is observed in experiments to be enshrouded in actin, and this is implemented in the model by making these cells very stiff. Compare these simulations with the experimental images in (**C**), (**C′**) and (**C″**). ExE-VE denotes visceral endoderm overlying the extra-embryonic ecotoderm: Epi-VE the visceral endoderm overlying the epiblast

Migration of the AVE cells is implemented in the model by assuming that there is a directed force on the leading vertex of the AVE cells enabling them to move upwards from the base of the embryo. The source of this force is unspecified as the issue at hand is to understand how rosettes affect the ability of the cells to undergo coherent migration, rather than how directed motion is achieved. Despite the phenomenological nature of the model, it was found that simulations agree qualitatively with data on the polygon distribution, which varies in space and time. An investigation of the difference in model predictions between embryos in which rosettes were and were not allowed to form was then carried out. This was encoded in the model by varying the probability of rosette formation. Remarkably, it was found that simulations without rosettes predicted the break-up of coherent AVE cell migration in a way which was very similar to that observed in a mutant mouse which does not form rosettes (see Figs. 1 and 2).

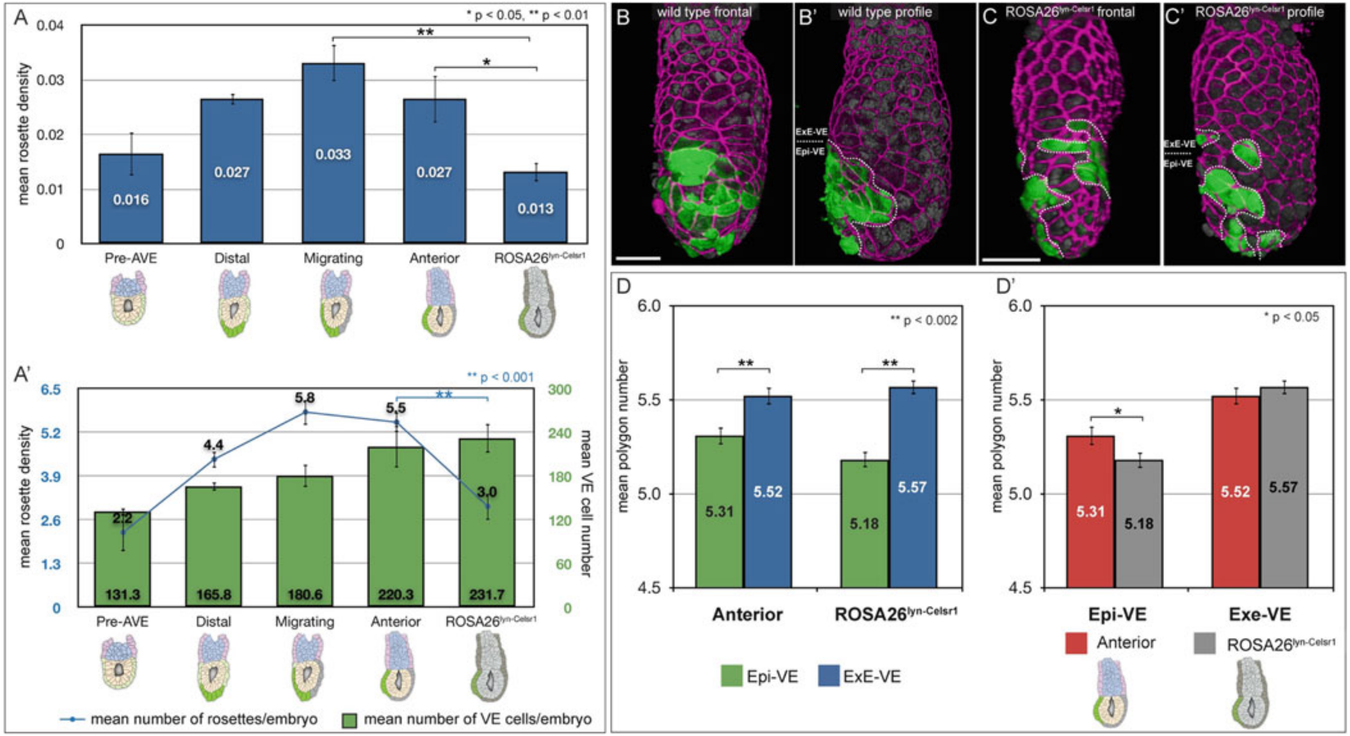


Fig. 2 (**A**) and (**A′**) show experimental data on rosette densities at different stages of AVE migration in a normal mouse. (**B**) and (**B′**) show two different views of a normal embryo while (**C**) and (**C′**) are the corresponding views of a mutant mouse which does not form rosettes. In both cases the AVE cells are in light grey and it can be seen that while in the normal mouse the AVE cells migrate in a coherent, ordered way, in the mutant they break up into small clusters as predicted by the model (Fig. 1). ExE-VE denotes visceral endoderm overlying the extra-embryonic ecotoderm: Epi-VE the visceral endoderm overlying the epiblast. (**D**) and (**D′**) show experimental data on the mean polygon number for each case

2.2 Other Approaches

There are, of course, other approaches to modelling epithelial sheet dynamics. For example, the spring-based model (see, for example, [8]) assumes cells to be point masses attached by springs. The cellular Potts model [9], on the other hand, assumes that cells are composed of elements which move to minimise a Hamiltonian representing an energy, and then there are the more sophisticated subcellular element models [10]. A natural question to ask is how do these models compare, but this is a difficult question to answer because they are not built from a common underlying framework. However, for the very simple case of a row cells, connected by linear springs, it can be shown that, in the overdamped case, the continuum limit of the discrete model is the nonlinear diffusion equation

$$\frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left[\frac{k}{\eta n^2}\frac{\partial n}{\partial x}\right], \tag{1}$$

where $n(x,t)$ is the normalised cell density at position x and time t, k is the spring constant and η is the cell viscosity [11]. For the cellular Potts model, where cells interact via a hard-core potential, it can be shown that in the macroscopic limit this, too, reduces to a nonlinear diffusion equation but with the form

$$\frac{\partial n}{\partial t} = \frac{\partial}{\partial x}\left[C\frac{n_0^2 + n^2}{(n_o - n)^2}\frac{\partial n}{\partial x}\right], \tag{2}$$

where n_o^{-1} is the average cell length and C is a constant [12]. This allows, for the first time, an informed comparison to be made between these modelling frameworks. In fact, in [13] the spring-based formalism is extended to include nonlinear springs, so that it is now possible to compare, in a consistent framework, cellular Potts models, spring-based models, and continuum models with phenomenologically hypothesised nonlinear diffusion coefficients. An open-source cell-based computational framework has also been established in which to compare these models [14, 15]. An open, and very challenging, question is how to derive the continuum limit for more complicated forms of cell-based models which incorporate biochemical signals [16].

3 Cranial Neural Crest Cell Migration

Cells migrate long distances during normal development and cancer metastasis, but little is understood about the mechanisms that control such behaviour. The paper [17] carries out an interdisciplinary study of one such example, cranial neural crest cell migration. Experiments are used to determine how the domain on which the cells migrate grows, and this is then incorporated into a hybrid model. Here a two-dimensional off-lattice individual-based model is used to describe cell migration,

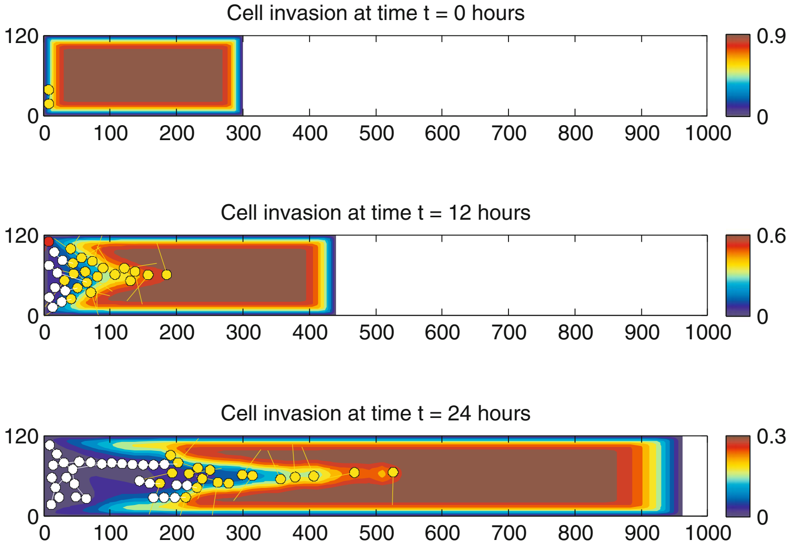


Fig. 3 A hybrid discrete cell-based model for neural crest cell invasion. Cells emerge from the neural tube (*right-hand boundary*) and consume VEGF, setting up a gradient (*dark shading*—low concentration of VEGF, *light shading*—high concentration of VEGF). A sub-group of the cells (*leaders*—at the front) move up the gradient in VEGF while the trailing group of cells (*followers*—at the back) respond to the leaders. The result of this is successful invasion of the domain. See [17, 18] for full details

and it is coupled with a PDE model for the dynamics of vascular endothelial growth factor (VEGF) concentration. It is assumed that VEGF is produced by the underlying tissue, diffuses and is consumed by cells which, in turn, send out protrusions (filopodia) to sample the local VEGF concentration. The cells then move in the direction of the filopodium if it samples a concentration higher than the local average (chemoattraction). Such a model is used to test a number of hypotheses on possible migration mechanisms. It is shown that the simplest hypothesis, namely that cells emerge from the neural crest and, via consumption of VEGF, set up a VEGF gradient up which they migrate, is not robustly successful, as later emerging cells are left behind due to VEGF depletion. From this emerges a new model hypothesis, the cells at the back respond, not to VEGF, but to cells at the front, contacting them and moving in the direction in which they are moving. Experimental analyses of this follower–leader scenario reveal that cells at the front express a number of different genes to those at the back, with genes responsible for exploratory motion upregulated at the front, while those encoding for adhesive motion being upregulated at the back, precisely as predicted by the mathematical model (Fig. 3[1]).

[1] We thank Louise Dyson for providing this figure.

3.1 Volume Exclusion

To begin to understand the dynamics of the above type of model, the paper [19] considered a very simplified off-lattice individual-based model with volume exclusion. Using a master equation approach and taking a parabolic limit of the resulting conservation equation the authors showed how, in the macroscopic limit, a position jump process with volume exclusion could be written as a diffusion equation with a nonlinear diffusion coefficient. The particular details of the diffusion coefficient depend on the details of the jump probability density function. For example, if individuals hopped with rate α a constant distance d to the left or right, then the macroscale diffusion equation, under certain simplifying assumptions, takes the form

$$\frac{\partial n}{\partial t} = D\frac{\partial}{\partial x}\left[1 + 4R\frac{N-1}{N}n\right], \tag{3}$$

where $n(x,t)$ is the average total cell density at position x and time t, N is the total cell number, D is $\alpha D^2/2$ (in the limit of d tending to 0), and R is the cell radius (assumed the same for all cells). Note that in this case, crowding actually enhances diffusion but of course, as N increases the limiting procedure becomes invalid and in fact we reach the jamming limit, where crowding inhibits movement.

On the other hand, if it is assumed that the distance moved is normally distributed with zero mean and variance σ^2, then the diffusion equation, under the same simplifying assumptions, takes the form

$$\frac{\partial n}{\partial t} = \frac{\alpha\sigma^2}{2}\frac{\partial}{\partial x}\left[1 + \frac{N-1}{N}\left(4R - 2\sigma\sqrt{\frac{2}{\pi}}\right)n\right]. \tag{4}$$

One can see that, although this is a very different model and is based on very different biology, in the continuum limit, it leads, as for the case in the previous section, to a nonlinear diffusion equation for macroscopic (global) behaviour, in which the microscopic (local) details are encapsulated in the precise form of nonlinear diffusion coefficient.

4 Acid-Mediated Invasion Hypothesis

In 1996, Gatenby and Gawlinski [20] proposed a novel model for tumour cell invasion. They investigated the seemingly paradoxical phenomenon of tumour cells undergoing glycolytic metabolism even in the presence of oxygen (known as the Warburg effect [21]). If one considers the body as an ecosystem, with cancer cells attempting to invade the host (normal cells), then in the resultant competition, it is puzzling as to why tumour cells, in the presence of oxygen, do not undergo aerobic respiration, which is much more energy efficient than anaerobic (glycolytic) metabolism. They made the hypothesis that tumour cells gained a competitive advantage in doing so because a byproduct of glycolysis, lactic acid, is more toxic to normal cells than to cancer cells. In effect, the cancer cells change the environment to gain an advantage. They showed that this model could give rise to travelling waves of invasion which, in certain parameter regimes, predicted that there would be an acellular gap between the invading tumour front and the regressing normal cells. They validated this model prediction in the laboratory. The travelling waves exhibited by this model were mathematically analysed in [22] where it was shown that the system exhibited both slow and fast waves.

More recently, the paper by McGillen et al. [23] considered an extended version of the above model which assumed that the tumour cells were not totally resistant to acid (a more realistic description). The model, in one spatial dimension for simplicity, takes the form

$$\frac{\partial U}{\partial t} = \rho_1 U \left(1 - \frac{U}{\kappa_1} - \alpha_2 \frac{V}{\kappa_1}\right) - \delta_1 U W, \tag{5}$$

$$\frac{\partial V}{\partial t} = \rho_2 V \left(1 - \frac{V}{\kappa_2} - \alpha_1 \frac{U}{\kappa_2}\right) - \delta_2 V W + \frac{\partial}{\partial x}\left[D_2 \left(1 - \frac{U}{\kappa_1}\right)\frac{\partial V}{\partial x}\right], \tag{6}$$

$$\frac{\partial W}{\partial t} = \rho_3 V - \delta_3 W + D_3 \frac{\partial^2 V}{\partial x^2}, \tag{7}$$

where $U(x,t)$, $V(x,t)$, and $W(x,t)$ are, respectively, normal tissue density, cancer cell density, and excess acid concentration at position x and time t. All the parameters are non-negative and constant.

The first equation assumes that normal cells grow logistically in the absence of tumour cells, compete with the tumour cells, and die due to the presence of excess acid. It also assumes that normal cells do not move which, in the adult, is a biologically realistic assumption. The second equation models growth of tumour cells in a way similar to that of normal cells, but allows the tumour cells to diffuse when there is space available. The third equation assumes that excess lactic acid growth is linearly dependent on tumour cell density and that it degrades linearly and diffuses.

This model recaptures the key results of the original paper (Fig. 4). Exploiting the existence of a small parameter in this system (the ratio of tumour cell diffusion coefficient to that of lactic acid) in [23], a very detailed perturbation analysis was used to show that, in the asymptotic limit, the problem reduces to the famous Fisher–KPP problem (see, for example, [24]). Using the biological literature to estimate as many parameters as possible, a detailed analysis was carried out of the remaining parameter space to determine under what biologically realistic conditions the model could predict a gap of the size that was observed experimentally. It was found that the parameters had to be very finely tuned to achieve this behaviour. However, a larger parameter space was found to permit the existence of a smaller gap. This is in agreement with the original experiments, where the authors were aware of the fact that the gap they observed may be an experimental artefact (caused by fixing). However, when they did not fix the tissue they still observed a gap, but it was smaller, which is consistent with the model calculations.

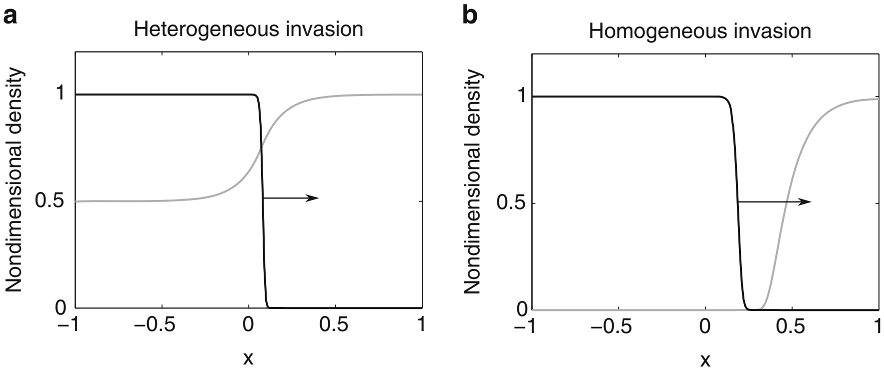


Fig. 4 The acid-mediated tumour cell invasion model predicts the possibilities of an initially small compact tumour invading the host tissue either leading to coexistence (**a**) or competitive exclusion (**b**). Note that in (**b**) there is also an acellular gap between the two cell types. *Dark lines* represent tumour cell density and *light lines* normal cell density. See [23] for full details

This study suggests that in mice, bicarbonate treatment can significantly reduce metastatic tumour invasion and this has been experimentally validated [25]. Whether or not such a treatment would be effective in humans remains controversial [26].

5 Conclusions

Biological systems are, of course, extremely complex, with myriad processes interacting across a multitude of spatial and temporal scales. The inherent nonlinearities present in biology mean that the traditional verbal reasoning approaches used within the field can lead to incorrect conclusions. Therefore, it is necessary to address these problems mathematically. While bioinformatics has already had a huge impact on biology through using statistical approaches for data mining, the recent technological advances that have led to acquisition of spatiotemporal data now mean that dynamical models can be tested and validated. There are conflicting philosophies in the latter approach. One camp has the view that large multiscale models should be built to try to capture as many features as possible of the system being modelled. The other camp favours small models, developed to answer specific biological questions. This paper reviews three problems addressed using the latter approach. It is shown that the formation of multicellular rosettes in the mouse embryo may facilitate orderly migration of cells, that cranial neural crest cell invasion requires different cell phenotypes, and that the glycolytic phenotype can lead to acid-mediated tumour cell invasion.

All the above models are based on very different biological hypotheses and very different mathematical/computational modelling frameworks. However, we see that, in some limits, all these models share the same unifying underlying mathematical structure, namely reaction–diffusion equations with a nonlinear diffusion coefficient. These equations describe the global, macroscopic (coarse-grained) dynamical behaviour of the system, where the local, microscopic (cell)-level behaviour manifests itself in the specific form of the nonlinear diffusion coefficient.

References

1. Odell, G.M., Oster, G., Alberch, P., Burnside, B.: The mechanical basis of morphogenesis. Dev. Biol. **85**, 446–462 (1981)
2. Honda, H.: Description of cellular patterns by Dirichlet domains: the two-dimensional case. J. Theor. Biol. **72**, 523–543 (1978)
3. Vincent, J.-P., Fletcher, A.G., Baena-Lopez, L.A.: Mechanisms and mechanics of cell competition in epithelia. Nat. Rev. Mol. Cell Biol. **14**, 581–591 (2013)
4. Fletcher, A.G., Osterfield, M., Baker, R.E., Shvartsman, S.Y.: Vertex models of epithelial morphogenesis. Biophys. J. **106**(11), 2291–2304 (2014)
5. Blankenship, J.T., Backovic, S.T., Sanny, J.S., Weitz, O., Zallen, J.A.: Multicellular rosette formation links planar cell polarity to tissue morphogenesis. Dev. Cell **11**, 459–470 (2006)
6. Trichas, G., Smith, A.M., White, N., Wilkins, V., Watanabe, T., Moore, A., Joyce, B., Sugnaseelan, J., Rodriguez, T.A., Kay, D., Baker, R.E., Maini, P.K., Srinivas, S.: Multi-cellular rosettes in the mouse visceral endoderm facilitate the ordered migration of anterior visceral endoderm cells. PLoS Biol. **10**(2), e1001256 (2012)
7. Farhadifar, R., Roper, J.C., Aigouy, B., Eaton, S., Julicher, F.: The influence of cell mechanics, cell-cell interactions, and proliferation on epithelial packing. Curr. Biol. **17**, 2095–2104 (2007)
8. Meineke, F., Potten, C.S., Loeffler, M.: Cell migration and organization in the intestinal crypt using a lattice-free model. Cell Prolif. **34**(4), 253–266 (2001)
9. Graner, F., Glazier, J.A.: Simulation of biological cell sorting using a two-dimensional extended Potts model. Phys. Rev. Lett. **69**(13), 2013–2016 (1992)
10. Newman, T.J.: Modeling multi-cellular systems using sub-cellular elements. Math. Biosci. Eng. **2**, 611–622 (2005)
11. Murray, P.J., Edwards, C.M., Tindall, M.J., Maini, P.K.: From a discrete to a continuum model of cell dynamics in one dimension. Phys. Rev. E **80**, 031912 (2009)
12. Lushnikov, P.M., Chen, N., Alber, M.: Macroscopic dynamics of biological cells interacting via chemotaxis and direct contact. Phys. Rev. E **78**, 061904 (2009)
13. Murray, P.J., Edwards, C.M., Tindall, M.J., Maini, P.K.: Classifying general nonlinear force laws in cell-based models via the continuum limit. Phys. Rev. E **85**, 021921 (2012)
14. Pitt-Francis, J., Pathmanathan, P., Bernabeu, M.O., Bordas, R., Cooper, J., Fletcher, A.G., Mirams, G.R., Murray, P., Osborne, J.M., Walter, A., Chapman, S.J., Garny, A., van Leeuwen, I.M., Maini, P.K., Rodriguez, B., Waters, S.L., Whiteley, J.P., Byrne, H.M., Gavaghan, D.J.: Chaste: a test-driven approach to software development for biological modelling. Comput. Phys. Comm. **180**, 2542–2471 (2009)
15. Murray, P.J., Walter, A., Fletcher, A.G., Edwards, C.M., Tindall, M.J., Maini, P.K.: Comparing a discrete and continuum model of the intestinal crypt. Phys. Biol. **8**, 026011 (2011)
16. Murray, P.J., Kang, J.-W., Mirams, G.R., Shin, S.-Y., Byrne, H.M., Maini, P.K., Cho, K.-H.: Modelling spatially regulated β-catenin dynamics and invasion in intestinal crypts. Biophys. J. **99**, 716–725 (2010)
17. McLennan, R., Dyson, L., Prather, K.W., Morrison, J.A., Baker, R.E., Maini, P.K., Kulesa, P.M.: Multiscale mechanisms of cell migration during development: Theory and experiment. Development **139**, 2935–2944 (2012)
18. Dyson, L.: Models of cranial neural crest cell migration. D.Phil. thesis, University of Oxford (2013)

19. Dyson, L., Maini, P.K., Baker, R.E.: Macroscopic limits of individual-based models for motile cell populations with volume exclusion. Phys. Rev. E **86**, 031903 (2012)
20. Gatenby, R.A., Gawlinski, E.T.: A reaction-diffusion model of cancer invasion. Cancer Res. **56**, 5745–5753 (1996)
21. Warbug, O.: The Metabolism of Tumors. Arnold Constable, London (1930)
22. Fasano, A., Herrero, M.A., Rodrigo, M.R.: Slow and fast invasion waves in a model of acid-mediated tumour growth. Math. Biosci. **220**, 45–56 (2009)
23. McGillen, J.B., Gaffney, E.A., Martin, N.K., Maini, P.K.: A general reaction-diffusion model of acidity in cancer invasion. J. Math. Biol. **68**, 1199–1224 (2014)
24. Murray, J.D.: Mathematical Biology II: Spatial Models and Biomedical Applications. Springer, New York (2003)
25. Robey, I.F., Baggett, B.K., Kirkpatrick, N.D., Roe, D.J., Dosescu, J., Sloane, B.F., Gatenby, R.A., Raghunand, N., Gillies, R.J.: Bicarbonate increases tumor pH and inhibits spontaneous metastases. Cancer Res. **69**, 2260–2268 (2009)
26. Martin, N.K., Gaffney, E.A., Gatenby, R.A., Gillies, R.J., Robey, I.F., Maini, P.K.: A mathematical model of tumour and blood pHe regulation: the $HCO3^{-}/CO2$ buffering system. Math. Biosci. **230**, 1–11 (2011)

Modelling Oxygen Capillary Supply to Striated Muscle Tissues

A.A. Al-Shammari, E.A. Gaffney, and S. Egginton

Abstract The ability to characterise functional capillary supply (FCS) plays a key role in developing effective therapeutic interventions for numerous pathological conditions, such as chronic ischaemia in skeletal or cardiac muscle. Detailed tissue geometry, such as muscle fibre size, has been incorporated into indices of FCS by considering the distribution of Voronoi tessellations ('capillary domains') generated from vessel locations in a plane perpendicular to muscle fibre orientation, implicitly assuming that each Voronoi polygon represents the area of supply of its enclosed capillary. However, to assess the capacity of FCS in muscle, we are naturally led to use a modelling framework that can account for the local anatomic and metabolic heterogeneities of muscle fibres. Such a framework can be used to explore the validity of the Voronoi polygon representation of FCS regions while also providing a general platform for robust predictions of FCS.

Keywords Mathematical modelling • Oxygen transport • Capillary supply • Capillary domains • Voronoi polygons • Trapping regions

A.A. Al-Shammari (✉)
Wolfson Centre for Mathematical Biology, Mathematical Institute,
University of Oxford, Oxford OX2 6GG, UK

Department of Mathematics, Faculty of Sciences, Kuwait University,
P.O. Box 5969, Khaldiya 13060, Kuwait
e-mail: alshammari@maths.ox.ac.uk

E.A. Gaffney
Wolfson Centre for Mathematical Biology, Mathematical Institute,
University of Oxford, Oxford OX2 6GG, UK

S. Egginton
School of Biomedical Sciences, Faculty of Biological Sciences,
University of Leeds, Leeds LS2 9JT, UK

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_2,

1 Introduction

The availability of energy within striated muscle cells (fibres) is essential for sustaining healthy function. The cellular preference for high energy aerobic metabolism necessitates a continuous supply of oxygen (O_2) for matching the local cellular demand. Such a match is ensured by allowing adequate O_2 delivery from the microcirculation and through a local capillary bed. In particular, capillaries provide the terminal sites for O_2 delivery to and metabolite waste removal from cells, where O_2 diffuses passively across capillary walls and into tissue to meet the local cellular demand (Fig. 1a, b). Hence a healthy capillary supply is essential for healthy tissue function, thus highlighting the importance of capillary distributions for adequate tissue oxygenation.

Capillary delivery of oxygen is a major limiting factor in the oxygen transport pathway to muscle tissue, especially in the presence of vascular and tissue pathologies. For example, *ischaemia*, a vascular disease involving a restriction in arterial blood supply to tissues (e.g. coronary artery disease), leads to a vascular shortage in oxygen (hypoxemia), which, if left untreated, can further lead to insufficient tissue O_2 supply (hypoxia), complete deprivation of O_2 supply (anoxia), and ultimately necrosis (tissue death). In particular, according to recent estimates from the World Health Organization, ischaemic heart disease is the leading cause of global human death [3]. While treatment from chronic ischaemia in skeletal and cardiac muscles would certainly benefit from a local enhancement of functional capillary supply (FCS) of oxygen by inducing capillary growth (*angiogenesis*) to match the local tissue demand, we still lack a complete understanding of such interventions. However, even quantifying FCS is fraught with difficulties. While measures of gross capillary supply may highlight a global tissue ischaemia [4], their spatial resolution cannot capture the local tissue pathologies associated with the underlying capillary distribution. At such local resolutions, analyses based on

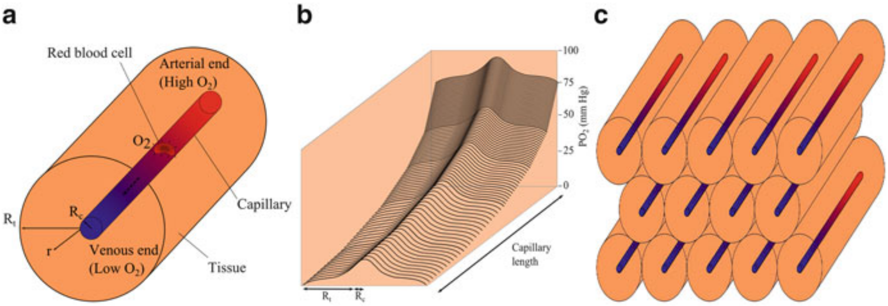


Fig. 1 (**a**) Traditional view of tissue oxygenation. Estimation of PO_2 within a circular cylinder of tissue surrounding a capillary of radius R_c; r is the distance from the capillary centre; R_t denotes the cylinder radius where oxygen flux becomes zero [1]. (**b**) PO_2 at capillary declines monotonically both around and within the fibre; the minimum PO_2 is at the centre of a fibre [2]. (**c**) Krogh's view of circular tissue cylinder stacking, where tissue supply voids are inevitable

conventional FCS measures can give conflicting results [5], thus potentially leading to poor interpretations of experimental findings.

There has been a growing interest in improving the classification of FCS to tissue and using measures that take into account the local anatomical and metabolic details in experimental studies seeking to assess the extent and location of angiogenesis in striated muscle tissues [2, 6, 7]. Recognising the importance of such attempts, we present a brief account that highlights the modelling developments seeking to quantify the regions of muscle tissue exclusively supplied by individual capillaries as a basis for analysing FCS.

2 Theory

2.1 Krogh Cylinder

The idea of quantifying capillary supply by assigning a region of tissue to each capillary was initially conceived by August Krogh in 1919 [1] and subsequently led to his Nobel Prize in physiology. Essentially a *capillary supply region* was defined as the extent of tissue volume diffusively supplied by a capillary. Based on anatomical observations, each capillary was assumed to concentrically supply a hexagonal cylinder. This was further conveniently reduced to an circular cylinder (Krogh Cylinder) with a predefined radius (Fig. 1a), thus leading to a 3D arrangement where capillaries parallel to skeletal muscle fibre axes are symmetrically distributed with their Krogh cylinders stacked evenly (Fig. 1c), inevitably giving rise to tissue supply voids. Along with other simplifications [8], these led to a simple 1D steady-state diffusion problem for oxygen tension, p, with the solution (Krogh–Erlang equation)

$$p(r) = p(R_c) - \frac{M_0}{4K}\left[R_t^2 \log \frac{r^2}{R_c^2} - (r^2 - R_c^2)\right],$$

where R_t and R_c are the tissue and capillary radii with $R_c \leq r \leq R_t$, K is Krogh's O_2 diffusion coefficient in tissue, and M_0 is a constant tissue demand for O_2. Combining experimental measurements and geometrical observations of the microvasculature with this formula has led to estimates of the minimum tissue oxygen tension and capillary density [1, 9].

2.2 Capillary Domains

Krogh's attempt to close pack circular tissue cylinders has led to tissue voids where diffusive supply was geometrically excluded. Gonzalez-Fernandez and Atta [9] addressed this by reformulating Krogh's original problem to allow for oxygen supply to the entire domain *via* hexagonal, square, and triangular tissue cylinders.

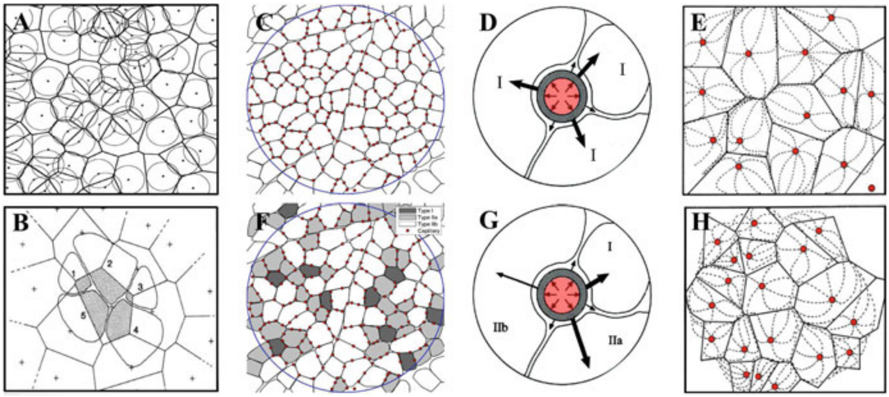


Fig. 2 (**a**) Digitised rat EDL muscle section showing capillary locations (*black dots*), capillary domains (DOM, polygons), and Krogh cylinders (*circles*). (**b**) Fibres partition capillary supply unambiguously by overlapping DOM. (**c**, **d**) Uniform muscles have only one type of muscle fibre (e.g. Type I) with spatially homogeneous tissue oxygen demand (MO_2). (**e**) If capillaries (*red discs*) have identical transport capacity, the predicted O_2 flux lines (*dotted lines*) coalesce at the no-flux points that match DOM boundaries (*solid lines*). (**f**) Mixed muscles have at least three distinct fibre types (I, IIa, and IIb) with distinct MO_2. (**g**) O_2 diffusion depends on the local extraction pressures established by differences in MO_2. (**h**) Given any capillary may be surrounded by distinct fibres, the heterogeneity in fibre composition and MO_2 reduces the fit between no-flux and DOM boundaries for mixed muscles. The model geometry is obtained by considering a muscle tissue cross section (Fig. **c**, **f**). The tissue region excluding capillaries is denoted by Ω with an external boundary $\partial\Omega$. Capillaries, Ω_i, are treated as circular inclusions within the tissue with a boundary $\partial\Omega_i$ and a uniform radius. Data from [2, 11, 12], with permission

This essentially marked the first formal attempt for modelling capillary supply regions as *capillary domains* (DOM). While the use of such domains had clearly solved the tissue void problem, it still maintained the assumption that capillary arrangements within tissue are highly symmetrical. In contrast, capillaries in skeletal muscles are often asymmetrically distributed, thus breaking the symmetry of Krogh's cylinders.

Hoofd and colleagues [10] tackled this question by generalising the symmetry in Krogh's geometrical formalism by allowing each capillary to have a distinct edge of symmetry with each of its neighbours (the bisector of the line connecting neighbouring capillaries). Such a construction identified DOM with the Voronoi tessellation [4, 10] of capillary locations in the plane perpendicular to muscle fibre orientation (see polygons in Fig. 2a). Consequently, the tissue cylinders formed by DOM may have distinct geometries (loss of symmetry), indicating that the Krogh–Erlang equation will assume different solutions for geometrically distinct tissue cylinders. In addition, an 'equivalent' Krogh cylinder, whose cross-sectional area is identically set to the average capillary domain, was alternatively used for all capillaries (compare cylinders to polygons in Fig. 2a). However, the large voids and overlaps associated with these cylinders highlight the inadequacy of using Krogh cylinders to represent regions of capillary supply.

As noted previously, within the framework of DOM, capillaries supply the tissue regions nearest to them, thereby generating a complete tessellation of the tissue plane. This, in turn, allows the detailed anatomical geometry to be incorporated into measures of FCS by considering the overlap of DOM with muscle fibres [2], implicitly assuming that a capillary domain represents the diffusive area of supply of its enclosed capillary (Fig. 2b). However, such geometrical constructs are still simplifications to the diffusive supply regions, which may well be affected by spatial heterogeneities of capillaries and oxygen uptake (Fig. 2f–h).

2.3 Flux Trapping Regions

Hoofd and colleagues [12] assessed the accuracy of DOM by taking capillaries to be O_2 point sources, which in turn led to an analytical expression for O_2 flux. For a capillary distribution embedded in a striated muscle with spatially uniform oxygen uptake (Fig. 2c, d), e.g. cardiac muscle, they found that DOM accurately capture the predicted flux lines (Fig. 2e; [2, 12]). However, it was not clear whether this representation will generalise to all striated muscle tissues, especially in the presence of a feedback between capillaries and tissue. For example, asymmetries in the spatial distribution of capillaries and blood oxygen content as well as heterogeneities in intracellular metabolic and diffusive characteristics are expected to affect the flux of oxygen at the prescribed boundaries of DOM (Fig. 2f–h). In addition, a recent mathematical exploration of this problem has led to the conclusion that DOM are inaccurate for capillary supply representation [13], though based on predictions that were heavily influenced by boundary conditions [14]. Hence, this leaves the question of whether DOM are appropriate in physiological settings.

3 Mathematical Model

Here we present a brief description of our recent mathematical modelling framework which was aimed at assessing the capillary domain approximation and generalising it to capture tissue heterogeneities.

Under maximal aerobic capacity, O_2 transport is effectively 2D and governed by Michaelis–Menten O_2 consumption within muscle fibres, free O_2 diffusion, and O_2-facilitated diffusion by myoglobin (a protein carrier). Averaged intravascular dynamics is fed into the model through a Robin boundary condition at the capillary wall.

Striated muscle tissues are composed of two distinct regions: (1) interstitial spaces and (2) muscle fibres. In addition, muscle fibres can have different intracellular composition which leads to further local specialisations giving rise to distinct fibre types (I, IIa, and IIb). Letting Ω denote the tissue domain exclusive of capillaries (Ω_i), with external boundary $\partial\Omega$, we seek to explore the 2D profile of oxygen tension (PO$_2$) in Ω (Fig. 2c, f)

$$\nabla\cdot\left[\underbrace{D(x)\nabla(\alpha(x)p)}_{\text{free diffusive flux}}+\underbrace{C^{Mb}(x)D^{Mb}(x)\left(\frac{dS_{Mb}}{dp}\nabla p\right)}_{\text{myoglobin-facilitated flux}}\right]=\underbrace{M(x,p),}_{\text{Tissue consumption}}\quad x\in\Omega, \tag{1}$$

$$n_i\cdot\left[\alpha(x)D(x)\nabla p\right]=k\left(p_{cap}-p\right),\ x\in\partial\Omega_i, \tag{2}$$

$$n_{\text{tissue}}\cdot\left[\alpha(x)D(x)\nabla p\right]\Big|_{\partial\Omega}=0, \tag{3}$$

$$S_{Mb}(p)=\frac{p}{p+p_{50,Mb}},\ M(x,p)=\frac{M_0(x)p}{p+p_c}, \tag{4}$$

where D and α are the molecular diffusivity and solubility of free oxygen, C^{Mb} and D^{Mb} are the bulk myoglobin (Mb) concentration and diffusivity, S_{Mb} is the equilibrium O$_2$ saturation of Mb, $p_{50,Mb}$ is the tissue oxygen partial pressure at half Mb saturation, M is the rate of O$_2$ consumption in muscle tissue, M_0 is the maximal consumption rate (VO$_{2\text{max}}$) of a muscle fibre, and p_c is the tissue PO$_2$ value which reflects the partial pressure scale where fibre mitochondria are no longer able to extract oxygen at maximal rate. Parameter values are detailed in [11, 14].

4 Computational Solution

4.1 PO$_2$, Oxygen Flux, and Trapping Regions

A direct numerical exploration of the oxygen transport problem within tissue cross sections can be pursued via image capture, overlaying a mesh which is faithful to the geometry captured from biopsies and refined within regions of complex geometry (see Fig. 3a–c). This allows a numerical solution of our oxygen transport equations, which capture the biophysics of oxygen delivery while accounting for histological detail. However, the complexity at the microvascular level limits the length scales which may be readily explored in this manner, especially for 3D simulations or for simulations within a large parameter space.

To determine the supply regions of our model (trapping regions, TR; Fig. 3d), oxygen flux can be first computed by solving the gradient dynamical system, $\frac{d\mathbf{x}}{ds}=\nabla p$ where $\mathbf{x}(s)$ is a parameterisation of the trapping region boundary, via Heun's method. The Hartman–Grobman theorem can then be employed to estimate TR as detailed in [14].

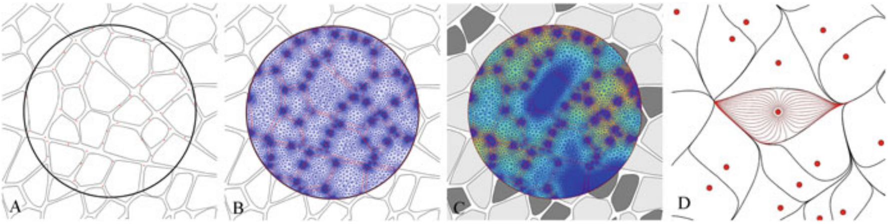


Fig. 3 Computational framework. (**a**) Post-segmentation digitised image of tissue cross section. (**b**) Finite element mesh generation. (**c**) Numerical solution to (1)–(4) with fibre-specific parameters using Matlab's PDE Toolbox [15]. (**d**) PO_2 flux lines (*red*) generated for each capillary (*disc*) by numerically solving $\frac{d\mathbf{x}}{ds} = -\nabla p$ with trapping regions delimited (*black*), where s parameterises the flux lines

4.2 Capillary Domains vs Trapping Regions

Using the above framework we can qualitatively and quantitatively assess the area of capillary supply in the presence of heterogeneities (Fig. 4). For example, DOM are a generally accurate approximation of TR (Figs. 4a–c), with lower accuracy correlating with increased spatial heterogeneities of capillary locations (Fig. 4i). Nonetheless, DOM breakdown in the presence of significant capillary rarefaction (Fig. 4d). In addition, increasing the metabolic heterogeneity further accentuates DOM's inaccuracy (Figs. 4e–h, j). In particular, the heterogeneity in capillary arrangements is observed to have a much more pronounced effect on the accuracy of DOM than that of metabolic heterogeneities.

5 Discussion

Voronoi tessellations (capillary domains) may be a useful method for assessing oxygen capillary supply in homogeneous tissue, but their use may be problematic in the presence of extensive capillary rarefaction (functional and structural). Calculation of diffusive oxygen fluxes provides a computationally more intensive alternative. In cases of heterogeneous perfusion, such trapping regions provide a more general representation of capillary supply regions. In addition, this approach will allow incorporation of additional influences of heterogeneity that are absent in the consideration of capillary domains, such as differences in local metabolism or muscle fibre size. Therefore, trapping regions may be used to better inform experimental studies assessing microvascular and tissue dysregulations and pathologies.

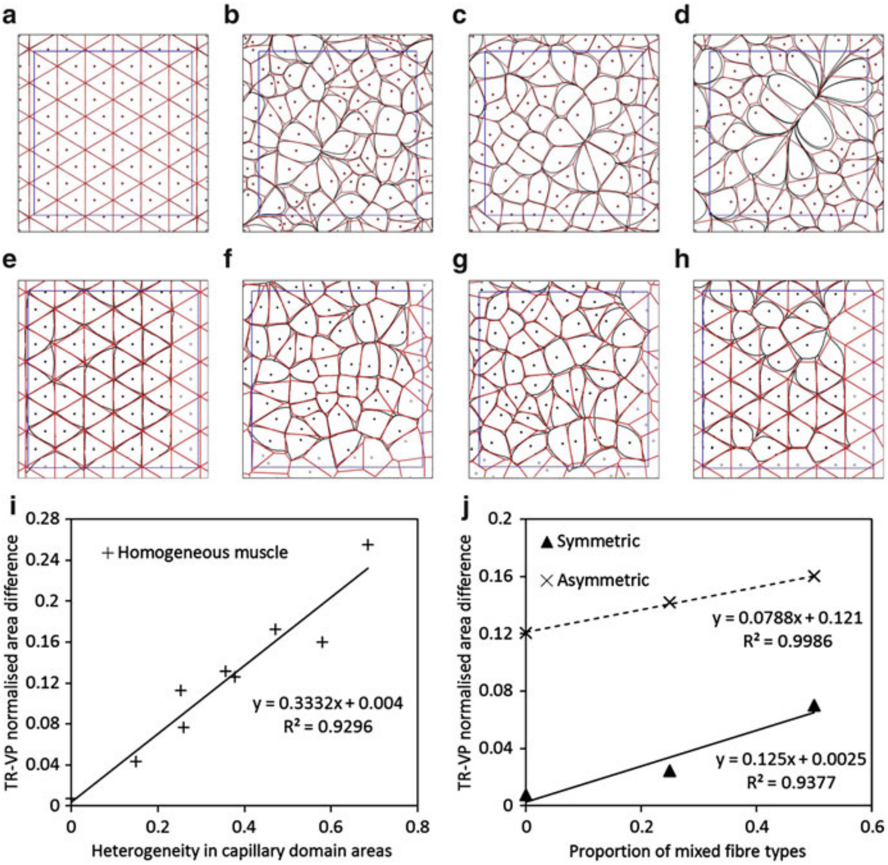


Fig. 4 Investigation of the effect of structural and metabolic heterogeneities on the correlation between capillary domains (DOM; *red*) and trapping regions (TR; *black*). Capillary arrangement is symmetric (a,e), asymmetric (b,f), extensor digitorum longus muscle (c,g), or rarefied (d,h). Oxygen demand is homogeneous in (a–d), and heterogeneous in (e–h). Plots of the difference between DOM and TR are given for variation in (i) the spread of DOM areas in homogeneous muscle, and (j) the proportion of mixed fibres in heterogeneous muscle. Data from [11, 14] with permission

Acknowledgements AA is supported by a studentship from Kuwait University, Kuwait, and by travel grants from the Mathematical Institute and Linacre College, University of Oxford, UK.

References

1. Krogh, A.: The number and distribution of capillaries in muscles with calculations of the oxygen pressure head necessary for supplying the tissue. J. Physiol. **52**, 391–408, 409–415, 457–474 (1919)

2. Egginton, S., Ross, H.F.: Planar analysis of tissue capillary supply. In: Oxygen Transport in Biological Systems. Society for Experimental Biology Seminar Series, vol. 51, pp. 165–195. Cambridge University Press, Cambridge (1992)
3. World Health Organization: The top 10 causes of death. http://www.who.int/mediacentre/factsheets/fs310/en/. Accessed 22 Dec 2013
4. Egginton, S.: Morphometric analysis of tissue capillary supply. In: Boutilier, R.G. (ed.) Vertebrate Gas Exchange from Environment to Cell. Advances in Comparative and Environmental Physiology, vol. 6, pp. 73–141. Springer, Berlin (1990)
5. Egginton, S., Gaffney, E.A.: Tissue capillary supply—it's quality not quantity that counts! Exp. Physiol. **95**(10), 971–979 (2010)
6. Degens, H., Deveci, D., Botto-Van Bemden, A., Hoofd, L.J.C, Egginton, S.: Maintenance of heterogeneity of capillary spacing is essential for adequate oxygenation in the soleus muscle of the growing rat. Microcirculation **13**, 467–476 (2006)
7. W*ü*st, R.C.I., Gibbings, S.L., Degens, H.: Fiber capillary supply related to fiber size and oxidative capacity in human and rat skeletal muscle. In: Liss, P., Hansell, P., Bruley, D.F., Harrison, D.K. (eds.) Oxygen Transport to Tissue XXX. Advances in Experimental Medicine and Biology, vol. 645, pp. 75–80. Springer, New York (2009)
8. Kreuzer, F.: Oxygen supply to tissues: the krogh model and its assumptions. Experientia **38**, 1415–1426 (1982)
9. Gonzalez-Fernandez, J.M., Atta, S.E.: Concentration of oxygen around capillaries in polygonal regions of supply. Math. Biosci. **13**, 55–69 (1972)
10. Hoofd, L., Turek, Z., Kubat, K., Ringnalda, B.E.M., Kazda, S.: Variability of intercapillary distance estimated on histological sections of rat heart. Adv. Exp. Med. Biol. **191**, 239–247 (1985)
11. Al-Shammari, A.A., Gaffney, E.A., Egginton, S.: Modelling capillary oxygen supply capacity in mixed muscles: Capillary domains revisited. J. Theor. Biol. **356**, 47–61 (2014), DOI: 10.1016/j.jtbi.2014.04.016
12. Hoofd, L., Turek, Z., Olders, J.: Calculation of oxygen pressures and fluxes in a flat plane perpendicular to any capillary distribution. In: Rakusan, K., Biro, G., Goldstick, T.K., Turek, Z. (eds.) Oxygen Transport to Tissue XI, pp. 187–196. Plenum Press, New York (1989)
13. Wang, C.Y., Bassingthwaitghte, J.B.: Capillary supply regions. Math. Biosci. **173**, 103–114 (2001)
14. Al-Shammari, A.A., Gaffney, E.A., Egginton, S.: Modelling capillary oxygen supply capacity in mixed muscles. Capillary domains revisited. J. Theor. Boil. **356**, 47–61 (2014). doi:10.1016/j.jtbi.2014.o4.
15. The Mathworks, Inc.: Partial differential equations toolbox: user's guide (R2013b). http://www.mathworks.co.uk/help/pdf_doc/pde/pde.pdf (2013). Accessed 2 Nov 2013

Modeling Human Response to Bed–Net Promotion Campaigns and Its Impact on Malaria Transmission

Bruno Buonomo

Abstract We consider a malaria model including human response to health-promotion campaign for bed-net usage. We propose a formulation of the human–mosquito contact rate which is based on the idea of information-dependent epidemic models. We show that the model allows to easily determine optimal control strategies for implementing health campaigns. Moreover, the controlled system may predict a dramatic reduction of malaria incidence even when the uncontrolled system predicts stable endemicity.

Keywords Malaria • Mathematical model • Optimal control • Bed nets

1 Introduction

Malaria is a mosquito-borne infectious disease caused by parasites transmitted to susceptible humans through the bites of infected female mosquitoes of the genus *Anopheles*. Recent reports indicate that in spite of a substantial reduction of reported malaria cases and deaths in the last years, malaria is still a global emergency, with 3.3 billion people worldwide at risk of acquiring the disease in 2011 [1].

Mathematical modeling of malaria transmission, as part of the necessary multidisciplinary research approach, plays an important role for the understanding of malaria dynamics and the best strategies to control the disease [2–5].

Recently, the usage of non-pharmaceutical interventions (NPIs) for malaria control has received much attention from modelers. Such interventions aim to limit the disease spread by reducing the contacts between infectious and susceptible individuals [6]. Among the NPIs, the insecticide-treated bed nets (ITNs) are the

B. Buonomo (✉)
Department of Mathematics and Applications, University of Naples Federico II,
via Cintia, I-80126 Naples, Italy
e-mail: buonomo@unina.it

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_3,

most prominent malaria preventive measure for large-scale deployment in highly endemic areas [7]. An intriguing aspect of ITN usage is that its effectiveness is largely influenced by behavioral factors. Improper handling, nuisance and discomfort in using them, or simply personal habits and convictions may be reasons for not using ITNs [8]. Therefore, a modeling approach in the framework of *Behavioral Epidemiology*, where the key aspect is the impact of human behavior on epidemics [9], seems to be appropriate when assessing the impact of ITN usage to malaria transmission.

In a recent paper, Agusto et al. [10] proposed a malaria model where ITN usage is assumed to increase mosquito mortality and to reduce the human–mosquito contact rate (i.e. the average number of bites per mosquito per unit time, denoted here by β). This last effect is represented by the relation

$$\beta(b) = \beta_{\max} - b\,(\beta_{\max} - \beta_{\min})\,, \tag{1}$$

where $\beta_{\max}$ and $\beta_{\min}$ are the maximum and the minimum contact rate, respectively, and b is the proportion of ITN usage. The parameter b is a positive constant, the value of which may range from 0 (no ITN usage) to 1 (the whole population is protected by ITNs). As a consequence, the contact rate ranges between $\beta(0) = \beta_{\max}$ and $\beta(1) = \beta_{\min}$.

In this paper, we propose a different formulation of human–mosquito contact rate based on the idea of information-dependent epidemic models [11–14]. We show that the new model allows us to easily determine optimal strategies for implementing health campaigns. In particular, we show that the controlled system may predict dramatic reductions of malaria prevalence, even when the uncontrolled system predicts stable endemicity.

2 The Model

Agusto et al. [10] considered a malaria model where both the host and vector population are divided into two compartments, susceptibles and infectious individuals. The dynamics is ruled by the following system of nonlinear ordinary differential equations:

$$\begin{aligned}
\dot{S}_h &= \Lambda_h - \lambda_h(b)S_h - \mu S_h + \delta I_h \\
\dot{I}_h &= \lambda_h(b)S_h - (\alpha + \mu + \delta)I_h \\
\dot{S}_v &= \Lambda_v - \lambda_v(b)S_v - \eta(b)S_v \\
\dot{I}_v &= \lambda_v(b)S_v - \eta(b)I_v,
\end{aligned} \tag{2}$$

where the upper dot denotes the time derivative. The state variables are given by susceptible humans, S_h, infectious humans, I_h, susceptible vectors, S_v, and infectious vectors, I_v. The parameter $b \in [0, 1]$ is the proportion of ITN usage.

Table 1 Description of parameters in system (2) and baseline values (taken from [10])

Parameter	Description	Baseline value
Λ_h	Immigration rate in humans	$10^3/(70 \times 365)$
Λ_v	Immigration rate in mosquitoes	$10^4/21$
b	Proportion of ITN usage	Varies
μ	Natural mortality rate in humans	$1/(70 \times 365)$
η_{nat}	Natural mortality rate in mosquitoes	$1/21$
η_{bn}	Maximum ITN-induced death rate in mosquitoes	$1/21$
α	Disease-induced death rate in humans	10^{-3}
p_1	Prob. of disease transm. from mosquito to human	1
p_2	Prob. of disease transm. from human to mosquito	1
$\beta_{\max}$	Maximum transmission rate	0.1
$\beta_{\min}$	Minimum transmission rate	0
δ	Recovery rate of infectious humans to be susceptible	1/4

All the parameters in (2) are strictly positive constants and their meaning is described in Table 1. The *forces of infection* are given by

$$\lambda_h(b) = p_1\beta(b)\frac{I_v}{N_h}, \qquad \lambda_v(b) = p_2\beta(b)\frac{I_h}{N_h}, \tag{3}$$

where $\beta(b)$ represents the human–mosquito contact rate.

Using bed nets reduces the probability for humans to be bitten. Moreover, the nets are treated with insecticide. Therefore, in [10], it is assumed that ITN usage reduces the contact rate β according to (1) and increases the mosquito death rate η according to the relation

$$\eta(b) = \eta_{\text{nat}} + \eta_{bn}b. \tag{4}$$

Here, our aim is to incorporate human behavior in model (2) by employing the approach of information-dependent epidemic models [11–14]. The basic idea is to consider the feedback that the information about an infectious disease has on its spreading. Here, we model this feedback as the actions taken by individuals as consequence of a health-promotion campaign aimed at using ITNs.

The first step is to assume that the actions taken for the health-promotion campaign, such as advertising, counseling, hygienic aid, etc., summarized by the *effort* function $u(t)$, build up a *goodwill* $w(t)$, like the classical concept in marketing literature [15]. In this setting, w should be interpreted as concern or the willingness to use ITNs. As for the *information variable* employed in [11–14], we assume that w is not instantaneous but depends on the *past history* of the campaign in a way prescribed by a function ψ and distributed in the recent or far past by a delay kernel K_ξ^p. Therefore we set

$$w(t) = \int_{-\infty}^{t} \psi(I_h(\tau), u(\tau))\, K_\xi^p(t-\tau)d\tau. \tag{5}$$

In (5), the effort function $u(t)$ is assumed to be bounded $0 \leq u \leq u_{\max}$. As in [14] the kernel K_ξ^p is assumed to be an Erlangian kernel defined by the probability density function

$$K_\xi^p(x) = \frac{\xi^p x^{p-1} e^{-\xi x}}{(p-1)!}, \quad x, \xi \in \mathbf{R}_+, \quad p \in \mathbf{N}_+,$$

where $\xi > 0$, $p = 1, 2, \ldots$. In this case the delay is infinite and centered at p/ξ, which is the average delay [16]. The function ψ describes the role played by the state variables and health campaign measures in the goodwill dynamics and it may be generally assumed to be continuous and increasing with respect to I_h (actually ψ might be independent of I_h) and increasing respect to u. We also set $\psi(I_h, 0) = 0$ for all I_h.

As a particular case, if $p = 1$, we get $K_\xi^1(t) = \xi\, e^{-\xi t}$ that is an *exponentially fading memory*. In this case, from (5) by using the *linear chain trick* [16], we have $\dot{w} = \xi \psi\,(I_h(t), u(t)) - \xi w$. Note that here it must be

$$w(0) = \int_{-\infty}^{0} \psi\,(I_h(\tau), u(\tau))\, K_\xi^1(-\tau) d\tau.$$

By choosing $\psi(I_h, u) = u/\xi$ (which means that the "history" of the goodwill is affected only by campaign effort u) we get

$$\dot{w} = u(t) - \xi w. \tag{6}$$

Once that the goodwill has been introduced, it remains to specify how it affects the forces of infection [for the sake of simplicity, we neglect the mosquito killing effect of NTIs, i.e. we set $\eta_{bn} = 0$ in (4)]. As proposed in [15], we assume that the contact rate decays exponentially with w, i.e. we consider the forces of infection (3), where now

$$\beta(w) = \beta_{\max} - \frac{\beta_{\max} - \beta_{\min}}{1 - e^{-\gamma w_{\max}}} \left(1 - e^{-\gamma w}\right), \tag{7}$$

where γ is a positive constant. Observe that β in (7) ranges from $\beta_{\max}$ to $\beta_{\min}$ as w ranges from 0 to $w_{\max} = u_{\max}/\xi$.

3 The Optimal Control Problem

In this section, we will consider the model given by (2)–(4), (6), and (7) and look for optimal strategies for implementing health campaigns. In other words, we aim to determine the *optimal* effort $u(t)$ over a finite horizon t_f. "Optimal" in the sense that the campaign target is to minimize the total costs associated to both the disease and the controls. The costs associated to disease are assumed to be linearly dependent on the size of human infectious compartment, whereas the intervention costs are assumed to be quadratic (quadratic expressions of the control are the simplest and most widely used nonlinear representation of intervention costs. For more details see, e.g., [17–20]). The objective functional to be minimized is

$$J(u) = \int_0^{t_f} \left(AI_h + \frac{B}{2}u^2 \right) dt, \tag{8}$$

where the control $u(t)$, $i = 1, 2$, is a Lebesgue measurable functions such that $0 \leq u(t) \leq u_{\max}$, for $t \in [0, t_f]$. In (8), the (positive) constants A and B are *weight* parameters describing the comparative importance of the two terms in the functional [20]. This optimal control problem may be addressed by the well-known Pontryagin's maximum principle, where the Hamiltonian

$$H = g(\mathbf{x}, u, t) + \sum_{i=1}^{5} \lambda_i(t)\varphi_i(\mathbf{x}, \mathbf{u}, t),$$

must be minimized pointwise [20]. Here g is the integrand of the objective functional, $\mathbf{x}$ denotes the state-variable vector, λ_i, $i = 1, \ldots, 5$, are the adjoints, and φ_i denotes the right-hand side of the ith equation of system (2).

A similar optimal control approach, applied to malaria models or general host–vector models, can be also found in [21–24].

4 Numerical Results

We omit all the details concerning the method used to numerically solve the optimality system (the procedure is analogous to that in [18–20, 25] and many other papers). In our simulations, the parameter values are given in Table 1 (except that $\eta_{bn} = 0$). We assume that $u(t) = 0$ for $t < 0$ and $\gamma = 0.01$, $A = 1$, $B = 10$, $t_f = 250$ and the initial values $S_h(0) = 950$, $I_h(0) = 5$, $S_v(0) = 4,000$, $I_v(0) = 1,000$, $w(0) = 0$.

The optimal control profile $u(t)$ together with the corresponding goodwill $w(t)$ is shown in Fig. 1. It can be seen that the effort must be applied at its upper bound and this level must be maintained for almost all the total period, before dropping to zero. The goodwill of the ITN usage campaign affects the contact rate and, in turn, the disease transmission as shown in Fig. 2, where the state variables are plotted in both the cases of controlled and uncontrolled dynamics. In the shown simulation, the control is able to produce a dramatic reduction of malaria prevalence, even when the uncontrolled system predicts stable endemicity.

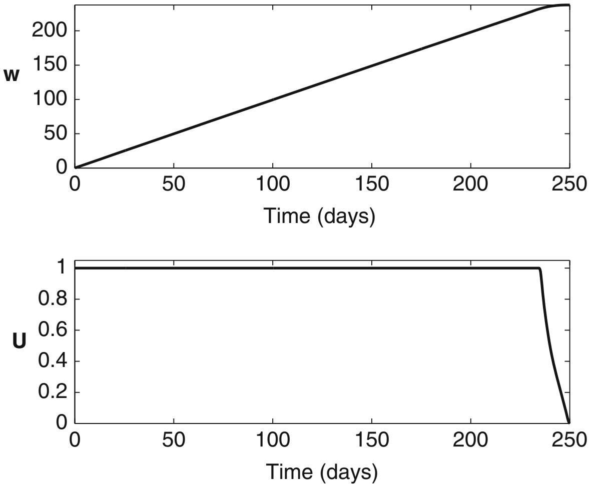


Fig. 1 The *goodwill* w and the optimal control u

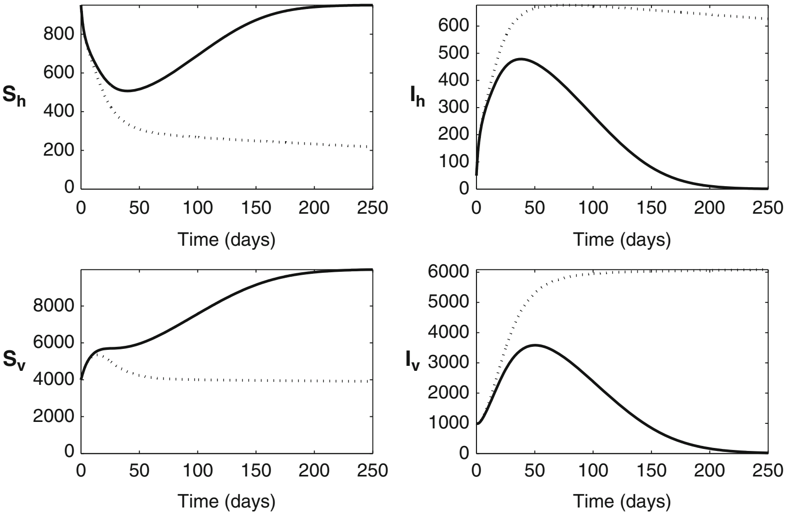


Fig. 2 *Dotted line*: uncontrolled dynamics. *Solid lines*: controlled dynamics

5 Conclusions

In this short note we used a mathematical modeling approach to investigate the effects of human behavior on malaria transmission. Motivated by the well-documented strong influence of behavioral factors in ITN usage, we propose a formulation of the human–mosquito contact rate that is based on the idea of information-dependent epidemic models. In particular, we assume that the goodwill of a health campaign for ITN usage depends on the past history and is distributed in the recent or far past by a delay kernel. As a particular example, we have considered the case where the goodwill is affected by campaign effort only and the memory is exponentially fading. We have shown that the new model allows to easily determine optimal strategies for implementing health campaigns.

We remark that an optimal control problem applied to model (2), together with (1), (3) and (4), has been recently considered by Silva and Torres [25]. In their model, the contact rate (1) is reduced by a coefficient $1 - \zeta(t)$, where the function $\zeta(t)$ is analogous to campaign effort $u(t)$ used in our model and must be chosen optimally in the same way (the costs (8) must be minimized). They show that the control is not able to radically change the dynamics of the system, although it may make faster the decay of infected humans when compared to the case where no controls are used. Instead, with our modeling approach, the controlled system may predict dramatic changes of malaria prevalence, as shown in Fig. 2.

Our study is, of course, a theoretical one. Real data, when available, could validate our findings. As far as future investigations are concerned, it might be of relevance to investigate the model in case of a general memory kernel, thus extending the present study.

Acknowledgements The author acknowledges the Italian National Group for the Mathematical Physics (GNFM) of National Institute for Advanced Mathematics (INdAM) for partially funding his participation at Gulf International Conference on Applied Mathematics, at the Gulf University for Science, Technology, Kuwait, held in November 2013. The anonymous referee is thanked for his/her useful comments, which helped to improve the quality of the paper.

References

1. World Health Organization: World Malaria Report 2012. Global Malaria Programme (2012)
2. Aron, J.L., May, R.M.: The population dynamics of malaria. In: Anderson, R.M. (ed.) The Population Dynamics of Infectious Disease: Theory and Applications, pp. 139–179. Champman and Hall, London (1982)
3. Koella, J.C.: On the use of mathematical models of malaria transmission. Acta Trop. **49**, 1–25 (1991)
4. Mandal, S., Sarkar, R.R., Sinha, S.: Mathematical models of malaria—a review. Malar. J. **10**, 202 (2011)
5. Nedelman, J.: Introductory review: Some new thoughts about some old malaria models. Math. Biosci. **73**, 159–182 (1985)
6. Lin, F., Muthuraman, K., Lawley, M.: An optimal control theory approach to non-pharmaceutical interventions. BMC Infect. Dis. **10**, 32–45 (2010)
7. Lengeler, C.: Insecticide-treated bed nets and curtains for preventing malaria. Cochane Database Syst. Rev. (2), Article No. CD000363 (2004)
8. Frey, C., Traoré, C., De Allegri, M., Kouyaté, B., Müller, O.: Compliance of young children with ITN protection in rural Burkina Faso. Malar. J. **5**, 70 (2006)
9. Manfredi, P., d'Onofrio, A. (eds.): Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases. Springer, New York (2013)
10. Agusto, F.B., Del Valle, S.Y., Blayneh, K.W., Ngonghala, C.N., Goncalves, M.J., Li, N., Zhao, R., Gong, H.: The impact of bed-net use on malaria prevalence. J. Theor. Biol. **320**, 58–65 (2013)

11. Buonomo, B., d'Onofrio, A., Lacitignola, D.: Global stability of an SIR epidemic model with information dependent vaccination. Math. Biosci. **216**, 9–16 (2008)
12. Buonomo, B., d'Onofrio, A., Lacitignola, D.: Globally stable endemicity for infectious diseases with information-related changes in contact patterns. Appl. Math. Lett. **25**, 1056–1060 (2012)
13. d'Onofrio, A., Manfredi, P.: Information-related changes in contact patterns may trigger oscillations in the endemic prevalence of infectious diseases. J. Theor. Biol. **256**, 473–478 (2009)
14. d'Onofrio, A., Manfredi, P., Salinelli, E.: Vaccinating behaviour, information, and the dynamics of SIR vaccine preventable diseases. Theor. Popul. Biol. **71**, 301–317 (2007)
15. Behncke, H.: Optimal control of deterministic epidemics. Optim. Control Appl. Meth. **21**, 269–285 (2000)
16. Smith, H.: An Introduction to Delay Differential Equations with Applications to the Life Sciences. Text in Applied Mathematics, vol. 57. Springer, New York (2011)
17. Anita, S., Arnautu, V., Capasso, V.: An Introduction to Optimal Control Problems in Life Sciences and Economics. Birkhäuser, Boston (2010)
18. Buonomo, B.: A simple analysis of vaccination strategies for rubella. Math. Biosci. Eng. **8**, 677–687 (2011)
19. Buonomo, B.: On the optimal vaccination strategies for horizontally and vertically transmitted infectious diseases. J. Biol. Syst. **19**, 263–279 (2011)
20. Lenhart, S., Workman, J.T.: Optimal Control Applied to Biological Models. Chapman and Hall/CRC Mathematical and Computational Biology Series. Chapman and Hall/CRC, Boca Raton (2007)
21. Agusto, F.B., Marcus, N., Okosun, K.O.: Application of optimal control to the epidemiology of malaria. Electron. J. Diff. Equat. **2012**, 1–22 (2012)
22. Kong, Q., Qiu, Z., Sang, Z., Zou, Y.: Optimal control of a vector-host epidemics model. Math. Control Rel. Fields **1**, 493–508 (2011)
23. Okosun, K.O., Ouifki, R., Marcus, N.: Optimal control analysis of a malaria disease transmission model that includes treatment and vaccination with waning immunity. Biosystems **106**, 136–145 (2011)
24. Ozair, M., Lashari, A.A., Jung, I.H., Okosun, K.O.: Stability analysis and optimal control of a vector-borne disease with nonlinear incidence. Discrete Dyn. Nat. Soc. **2012**, Article ID 595487 (2012)
25. Silva, C.J., Torres, D.F.M.: Conference Papers in Mathematics vol. 2013, Article ID 658468 (2013). arXiv:1306.2039v1

Computational Modelling and Optimal Control of HIV/AIDS Transmission in a Community with Substance Abuse Problem

I. Takaidza, O.D. Makinde, and K.O. Okosun

Abstract Abuse of substances continues to be ubiquitous in communities leading to high-risk sexual behaviour mainly due to impaired decision-making capacity. The abuse may also have numerous effects on neurocognitive function resulting in HIV infection and ultimately AIDS. In this paper, a compartmental deterministic model for the transmission dynamics of HIV/AIDS in a community plagued with substance abuse is proposed. The nonlinear problem is tackled using stability theory of differential equations and a basic reproduction number for the elimination of HIV infection is determined. The implementation of optimal control strategies involving treatment of substance-abusing susceptibles, counselling and prevention to combat the spread of HIV infection is determined using Pontryagin's maximum principle. Numerical simulations are performed and the pertinent results are presented graphically and discussed quantitatively.

Keywords HIV/AIDS model • Substance abuse • Reproduction number • Optimal control • Numerical simulation

1 Introduction

The use of any drug or combination of drugs to such an extent that drug effects seriously interfere with health or occupational and social functioning is considered abuse. Substance abuse is linked with poor adherence to taking ARV doses,

I. Takaidza (✉) • K.O. Okosun
Maths Department, Vaal University of Technology, P. Bag X021,
Vanderbijlpark 1900, Republic of South Africa
e-mail: isaact@vut.ac.za; kazeemo@vut.ac.za

O.D. Makinde
Stellenbosch University, P. Bag X2, Saldanha 7395, Republic of South Africa
e-mail: makinded@gmail.com

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_4,

which can lead to treatment failure. Mixing recreational drugs and ARVs can be dangerous as drug interactions can cause serious side effects or dangerous overdoses. In addition, drug use and abuse can facilitate the progress of HIV infection by further compromising the immune system. Treatment enables people to counteract addiction's powerful disruptive effects on the brain and behaviour and regain control of their lives [1]. HIV and substance abuse treatment and prevention services must thus be better integrated to take advantage of the multiple opportunities for intervention.

Plant [2] concluded that chronic heavy drinking or alcohol consumption levels consistent with alcohol dependence or alcohol-related liver disease do damage to the immune system. Hence, drinking appears to be a risk factor for potential exposure to HIV infection and for relapse into 'high-risk' sexual activities. McManus and Weatherburn [3] assessed the relationship between alcohol use and the likelihood of engagement in 'unsafe' sexual behaviour, the impact of alcohol on immune function and its importance as a co-factor for AIDS-related illness. A model for the spread of HIV/AIDS amongst a population of injecting drug users is developed and analysed in [4]. Logistic regression has been used to identify the independent influences of drug dependence symptoms or heavy drinking and HIV-related variables on comorbidity [5]. The substance abuse epidemic can be reduced by intervention programmes targeted at light drug users and by increasing the uptake rate into treatment for those addicted [6]. A deterministic model for the case where there is no interaction between misusers and non-misusers in the susceptible and infectious classes allowing only for transition of drug misusers with AIDS to non-misusers with AIDS is considered in [7].

The paper is organized as follows. In Sect. 2, we present a compartmental deterministic model consisting of ordinary differential equations describing the transmission dynamics of the disease given the underlying assumptions and also provide the basic properties of the model. Section 3 is devoted to the optimal control of the disease, making use of Pontryagin's maximum principle. In Sect. 4, we present and discuss the numerical simulation results. Cost-effective analysis is the subject of Sect. 5.

2 Mathematical Model

The population, $N(t)$, is divided as follows: susceptible individuals who do not abuse substances, $S_1(t)$, the susceptibles who abuse substances, $S_2(t)$, individuals who do not abuse substances and are infected by HIV, $I_1(t)$, infected individuals who also abuse substances, $I_2(t)$, and the full-blown AIDS group, $A(t)$. We assume that the susceptibles are recruited into the community through birth or immigration at a rate Q_0. A proportion P of the immigrants abuse substances. The non-abusing susceptibles become abusers at a rate θ_1 and the non-abusing infected become abusers at a rate θ_2 where $\theta_2 > \theta_1$ due to HIV infection. We denote by β the contact rate between the susceptible and the HIV infected while $\sigma > 1$ denotes the rate of

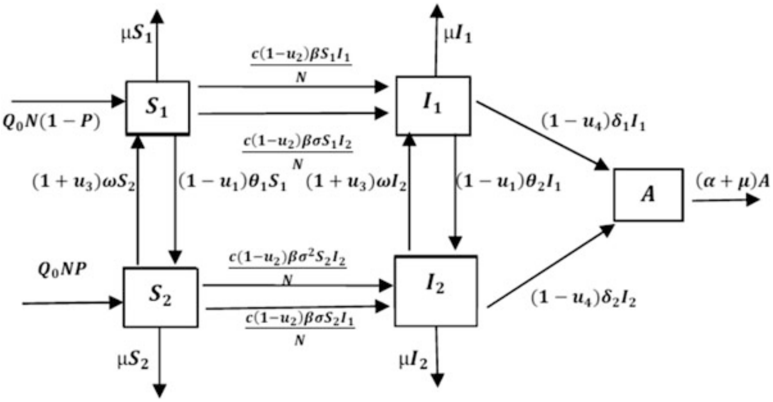


Fig. 1 Model flow diagram

increase in contact between the susceptible and the HIV infected due to substance abuse. c is the number of sexual partners an infected person is having.

The control variable for reducing the recruitment to substance abuse is denoted by u_1 while the control variable for reducing the spread of HIV infection is denoted by u_2. The non-abusing and abusing infected move to the AIDS class at rates δ_1 and δ_2, respectively, $\delta_2 > \delta_1$ due to substance abuse. The recovery rate for substance abuse is denoted by ω, while u_3 is the control on treatment for substance abuse. u_4 is the control on treatment of the infected. Note that $0 \leq u_i \leq 1, i = 1, 2, 3, 4$. μ and α represent the natural and disease-induced mortality rates, respectively. The dynamics are depicted in Fig. 1.

Following the above discussion, the resulting *state system* is given by

$$
\begin{aligned}
\frac{dS_1}{dt} &= (1-P)Q_0N - ((1-u_1)\theta_1 + \mu)S_1 - \frac{c(1-u_2)\beta S_1(I_1+\sigma I_2)}{N} \\
&\quad + (1+u_3)\omega S_2 \\
\frac{dS_2}{dt} &= PQ_0N + (1-u_1)\theta_1 S_1 - \frac{c(1-u_2)\beta\sigma S_2(I_1+\sigma I_2)}{N} \\
&\quad - ((1+u_3)\omega + \mu)S_2 \\
\frac{dI_1}{dt} &= -(1-u_1)\theta_2 I_1 + \frac{c(1-u_2)\beta S_1(I_1+\sigma I_2)}{N} + (1+u_3)\omega I_2 \\
&\quad -((1-u_4)\delta_1 + \mu)I_1 \\
\frac{dI_2}{dt} &= (1-u_1)\theta_2 I_1 + \frac{c(1-u_2)\beta\sigma S_2(I_1+\sigma I_2)}{N} \\
&\quad - ((1+u_3)\omega + (1-u_4)\delta_2 + \mu)I_2 \\
\frac{dA}{dt} &= (1-u_4)\delta_1 I_1 + (1-u_4)\delta_2 I_2 - (\alpha+\mu)A
\end{aligned}
\tag{1}
$$

with the initial conditions given by $S_1(0) = S_{01}, S_2(0) = S_{02}, I_1(0) = I_{01}, I_2(0) = I_{02}$ and $A(0) = A_0$. The model is epidemiologically meaningful since all solutions with non-negative initial data will remain non-negative for all time.

Theorem. *If $S_{01}, S_{02}, I_{01}, I_{02}$ and A_0 are non-negative, then so are $S_1(t), S_2(t), I_1(t), I_2(t)$ and $A(t)$ for all time $t > 0$. Moreover,*

$$\limsup_{t\to\infty} N(t) \leq \frac{Q_0}{\mu}. \tag{2}$$

Furthermore, if $N(0) \leq \frac{Q_0}{\mu}$, then $N(t) \leq \frac{Q_0}{\mu}$.

Local existence of solutions follows from standard arguments since the right-hand sides of the model system (1) are locally Lipschitz. Global existence follows from the a priori bounds.

The disease-free equilibrium $\xi_0 = (S_1^0, S_2^0, I_1^0, I_2^0, A^0)$ entails $Q_0 = \mu$ and is given by

$$\xi_0 = \left(\frac{N[Q_0(1-P) + (1+u_3)\omega]}{(1-u_1)\theta_1 + \mu + (1+u_3)\omega}, \frac{N[(1-u_1)\theta_1 + Q_0 P]}{(1-u_1)\theta_1 + \mu + (1+u_3)\omega}, 0, 0, 0 \right). \tag{3}$$

The linear stability of ξ_0 is governed by the basic reproduction number R_0 which is defined as the expected number of secondary infections produced by a single infectious individual during his/her entire infectious period. To compute the basic reproduction number we only consider the states that apply to infected individuals and thus focus on those equations of (1) that describe the production of new infections and changes in state amongst infected individuals. The matrix describing this infected subsystem is decomposed as $\mathbf{F} - \mathbf{V}$, where

$$\mathbf{F} = \begin{bmatrix} \frac{c(1-u_2)\beta S_1}{N} & \frac{c(1-u_2)\beta\sigma S_1}{N} & 0 \\ \frac{c(1-u_2)\beta\sigma S_2}{N} & \frac{c(1-u_2)\beta\sigma^2 S_2}{N} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4}$$

is the transmission part, describing the production of new infections, and

$$\mathbf{V} = \begin{bmatrix} \mu + (1-u_1)\theta_1 + (1-\rho u_4)\delta_1 & -u_3\omega & 0 \\ -(1-u_1)\theta_2 & u_3\omega + (1-\rho u_4)\delta_2 + \mu & 0 \\ -(1-\rho u_4)\delta_1 & -(1-\rho u_4)\delta_2 & \alpha + \mu \end{bmatrix} \tag{5}$$

is the transition part, describing changes in state (including removal by death). The control reproduction number, R_c, in the presence of drug-abusing individuals is then computed as the dominant eigenvalue, or more precisely the spectral radius, of the next-generation matrix $\mathbf{K} = \mathbf{F} \times \mathbf{V}^{-1}$ about the infection-free steady state [8], that is, the disease-free equilibrium and is given by

$$R_c = \frac{c\beta\sigma(1-u_2)[(1-u_1)\theta_1 + Q_0P][((1-u_1)\theta_1 + (1-u_4)\delta_1 + \mu)\sigma + (1+u_3)\omega]}{[(1-u_4)\delta_1 + \mu + (1-u_1)(\theta_2+\theta_1)][(1-u_4)\delta_2 + \mu + (1+u_3)\omega][(1-u_1)\theta_1 + \mu + (1+u_3)\omega]}$$

with the basic reproduction number given by

$$R_0 = \frac{c\beta\sigma[\theta_1 + Q_0P][(\theta_1+\delta_1+\mu)\sigma+\omega]}{[\delta_1+\mu+\theta_2+\theta_1][\delta_2+\mu+\omega][\theta_1+\mu+\omega]}. \tag{6}$$

Analysis of the reproduction number shows that the additional pathways of disease transmission for substance abusers increase the likelihood of disease spread.

3 Optimal Control

To investigate the optimal level of efforts that would be needed to control the disease, we wish to minimize the number of substance-abusing and infectious individuals and the cost of applying the controls u_1, u_2, u_3 and u_4 over a finite time interval $[0, T]$. We achieve this by defining an objective functional, J, by choosing a quadratic cost on the controls

$$J = \int_0^T (mS_2 + nI_1 + kI_2 + b_1u_1^2 + b_2u_2^2 + b_3u_3^2 + b_4u_4^2)dt, \tag{7}$$

where m, n, k, b_1, b_2, b_3 and b_4 are positive weights. We seek optimal controls such that the objective functional is minimized. The necessary conditions that $\mathbf{u}^* = (u_1^*, u_2^*, u_3^*, u_4^*)$ and $\mathbf{x}^* = (S_1^*, S_2^*, I_1^*, I_2^*, A^*)$ must satisfy come from Pontryagin's maximum principle [9]. We use this principle to convert the problem of minimization of the objective functional coupled with the state variables into a problem of minimizing point-wise a Hamiltonian, H, with respect to the controls u_1, u_2, u_3 and u_4.

$$H = mS_2 + nI_1 + kI_2 + \sum_{i=1}^{4} b_iu_i^2 + \lambda_{S_1}\frac{dS_1}{dt} + \lambda_{S_2}\frac{dS_2}{dt} + \lambda_{I_1}\frac{dI_1}{dt} + \lambda_{I_2}\frac{dI_2}{dt} + \lambda_A\frac{dA}{dt}, \tag{8}$$

where $\lambda_{S_1}, \lambda_{S_2}, \lambda_{I_1}, \lambda_{I_2}$ and λ_A are adjoint or co-state variables. By applying Pontryagin's maximum principle and the existence result for the optimal control [10], we obtain the *adjoint system*

$$\begin{aligned}\frac{d\lambda_{S_1}}{dt} &= (\mu - Q_0)\lambda_{S_1} + (PQ_0 + (1-u_1)\theta_1)(\lambda_{S_1} - \lambda_{S_2}) \\ &\quad + \frac{\varphi_I}{N^2}\{\sigma S_2(\lambda_{I_2} - \lambda_{S_2}) + (N - S_1)(\lambda_{S_1} - \lambda_{I_1})\}\end{aligned}$$

$$\begin{aligned}
\frac{d\lambda_{S_2}}{dt} &= -m + Q_0P_\lambda + \mu\lambda_{S_2} + (1+u_3)\omega(\lambda_{S_2} - \lambda_{S_1}) \\
&\quad + \frac{\varphi_I}{N^2}\{S_1(\lambda_{I_1} - \lambda_{S_1}) + \sigma(N - S_2)(\lambda_{S_2} - \lambda_{I_2})\} \\
\frac{d\lambda_{I_1}}{dt} &= -n + Q_0P_\lambda + (1-u_1)\theta_2(\lambda_{I_1} - \lambda_{I_2}) + (1-u_4)\delta_1(\lambda_{I_1} - \lambda_A) \\
&\quad +\mu\lambda_{I_1} + \frac{c\beta(1-u_2)(N - (I_1 + \sigma I_2))}{N^2}\psi_{SI} \\
\frac{d\lambda_{I_2}}{dt} &= -k + Q_0P_\lambda + (1+u_3)\omega(\lambda_{I_2} - \lambda_{I_1}) + (1-u_4)\delta_2(\lambda_{I_2} - \lambda_A) \\
&\quad +\mu\lambda_{I_2} + \frac{c\beta(1-u_2)(\sigma N - (I_1 + \sigma I_2))}{N^2}\psi_{SI} \\
\frac{d\lambda_A}{dt} &= Q_0P_\lambda + \lambda_A(\alpha + \mu) - \frac{\varphi_I}{N^2}\psi_{SI},
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
P_\lambda &= P(\lambda_{S_1} - \lambda_{S_2}) - \lambda_{S_1}, \\
\varphi_I &= c\beta(1-u_2)(I_1 + \sigma I_2), \\
\psi_{SI} &= S_1(\lambda_{S_1} - \lambda_{I_1}) + \sigma S_2(\lambda_{S_2} - \lambda_{I_2}).
\end{aligned} \tag{10}$$

The adjoint system has *final values*

$$\lambda_{S_1}(T) = \lambda_{S_2}(T) = \lambda_{I_1}(T) = \lambda_{I_2}(T) = \lambda_A(T) = 0. \tag{11}$$

The values of the optimal control variables at each instant are found by noting that each minimizes the Hamiltonian and thus must satisfy the necessary condition $\dfrac{\partial H}{\partial u_i} = 0$. Coupled with standard control arguments involving the bounds on the controls yields the following expressions for the *optimal controls*:

$$\begin{aligned}
u_1^* &= \min\left\{1, \max\left\{0, \frac{\theta_1 S_1(\lambda_{S_2} - \lambda_{S_1}) + \theta_2 I_1(\lambda_{I_2} - \lambda_{I_1})}{2b_1}\right\}\right\}, \\
u_2^* &= \min\left\{1, \max\left\{0, \frac{c\beta(I_1 + \sigma I_2)[S_1(\lambda_{I_1} - \lambda_{S_1}) + \sigma S_2(\lambda_{I_2} - \lambda_{S_2})]}{2Nb_2}\right\}\right\}, \\
u_3^* &= \min\left\{1, \max\left\{0, \frac{\omega[S_2(\lambda_{S_2} - \lambda_{S_1}) + I_2(\lambda_{I_2} - \lambda_{I_1}]}{2b_3}\right\}\right\}, \\
u_4^* &= \min\left\{1, \max\left\{0, \frac{\delta_1 I_1(\lambda_A - \lambda_{I_1}) + \delta_2 I_2(\lambda_A - \lambda_{I_2})}{2b_4}\right\}\right\}.
\end{aligned} \tag{12}$$

4 Numerical Simulation and Discussion

An iterative scheme is used to solve the optimality system which consists of state and adjoint equations. We start to solve the state equations with a guess for the controls over the simulated time using the fourth-order Runge-Kutta scheme. The state equations (1) are solved using a forward method with given initial conditions, whereas the adjoint system (9) is solved using a backward scheme with the prescribed final conditions. Controls are updated by using a convex combination of the previous controls and the stationary value characterizations (12). This process is repeated and iterations stopped if the values of the unknowns at the previous iterations are very close to the ones at the present iterations [11].

We investigate and compare numerical results for the following combinations with at least three controls to ensure that we address either prevention or treatment of both substance abuse and HIV infection: (i) strategy A, when the substance abuse prevention control, u_1, is set to zero and the other controls optimized; (ii) strategy B, when the HIV prevention control, u_2, is set to zero and the other controls optimized; (iii) strategy C, when the substance abuse treatment control, u_3, is set to zero and the other controls optimized; (iv) strategy D, when the HIV treatment control, u_4, is set to zero and the other controls optimized; and (v) strategy E, when all controls are optimized.

For the simulations, we choose the model parameter values in Table 1. Q_0, P, ω and α are obtained from [7] while δ_1 and μ are as in [12] with the rest assumed.

We assume that the weight factor, b_1, associated with control u_1 is lower than b_2, b_3 and b_4 which are associated with controls u_2, u_3 and u_4, respectively. This assumption is based on the fact that it probably costs more to control the spread of HIV than substance abuse. The cost associated with treatment will include the cost of medical examinations, drugs and hospitalization with the treatment of HIV being lifelong, thus making it the costliest. We use the following objective functional parameter values (Table 2):

For illustrative purposes we make use of the initial state conditions $S_1(0) = 20$, $S_2(0) = 15$, $I_1(0) = 10$, $I_2(0) = 5$ and $A(0) = 0$.

Numerical simulations are consistent for all the scenarios under consideration, varying only in the margins of growth and reduction. We, consequently, only present and discuss results for the most cost-effective combination, which is no prevention to HIV infection. In Fig. 2a, the number of the substance-abusing individuals is lower under control as contrasted to without control. In fact, the abusing susceptible

Table 1 Model parameters

Parameter	Q_0	P	β	c	σ	θ_1	θ_2	δ_1	δ_2	ω	μ	α
Value	0.029	0.5	0.0753	1	1.02	0.15	0.18	0.1	0.12	0.05	0.02	0.4

Table 2 Objective functional parameters

Parameter	m	n	k	b_1	b_2	b_3	b_4
Value	12	20	25	20	60	140	170

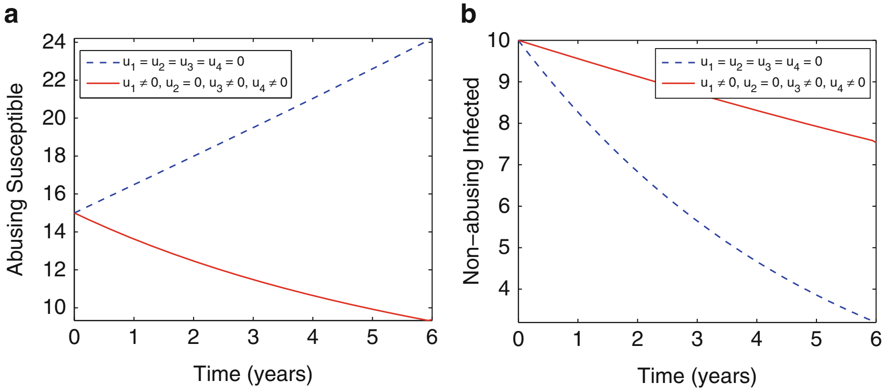


Fig. 2 (**a**) Abusing susceptible and (**b**) non-abusing infected

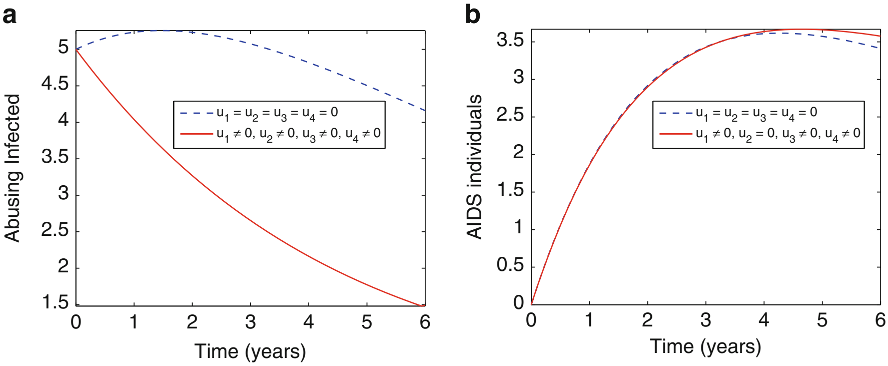


Fig. 3 (**a**) Abusing infected and (**b**) AIDS individuals

number reduces under control while there is increase in time without control. In Fig. 2b, the number of the non-abusers is more under control despite the lack of growth for the infected, which is reasonable due to the positive impact of prevention and treatment for substance abuse.

Figure 3a shows that there is slight growth for the first two years followed by a reduction in the number of the abusing infected in the absence of control but reduction from the onset under control. Figure 4a indicates that maximum efforts need to be employed to prevent and treat substance abuse. Shadow prices in Fig. 4b show that the infected abusers cost communities the worst followed respectively by the infected non-abusers and the susceptible abusers.

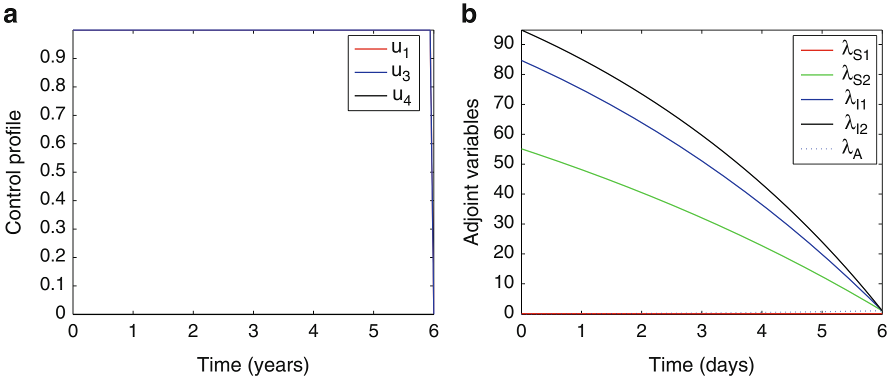


Fig. 4 (**a**) Control profiles and (**b**) adjoints

5 Cost-Effectiveness

Realizing positive impact for a population is an important goal of public health programmes and policies. Impact is measured using indicators related to a change in health status such as the estimated number of deaths and infections averted. Cost-effective analysis is one of several economic evaluation tools used to measure the costs and consequences of alternative programmes. The measures are then compared to assess how the greatest health benefits can be generated. To identify the strategy which realizes the most positive impact, we make use of incremental cost-effectiveness ratios (ICERs) defined as

$$\text{ICER} = \frac{\text{Difference in costs between strategies}}{\text{Difference in health effects between strategies}}.$$

We consider health effects as the cases averted in the S_2, I_1 and I_2 classes. Strategies are ranked from the least effective by considering health effects and then compared pairwise using ICERs.

Strategy	A	B	C	D	E
Cases averted	255	596	720	748	748
Costs	71,437	64,361	26,336	67,388	67,325

The ICER between A and C is -17.82. So it costs 17.82 less for each additional case averted from A to C, so A is excluded. Next, we calculate the ICER between C and B, which is -306.65. Hence, it costs 306.65 less for each additional case averted as we switch from C to B. So, we exclude C and calculate ICER for B and D. It costs 1,466.14 more for each additional case averted when switching from B to D.

D is now excluded and we calculate ICER between B and E. It costs 1,463.89 more for each additional case averted from B to E. Therefore, strategy B is the most cost effective.

6 Conclusion

Shadow prices show that the cost and impact of the infected substance abusers is very high; this may result in negative effects on the population. The results suggest that prevention and treatment of substance abuse coupled with treatment of the infected is the most effective strategy. However, budgetary provision still needs to be made to include the prevention of infection so as to reduce the risk of HIV transmission. Control programmes that follow these strategies can effectively reduce the spread of HIV attributable to substance abuse.

Acknowledgements The authors acknowledge the Vaal University of Technology Research Office and the National Research Foundation (NRF), South Africa, through the KIC Grant ID 87225 for the financial support to attend and present this paper at the Gulf International Conference on Applied Mathematics (GICAM13).

References

1. NIDA: Treatment and recovery. J. Neurosci. **21–23**, 9414–9418 (2001)
2. Plant, M.A.: Alcohol, sex and AIDS. Alcohol Alcohol. **25**(2/3), 293–301 (1990)
3. McManus, T.J., Weatherburn, P., Alcohol, AIDS and immunity. Br. Med. Bull. **50**(1), 115–123 (1994)
4. Greenhalgh, D., Hay, G.: Mathematical modelling of the spread of HIV/AIDS amongst injecting drug users. IMA J. Math. Appl. Med. Biol. **14**(1), 11–38 (1997)
5. Galvan, F.H., Burnam, M.A., Bing, E.G.: Co-occurring psychiatric symptoms and drug dependence or heavy drinking among HIV-positive people. J. Psychoac. Drugs **35**(Suppl 1), 153–160 (2003)
6. Kalula, A.S., Nyabadza, F.: A theoretical model for substance abuse in the presence of treatment. South African J. Sci.. **108**(3/4), 12 (2012). doi:10.4102/sajs.v108i3/4.654
7. Bhunu, C.P., Tchuenche, J.M., Lutscher, F., Mushayabasa, S., Bauch, C.T.: Assessing the effects of drug use on the transmission dynamics of HIV/AIDS. In: Mushayabasa, S., Bhunu, C.P. (eds.) Understanding the Dynamics of Emerging and Re-Emerging Infectious Diseases Using Mathematical Models, pp. 105–131. (2012). ISBN:978-81-7895-549-0
8. van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. **180**, 29–48 (2002)
9. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: The Mathematical Theory of Optimal Processes. Wiley, New York (1962)
10. Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1975)
11. Lenhart, S., Workman, J.T.: Optimal Control Applied to Biological Models. Chapman and Hall, New York (2007)
12. Okosun, K.O., Makinde, O.D., Takaidza, I.: Impact of optimal control on the treatment of HIV/AIDS and screening of unaware infective. Appl. Math. Model. **37**, 3802–3820 (2013). doi:10.1016/j.apm.2012.08.004

Standard Difference Scheme for a Singularly Perturbed Convection-Diffusion Equation in the Presence of Perturbations

G. Shishkin, L. Shishkina, and A. Petrenko

Abstract We consider a Dirichlet problem for a singularly perturbed ordinary differential convection-diffusion equation with a perturbation parameter ε ($\varepsilon \in (0, 1]$) multiplying the highest-order derivative in the equation. This problem is approximated by the standard monotone finite difference scheme on a uniform grid. Such a scheme does not converge ε-uniformly. Moreover, under its convergence, it is not ε-uniformly well conditioned and stable to perturbations in the data of the discrete problem and/or computer perturbations. For a model boundary value problem in the case of computer perturbations, we discuss results of numerical experiments and their conformity to theoretical results.

Keywords Singularly perturbed boundary value problem • Convection-diffusion equation • Standard difference scheme • Uniform grid • Uniform norm • Conditioning of difference scheme • Perturbed difference scheme • Computer perturbations

1 Introduction

Classical difference schemes on uniform grids (standard difference schemes) are widely used for solving applied problems [1]. Quite often such schemes are applied to solve singularly perturbed equations. However, for the convection-diffusion

G. Shishkin (✉) • L. Shishkina
Institute of Mathematics and Mechanics, Russian Academy of Sciences, Ekaterinburg, Russia
e-mail: shishkin@imm.uran.ru

A. Petrenko
Helsinki Metropolia University of Applied Sciences, PO BOX 4000 (Bulevardi 31), FI-00079, Metropolia, Finland
e-mail: mr.alexey.petrenko@gmail.com

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_5,

problems, standard schemes do not converge ε-uniformly, and in the case of their convergence they are not ε-uniformly well conditioned and not stable to perturbations in the data; see, e.g., [2–7]. Applicability of standard schemes for solving singularly perturbed problems requires further study.

In this paper, we consider a Dirichlet problem for a singularly perturbed ordinary differential convection-diffusion equation which is approximated by the standard monotone finite difference scheme on a uniform grid. For this problem, we develop a technique for theoretical and numerical studies of grid solutions in the presence of perturbations in the data of the discrete problem, as well as computer perturbations. We present and discuss the results of numerical experiments that illustrate theoretical results.

The contents of the paper are the following. Formulation of the boundary value problem for a singularly perturbed convection-diffusion equation and the aim of the study are presented in Sect. 2; here also, a standard scheme on a uniform grid is constructed. A difference scheme under perturbations in its data and also conditioning of the scheme are considered in Sect. 3. Standard scheme in the case of computer perturbations is studied in Sect. 4. In Sect. 5, numerical studying a model boundary value problem is performed; numerical results are compared with theoretical results.

Note that some results on the investigation of the standard scheme on a uniform grid, in particular, in the presence of perturbations in the data of grid problem from Sects. 2 and 3 are partial results of the papers [3, 5], adapted to the present study. Technique of numerical studying errors in the presence of perturbations in the data of the grid problem, as well as computer perturbations, that is described in Sect. 5, earlier has not been considered.

2 Problem Formulation, Standard Difference Scheme; Aim of Research

2.1 Problem Formulation

On the set $\overline{D} = D \cup \Gamma$, $D = (0, 1)$, we consider the Dirichlet problem for the singularly perturbed ordinary differential convection-diffusion equation[1]

$$L_{(1)}u(x) \equiv \left\{\varepsilon a(x)\frac{d^2}{dx^2} + b(x)\frac{d}{dx} - c(x)\right\} u(x) = f(x), \quad x \in D, \tag{1}$$

$$u(x) = \varphi(x), \quad x \in \Gamma.$$

[1] The notation $L_{(j.k)}$ ($M_{(j.k)}$, $G_{h(j.k)}$) means that these operators (constants, grids) are introduced in formula $(j.k)$.

Here $\Gamma = \Gamma_1 \cup \Gamma_2$, where Γ_1 and Γ_2 are the left and right parts of the boundary Γ; the functions $a(x)$, $b(x)$, $c(x)$, $f(x)$ are assumed to be sufficiently smooth on $\overline{D}$; moreover[2]

$$m \le a(x),\, b(x),\, c(x) \le M, \quad |f(x)| \le M, \quad x \in \overline{D}, \quad |\varphi(x)| \le M, \quad x \in \Gamma,$$

the parameter ε takes arbitrary values in $(0, 1]$. For small values of the parameter ε, a boundary layer appears in a neighborhood of the set Γ_1.

2.2 Standard Difference Scheme

We consider a standard difference scheme on the uniform grid $\overline{D}_h = \overline{D}^u_h$ with the step-size $h = 1/N$, where $N + 1$ is the number of nodes $x = x^i$ in the grid $\overline{D}^u_h$, $i = 0, 1, \ldots, N$.

Problem (1) is approximated by the difference scheme [1]

$$\Lambda z(x) \equiv \{\varepsilon\, a(x)\, \delta_{\overline{x}\hat{x}} + b(x)\, \delta_x - c(x)\} z(x) = f(x), \quad x \in D_h,$$
$$z(x) = \varphi(x), \quad x \in \Gamma_h; \tag{2}$$

here $D_h = D \cap \overline{D}_h$, $\Gamma_h = \Gamma \cap \overline{D}_h$, $\delta_{\overline{x}\hat{x}}\, z(x)$ is the central second-order difference derivative, and $\delta_x\, z(x)$ and $\delta_{\overline{x}}\, z(x)$ are the first-order difference (forward and backward) derivatives.

Difference scheme (2) is monotone ε-uniformly [1]. Using the maximum principle, for $z(x) - u(x)$, i.e., *the error of the solution to difference scheme* (2) (or, in short, *the error of the grid solution*), we obtain the following estimate (similar to estimate (3.3) in [3]):

$$\|u - z\|_{\overline{D}_h} \le M\, \delta_{st}; \quad \delta_{st} = \delta_{st}(\varepsilon, N) = \left(\varepsilon + N^{-1}\right)^{-1} N^{-1}. \tag{3a}$$

For the standard scheme, the value δ_{st} is determined by the product εN: $\delta_{st} = \delta_{st\,1}(\varepsilon N) = (\varepsilon N + 1)^{-1}$. Note that the estimate (3a) is equivalent to the following one:

$$\|u - z\|_{\overline{D}_h} \le M\, \delta; \quad \delta = \delta(\varepsilon, N) = \varepsilon^{-1} N^{-1}. \tag{3b}$$

We say that the value $\delta_{st} = \delta_{st\,(3a)}(\varepsilon, N)$ as well as $\delta = \delta_{(3b)}(\varepsilon, N)$ are *accuracy parameters of difference scheme* (2) (or, in short, *accuracy of the difference scheme*), and $M_{(3a)}$ and $M_{(3b)}$ are the *error constants*.

[2] By M (or m), we denote sufficiently large (small) positive constants independent of the parameter ε and of the discretization parameters.

The estimate (3) holds in the case of the following a priori estimate (which follows from a priori estimate (4.4) of [3]):

$$|d^k/dx^k\, u(x)| \le M\,(1+\varepsilon^{1-k}+\varepsilon^{-k}\ \exp^{-m\,\varepsilon^{-1}x}), \quad x \in \overline{D},\ k \le K,\ K = 3. \quad (4)$$

Thus, the following theorem on convergence of standard difference scheme (2) holds (similar to Theorem 1 from [3]).

Theorem 1. *Let the solution $u(x)$ of the problem* (1) *satisfy the estimate* (4)*. Then the solution of the standard finite difference scheme* (2) *converges to $u(x)$ with the estimate* (3)*.*

Remark 1. Standard scheme (2) does not converge ε-uniformly; for its convergence it is required to use grids with the number of nodes $N + 1$, growing indefinitely as $\varepsilon \to 0$ and $N = N(\varepsilon) \gg \varepsilon^{-1}$.

2.3 Aim of Research

Our aim for boundary value problem (1) is to consider conditioning of standard scheme on a uniform grid and convergence of its solution both under perturbation in the data of the grid problem and under computer perturbations; also it is required to compare theoretical results and results of numerical experiments.

3 Estimates of Grid Solution Under Data Perturbation; Conditioning of Difference Scheme (2)

In the case of difference scheme (2), we consider perturbations of solutions caused by perturbations in the data, as well as conditioning of the difference scheme.

3.1 Matrix Form of the Difference Scheme

Let the components of the function $z(x)$, $x \in \overline{D}_h$, be associated with an $(N + 1)$-dimensional vector Y. Ordering the elements $z(x)$ in scheme (2), we come to the system

$$A\,Y = F. \qquad (5)$$

Here A is a three-diagonal $(N+1) \times (N+1)$-matrix (a_{ij}); Y and F are vectors from the space $\mathbb{R}^{N+1}$ with the uniform vector norm $\|\cdot\|$. The components of the matrix A and vectors Y and F are determined by the relations

$$a_{i,i-1} = -\varepsilon\, h^{-2}\, a(x_i), \quad a_{i\,i} = 2\,\varepsilon\, h^{-2}\, a(x_i) + h^{-1}\, b(x_i) + c(x_i),$$

$$a_{i,i+1} = -\varepsilon\, h^{-2}\, a(x_i) - h^{-1}\, b(x_i), \quad 2 \le i \le N; \quad Y_i = z(x_i), \quad 1 \le i \le N+1;$$

$$F_1 = \varphi(x_1), \quad F_i = -f(x_i), \quad 2 \le i \le N, \quad F_{N+1} = \varphi(x_{N+1});$$

here $x_{i(5)} = x^{i+1}$, $x^i \in \overline{D}_h$. The matrix A is an M-matrix.

3.2 Perturbed Standard Difference Scheme

We consider the following perturbed problem corresponding to (5):

$$A^*\, Y^* = F^*. \tag{6}$$

Here A^* is the perturbed matrix (a^*_{ij}), Y^* and F^* are perturbed vectors, $A^* = A + \delta A$, $Y^* = Y + \delta Y$, $F^* = F + \delta F$. The perturbations of the coefficient $a(x_i)$ entering the components a_{ij}, $j = i-1, i, i+1$, $i = 2, \dots, N$ of the matrix A are, in general, different; we denote these perturbations in the components a_{ij} by δa_i^j. In a similar way, we denote the perturbations of the coefficient $b(x_i)$ in the components b_{ij}, $j = i, i+1$ and the perturbations of the coefficient $c(x_i)$ in the component $c_{i\,i}$ by δb_i^j and δc_i^i, respectively. Assume that the components equal to zero or one and also the values ε and h are not perturbed. Thus, in the componentwise notation of the matrix δA and the vectors δF and δY, we have

$$\delta a_{i,i-1} = -\varepsilon\, h^{-2}\, \delta a_i^{i-1}, \quad \delta a_{i\,i} = 2\,\varepsilon\, h^{-2}\, \delta a_i^i + h^{-1}\, \delta b_i^i + \delta c_i^i, \tag{7}$$

$$\delta a_{i,i+1} = -\varepsilon\, h^{-2}\, \delta a_i^{i+1} - h^{-1}\, \delta b_i^{i+1}, \quad 2 \le i \le N;$$

$$\delta F_1 = \delta\varphi(x_1), \;\; \delta F_i = -\,\delta f(x_i), \;\; 2 \le i \le N, \quad \delta F_{N+1} = \delta\varphi(x_{N+1}); \;\; \delta Y_i = \delta z(x_i).$$

In the presence of *perturbations in the data* of the grid problem, for the perturbed matrix problem (6) we have the following *perturbed standard difference scheme* (or, in short, *perturbed difference scheme*):

$$\begin{aligned} \Lambda^*\, z^*(x) &\equiv \{\varepsilon\, a^*(x)\, \delta_{\overline{x}\hat{x}} + b^*(x)\, \delta_x - c^*(x)\}\, z^*(x) = f^*(x), && x \in D_h, \\ z^*(x) &= \varphi^*(x), && x \in \Gamma_h. \end{aligned} \tag{8}$$

Here $x = x^i$, $x^i \in \overline{D}_h$, and in the relations below we have $x_i = x^{i-1}$, $x^i \in \overline{D}_h$, and

$$a^*(x_i) = a(x_i) + \delta a_i^{i-1}, \quad b^*(x_i) = b(x_i) + \delta b_i^{i+1} + \varepsilon\, h^{-1}\,(-\delta a_i^{i-1} + \delta a_i^{i+1}),$$

$$c^*(x_i) = c(x_i) + \delta c_i^i - \varepsilon\, h^{-2}\,(\delta a_i^{i-1} - 2\,\delta a_i^i + \delta a_i^{i+1}) + h^{-1}\,(\delta b_i^i - \delta b_i^{i+1}),$$

$$f^*(x_i) = f(x_i) + \delta f_i, \quad \varphi^*(x_i) = \varphi(x_i) + \delta\varphi_i;$$

$z^*(x)$ is the *perturbed grid solution*, i.e., the solution of the perturbed difference scheme (8).

Taking into account (7), for the value $z^*(x) - z(x)$, i.e., for the *perturbation of the grid solution*, we obtain the estimate

$$\|z^* - z\|_{\overline{D}_h} \le M\,\big[\,\varepsilon\, N^2\, |\delta a_i^j|^\wedge + N\, |\delta b_i^j|^\wedge + |\delta c_i^i|^\wedge + |\hat{\psi}_i^i|^\wedge\,\big]; \tag{9a}$$

$$|\delta a_i^j|^\wedge = \max_{i,j} |\delta a_i^j|, \quad |\delta b_i^j|^\wedge = \max_{i,j} |\delta b_i^j|, \quad |\delta c_i^i|^\wedge = \max_i |\delta c_i^i|,$$

$$|\hat{\psi}_i^i|^\wedge = \max\Big[\max_{i;i=1,N+1} |\delta a_i^i|,\ \max_i |\delta f_i|,\ \max_i |\delta\varphi_i|\Big].$$

Taking into account (3), for the value $z^*(x) - u(x)$, i.e., for the *error of the perturbed grid solution*, we obtain the estimate

$$\|u - z^*\|_{\overline{D}_h} \le M\,\big[(\varepsilon + N^{-1})^{-1}\, N^{-1} + \varepsilon\, N^2\, |\delta a_i^j|^\wedge + N\, |\delta b_i^j|^\wedge + |\delta c_i^i|^\wedge + |\hat{\psi}_i^i|^\wedge\,\big]. \tag{9b}$$

3.3 Estimates for the Perturbation of the Grid Solution

For the perturbation of the grid solution $z^*(x) - z(x)$, taking into account (9a), we obtain the estimate

$$\|z^* - z\|_{\overline{D}_h} \le M\, \eta(\varepsilon, \delta), \tag{10}$$

where $\eta(\varepsilon, \delta) = \eta(\varepsilon, \delta;\ \delta A,\ \delta F) = \varepsilon^{-1}\,\delta^{-2}\,|\delta a_i^j|^\wedge + \varepsilon^{-1}\,\delta^{-1}\,|\delta b_i^j|^\wedge + |\delta c_i^i|^\wedge + |\hat{\psi}_i^i|^\wedge$, $|\hat{\psi}_i^i|^\wedge = |\hat{\psi}_i^i|^\wedge_{(9)}$, $\delta = \delta_{(3)}(\varepsilon, N)$; the estimate is unimprovable up to a constant-factor.

3.4 Estimate for the Conditioning Number of the Difference Scheme

Definition 1 (See [3]). We write the unimprovable estimate (10) *in the variables* ε, δ in the form of the estimate to the *relative error* $\| z^* - z \|_{\overline{D}_h} \| z \|^{-1}_{\overline{D}_h}$ through *relative perturbations* in the data of the grid problem written in the matrix form (6)

$$\| z^* - z \|_{\overline{D}_h} \, / \, \| z \|_{\overline{D}_h} \leq æ_P(A; \overline{D}_h) \, (\|\delta \, F\| \, / \, \|F\| \, + \, \|\delta \, A\| \, / \, \|A\|).$$

We call the value $æ_P(A; \overline{D}_h)$ the conditioning number of the difference scheme (2) (see also the discussions of the matrix and problem conditioning in [8] for regular problems).

Taking into account estimate (10), we obtain the estimate $æ_P(A; \overline{D}_h)$ (similar to (5.13) from [3]):

$$æ_P(A; \overline{D}_h) \leq M \, \varepsilon^{-1} \, \delta^{-2}; \tag{11}$$

the estimate is unimprovable up to a constant-factor. The conditioning number $æ_P(A; \overline{D}_h)$ grows without bound as $\varepsilon \to 0$; the scheme (2) is not ε-uniformly well conditioned and it is not ε-uniformly stable to perturbations in the data of the grid problem.

4 Standard Difference Scheme Under Computer Perturbations

We consider the standard finite difference scheme in that case when perturbations of the solution are generated in the process of solving the discrete problem on a computer, for example, due to the finite number of computer word digits.

4.1 Computer Solution

We denote by Δ the maximum of perturbations in the data of the grid problem, caused by computer calculations. Let $z^*_\Delta(x), x \in \overline{D}_h$ be the corresponding *computer solution* (i.e., the perturbed solution obtained on a computer) of the difference scheme in the matrix form (6) and (7) under the condition

$$|\delta a_i^j|, \; |\delta b_i^j|, \; |\delta c_i^i|, \; |\delta f(x_i)| \leq \Delta, \;\; 2 \leq i \leq N; \; |\delta \varphi(x_i)| \leq \Delta, \;\; i = 1, N+1. \tag{12}$$

The function $z^*_\Delta(x)$, $x \in \overline{D}_h$ is the solution of difference scheme (8), in which the computer data perturbations δa^j_i, δb^j_i, δc^i_i, $\delta f(x_i)$, $\delta\varphi(x_i)$ satisfy the condition (12). We say that $z^*_\Delta(x)$ is the solution of the difference scheme (8) and (12), i.e., *the standard difference scheme under computer perturbations*. The condition (12) can also be seen as a condition imposed on perturbations in the data of the standard difference scheme under the perturbation of its data discussed in Sect. 3.

4.2 Estimates for the Computer Perturbation and for the Computer Solution

For the grid function $z^*_\Delta(x) - z(x)$, i.e., the *perturbation of the grid solution* caused by the *computer solution* $z^*_\Delta(x)$ (or, in short, the *computer perturbation*), taking into account (10), we obtain the following estimate in the variables ε, δ:

$$\|z^*_\Delta - z\|_{\overline{D}_h} \leq M\,\varepsilon^{-1}\,\delta^{-2}\;\;\Delta, \tag{13a}$$

where $M = 4M_{(10)}$. This estimate is equivalent to the following estimate in the variables ε, N:

$$\|z^*_\Delta - z\|_{\overline{D}_h} \leq M\,\varepsilon\,N^2\;\;\Delta\;. \tag{13b}$$

The estimate is unimprovable with respect to orders of incoming values.

For the grid function $z^*_\Delta(x) - u(x)$, i.e., the *error of the computer solution*, the following estimate holds:

$$\|u - z^*_\Delta\|_{\overline{D}_h} \leq\; \|u - z\|_{\overline{D}_h} + \|z^*_\Delta - z\|_{\overline{D}_h} \equiv \sigma(u - z; z^*_\Delta - z), \tag{14a}$$

where $\sigma(u - z; z^*_\Delta - z)$ is *the total error* of the *computer solution* (sum of the error to the solution of the standard scheme $\|u - z\|_{\overline{D}_h}$ and perturbation $\|z^*_\Delta - z\|_{\overline{D}_h}$). In the variables ε, δ, taking into account estimates (3) and (13), we obtain the estimate

$$\|u - z^*_\Delta\|_{\overline{D}_h} \leq M_1\,\delta + M_2\,\varepsilon^{-1}\,\delta^{-2}\;\;\Delta \leq M\left[\,\delta + \varepsilon^{-1}\,\delta^{-2}\;\;\Delta\;\right], \tag{14b}$$

where $M_1 = M_{(3)}$, $M_2 = M_{(13a)}$. In the variables ε, N we have the following estimate which is similar to one (9b):

$$\|u - z^*_\Delta\|_{\overline{D}_h} \leq M_1\,(\varepsilon + N^{-1})^{-1}\,N^{-1} + M_2\,\varepsilon\,N^2\;\;\Delta\;. \tag{14c}$$

The estimate (14) is unimprovable with respect to orders of incoming values.

Thus, the following theorem is valid (similar to Theorem 5 from [3]).

Theorem 1. *Let the conditions of Theorem 1 be satisfied. Then for the perturbation of the discrete solution and the error of the computer solution, the estimates* (13) *and* (14) *hold, respectively.*

5 Numerical Investigation of a Model Boundary Value Problem

In this section, for a model boundary value problem, using results of numerical experiments, we study perturbations of the solution to the standard finite difference scheme in the case of computer perturbations; results of numerical experiments are compared with theoretical results.

5.1 Standard and Perturbed Difference Schemes for a Model Boundary Value Problem

Formulation of a model boundary value problem. Standard finite difference scheme and the perturbed difference scheme in the case computer perturbations.

Consider the boundary value problem

$$L_{(1)}u(x) \equiv \Big\{\varepsilon a(x)\frac{d^2}{dx^2}+b(x)\frac{d}{dx}\Big\}u(x) = f(x),\ \ x \in D, \quad u(x) = \varphi(x),\ \ x \in \Gamma. \tag{15a}$$

Here $\overline{D} = [0,1]$, $a(x) = 1$, $b(x) = 2$, $f(x) = -2$, $\varphi(x) = 0$. The solution of problem (15a) is written out explicitly:

$$u(x) = (1-e^{-2\,\varepsilon^{-1}})^{-1}\,(1-e^{-2\,\varepsilon^{-1}\,x}) - x, \quad x \in \overline{D}. \tag{15b}$$

We approximate problem (15) by the standard difference scheme

$$\Lambda z(x) \equiv \{\varepsilon\,\delta_{\overline{x}\hat{x}} + 2\,\delta_x\}z(x) = -2,\ \ x \in D_h, \quad z(x) = 0,\ \ x \in \Gamma_h; \tag{16}$$

here $\overline{D}_h$ is the uniform grid with the step-size $h = N^{-1}$.

In the case of perturbations in the data, the following perturbed standard difference scheme corresponds to difference scheme (16):

$$\begin{aligned}\Lambda^*\,z^*(x) &\equiv \{\varepsilon\,a^*(x)\,\delta_{\overline{x}\hat{x}} + b^*(x)\,\delta_x\}\ z^*(x) = f^*(x),\ x \in D_h,\\ z^*(x) &= \varphi^*(x), \qquad\qquad\qquad\qquad\qquad\quad x \in \Gamma_h.\end{aligned} \tag{17}$$

The perturbed data in the scheme (17) are determined by the relations

$$a^*(x) = a_{(15)}(x) + \delta a^i_{i+1}, \quad b^*(x) = b_{(15)}(x) = 2, \tag{18}$$

$$f^*(x) = f_{(15)}(x) = -2, \qquad x = x^i, \ x^i \in \overline{D}_h; \qquad \varphi^*(x) = 0, \quad x \in \Gamma_h;$$

in numerical experiments, we set

$$\delta a^j_i = -\delta a, \quad \delta a = 10^{-8}; \quad j = i-1, i, i+1, \quad i = 1, 2, \ldots, N. \tag{19}$$

It should be noted that, in accordance with the estimate (10), the most significant impact into the perturbation of the discrete solution $z^*(x) - z(x)$ and into the error of the perturbed solution $z^*(x) - u(x)$ is introduced by perturbations in the coefficient multiplying the second-order derivative in the differential equation (1).

In the case of the difference scheme in the presence of computer perturbations [scheme (8) and (12)], considering the computer solution $z^*(x) = z^*_\Delta(x)$ and the computer solution error $\delta^*_{u/\Delta} = \|u - z^*_\Delta\|_{\overline{D}_h}$, we assume that the computer perturbations satisfy the condition (12), where

$$\Delta = \delta a. \tag{20}$$

Thus, for the boundary value problem (15), we have the perturbed difference scheme (17)–(19) in the case of perturbations in the data of the grid problem, and we have the perturbed difference scheme (17)–(20) in the case of computer perturbations.

We are interested in the behavior of the errors in solutions to the standard difference scheme and of the perturbation to the computer solutions, depending on the parameter ε and the number of grid intervals of N, and we are also interested to compare experimental results with theoretical.

5.2 *Numerical Experiments in the Variables ε and N*

We discuss the results of numerical experiments for errors in solutions to the standard finite difference scheme in the absence of perturbations [scheme (16)] and the computer difference scheme [scheme (17)–(20)].

5.2.1 Errors in the Solution of the Standard Difference Scheme in the Absence of Perturbations

Consider the behavior of the solution error δ_u to the standard difference scheme in the absence of perturbations, i.e., the scheme (16),

$$\delta_u = \delta_u(\varepsilon, N) = \|u - z\|_{\overline{D}_h\,(3)} \tag{21a}$$

Table 1 Errors of the grid solution $\boldsymbol{\delta}_u = \boldsymbol{\delta}_u(\varepsilon, N)$ for various values ε and N

$\varepsilon \backslash N$	2^2	2^4	2^6	2^8	2^{10}
1	$3.96e^{-2}$	$1.27e^{-2}$	$3.37e^{-3}$	$8.54e^{-4}$	$2.14e^{-4}$
2^{-2}	$1.90e^{-1}$	$7.59e^{-2}$	$2.17e^{-2}$	$5.64e^{-3}$	$1.42e^{-3}$
2^{-4}	–	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$	$5.67e^{-3}$
2^{-6}	–	–	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$
2^{-8}	–	–	–	$1.98e^{-1}$	$7.65e^{-2}$
2^{-10}	–	–	–	–	$1.98e^{-1}$

depending on the parameter ε and the number of grid intervals of N. In Table 1, errors of the grid solution $\delta_u(\varepsilon,\ N)$ are given for various values ε and N. Note that in according to the estimate (3), the error $\delta_u(\varepsilon,\ N)$ essentially depends on the product $\varepsilon\, N$.

From the results in Table 1 it follows that under the condition $N^{-1} \le \varepsilon$ (the mesh step-size h is less than the value of the perturbation parameter ε) the error of the grid solution $\delta_u = \|u - z\|_{\overline{D}_h}$ tends to zero as N grows for fixed values of ε. Note that the error for the linear interpolant $\overline{z}(x)$, $x \in \overline{D}$, in the uniform continuous norm $\|\cdot\|$ is a quantity of order to the error of the discrete solution in the uniform grid norm $\|\cdot\|_{\overline{D}_h}$.

However, provided that $N^{-1} > \varepsilon$ (the mesh step-size h is greater than the value of the parameter ε), the linear interpolant of the grid solution in the uniform continuous norm $\|\cdot\|$ leads to the error of order one, even when the error in the uniform grid norm $\|\cdot\|_{\overline{D}_h}$ is small (see, e.g., discussions in [9]). For this reason, the errors for $\varepsilon < N^{-1}$ are not given in Table 1. Thus, results in Table 1 qualitatively agree with the assertion of Theorem 1.

5.2.2 Errors in the Solution of the Computer Difference Scheme (17)–(20)

Discuss the behavior of the computer perturbation to the grid solution

$$\delta_z = \delta_z(\varepsilon, N; \Delta) = \|z^*_\Delta - z\|_{\overline{D}_h\,(13b)} \tag{22a}$$

in the case of various values of the parameter ε and the number of intervals N for a fixed value of the perturbation Δ. The computer perturbation $\delta_z(\varepsilon, N; \Delta)$, according to estimate (13b), essentially depends on the product $\varepsilon\, N^2$.

In Table 2, computer perturbations $\delta_z = \delta_z(\varepsilon, N; \Delta)$ of the solution of computer difference scheme (17)–(20) are given for various values ε and N. Unlike the behavior in the error δ_u of the grid solution in Table 1 which tends to zero as N grows for fixed ε, the computer perturbations of the grid solution $\delta_z = \delta_z(\varepsilon, N; \Delta)$ in Table 2 increase as N grows. Thus, the results in Table 2 qualitatively agree with the estimate (13b) from Theorem 1.

Table 2 Computer perturbations $\boldsymbol{\delta}_z = \boldsymbol{\delta}_z(\varepsilon, N; \Delta)$ for various values ε and N

$\varepsilon \setminus N$	2^2	2^4	2^6	2^8	2^{10}
1	$2.16e^{-9}$	$5.35e^{-8}$	$9.18e^{-7}$	$1.49e^{-5}$	$2.39e^{-4}$
2^{-2}	$2.90e^{-9}$	$7.48e^{-8}$	$1.29e^{-6}$	$2.10e^{-5}$	$3.37e^{-4}$
2^{-4}	$1.31e^{-9}$	$3.10e^{-8}$	$5.16e^{-7}$	$8.32e^{-6}$	$1.33e^{-4}$
2^{-6}	$3.75e^{-10}$	$9.00e^{-9}$	$1.47e^{-7}$	$2.38e^{-6}$	$3.82e^{-5}$
2^{-8}	$9.67e^{-11}$	$2.31e^{-9}$	$3.87e^{-8}$	$6.22e^{-7}$	$9.99e^{-6}$
2^{-10}	$2.43e^{-11}$	$5.80e^{-10}$	$9.82e^{-9}$	$1.58e^{-7}$	$2.53e^{-6}$

From comparison of the behavior of the error $\delta_u = \|u - z\|_{\overline{D}_h}$ and the computer perturbation $\delta_z = \delta_z(\varepsilon, N; \Delta)$, it follows that for fixed values ε and large values N these computer perturbations will exceed the errors of grid solutions to the unperturbed difference scheme that is qualitatively consistent with the estimate (14c) from Theorem 1.

Tables 1 and 2 in the variables ε and N are rather complicated. Therefore, it is interesting to consider similar tables, but in other variables, the so-called "automodel" simplifies the structure of the tables.

5.3 *Numerical Experiments in the Automodel Variables*

Discuss the behavior of errors in the solution of the standard difference scheme and of the computer perturbation of the grid solution with regard to their theoretical estimates (3) and (13b).

5.3.1 Errors in the Solution of the Standard Difference Scheme in the Absence of Perturbations in the Variables $\boldsymbol{\varepsilon}$ and $\boldsymbol{\beta}$

Consider errors in the solution of the standard difference scheme, using the variables ε and β, where $\beta = \varepsilon\, N$ is the automodel variable.

In Table 3, errors of the grid solution $\overline{\delta}_u = \overline{\delta}_u(\varepsilon, \beta)$ are given for various values ε and β, where $\beta = \beta(\varepsilon,\ N) = \varepsilon\, N$. Here also the values $\{\beta \max\limits_{\varepsilon} \overline{\delta}_u(\varepsilon,\ \beta)\}$ are given for various values β. Note that

$$\overline{\delta}_u = \overline{\delta}_u(\varepsilon, \beta) = \overline{\delta}_u(\varepsilon, \beta(\varepsilon,\ N)) = \delta_{u\,(21a)}(\varepsilon,\ N).$$

From the results in Table 1 it follows that for a fixed β, the values $\overline{\delta}_u(\varepsilon, \beta)$ rather weakly depend on the values of the parameter ε, and they stabilize quickly with decreasing of ε. The values $\{\beta \max\limits_{\varepsilon} \overline{\delta}_u(\varepsilon,\ \beta)\}$ are weakly dependent on β, and they stabilize quickly with increasing of β; the maximum of these values does not exceed 0.369.

Table 3 Errors of the grid solution $\overline{\delta}_u = \overline{\delta}_u(\varepsilon, \beta)$ for various values ε and β and also the values $\{\beta \max_{\varepsilon} \overline{\delta}_u(\varepsilon, \beta)\}$ for various values β

$\varepsilon \setminus \beta$	2^0	2^2	2^4	2^6	2^8	2^{10}
1		$3.96e^{-2}$	$1.27e^{-2}$	$3.37e^{-3}$	$8.54e^{-4}$	$2.14e^{-4}$
2^{-2}	$1.90e^{-1}$	$7.59e^{-2}$	$2.17e^{-2}$	$5.64e^{-3}$	$1.42e^{-3}$	$3.57e^{-4}$
2^{-4}	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$	$5.67e^{-3}$	$1.43e^{-3}$	$3.59e^{-4}$
2^{-6}	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$	$5.67e^{-3}$	$1.43e^{-3}$	$3.59e^{-4}$
2^{-8}	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$	$5.67e^{-3}$	$1.43e^{-3}$	$3.59e^{-4}$
2^{-10}	$1.98e^{-1}$	$7.65e^{-2}$	$2.18e^{-2}$	$5.67e^{-3}$	$1.43e^{-3}$	$3.59e^{-4}$
$\{\beta \max_{\varepsilon} \overline{\delta}_u(\varepsilon, \beta)\}$	**0.198**	**0.306**	**0.358**	**0.366**	**0.368**	**0.369**

Thus, in the case of the model problem for the error of the grid solution $\delta_{u\,(21)}(\varepsilon,\ N)$, using the results in Table 3, we obtain the experimental estimate

$$\delta_u(\varepsilon,\ N) \le M_1\, \varepsilon^{-1}\, N^{-1}, \tag{21b}$$

where (according to Table 3) we have

$$M_1 = \max_{\beta} \left\{\beta\ \max_{\varepsilon} \overline{\delta}_u(\varepsilon,\ \beta)\right\} = 0.369. \tag{21c}$$

The estimate (21) for the error in the solution of the standard difference scheme is fully consistent with the estimate (3) from Theorem 1.

5.3.2 Errors in the Solution of the Computer Difference Scheme in the Variables ε and γ

Consider the perturbations of grid solutions, caused by computer calculations, i.e., computer perturbations of grid solutions, using the variables ε and γ, where $\gamma = \varepsilon\, N^2$ is the automodel variable.

In Table 4, computer perturbations of the grid solution $\tilde{\delta}_z = \tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta)$ are given for various values ε and γ. Here also the values $\{(\gamma\quad \Delta)^{-1} \max_{\varepsilon} \tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta)\}$ are given for various values γ. Note that

$$\tilde{\delta}_z = \tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta) = \tilde{\delta}_z(\varepsilon,\ \gamma(\varepsilon,\ N);\ \Delta) = \delta_z(\varepsilon,\ N;\ \Delta).$$

From Table 4, it follows that the computer perturbations of the grid solution $\delta_z = \tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta)$ sufficiently weakly depend on the parameter ε; moreover, they are stabilized with decreasing ε for fixed values γ. For fixed values ε, the perturbations $\tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta)$ change significantly with increasing γ; these perturbations grow with increasing γ at the rate close to linear one for all values ε. Here also the values

Table 4 Computer perturbations $\tilde{\delta}_z = \tilde{\delta}_z(\varepsilon,\ \gamma;\ \Delta)$ for various values ε and γ and also the values $\{(\gamma\ \Delta)^{-1} \max\limits_{\varepsilon}(\tilde{\delta}_z(\varepsilon, \gamma; \Delta))\}$ for various values γ

$\varepsilon \backslash \gamma$	2^8	2^{10}	2^{12}	2^{14}	2^{16}	2^{18}
1	$5.35e^{-8}$	$2.24e^{-7}$	$9.18e^{-7}$	$3.71e^{-6}$	$1.49e^{-5}$	$5.98e^{-5}$
2^{-2}	$3.15e^{-7}$	$1.29e^{-6}$	$5.22e^{-6}$	$2.10e^{-5}$	$8.42e^{-5}$	$3.37e^{-4}$
2^{-4}	$5.16e^{-7}$	$2.07e^{-6}$	$8.32e^{-6}$	$3.33e^{-5}$	$1.33e^{-4}$	$5.34e^{-4}$
2^{-6}	$5.93e^{-7}$	$2.38e^{-6}$	$9.54e^{-6}$	$3.82e^{-5}$	$1.53e^{-4}$	$6.12e^{-4}$
2^{-8}	$6.22e^{-7}$	$2.49e^{-6}$	$9.99e^{-6}$	$4.00e^{-5}$	$1.60e^{-4}$	$6.41e^{-4}$
2^{-10}	$6.33e^{-7}$	$2.53e^{-6}$	$1.01e^{-5}$	$4.06e^{-5}$	$1.62e^{-4}$	$6.51e^{-4}$
$\{(\gamma\ \Delta)^{-1} \max\limits_{\varepsilon}(\tilde{\delta}_z(\varepsilon, \gamma; \Delta))\}$	**0.247**	**0.249**	**0.249**	**0.249**	**0.249**	**0.250**

$\{(\gamma\ \ \Delta)^{-1} \max\limits_{\varepsilon}(\tilde{\delta}_z(\varepsilon,\ \gamma;\ \ \Delta))\}$ are given for various values γ. These values weakly depend on the γ, and they stabilize quickly with increasing γ; the maximum of that ratio does not exceed a value of 0.250.

Thus, in the case of the model problem for the computer perturbations of the grid solution $\delta_z(\varepsilon, N; \Delta)$, using the results in Table 4, we obtain the experimental estimate

$$\delta_z(\varepsilon, N; \Delta) \le M_2\, \varepsilon\, N^2\ \Delta, \tag{22b}$$

where (according to Table 4) we have

$$M_2 = \max_{\gamma} \{(\gamma\ \ \Delta)^{-1} \max_{\varepsilon}(\tilde{\delta}_z(\varepsilon, \gamma; \Delta))\} = 0.250. \tag{22c}$$

The estimate (22) is fully consistent with the estimate (13b) from Theorem 1.

5.3.3 Experimental Estimate for the Error of the Perturbed Computer Solution

Taking into account the estimates (21) and (22), for the error of the perturbed computer solution $\delta_u^* = \delta_{u/\Delta}^* = \|u - z_\Delta^*\|_{\overline{D}_h}$, we obtain the experimental estimate in the variables $\{\varepsilon,\ N,\ \Delta\}$

$$\delta_u^* \le M_1\, (\varepsilon + N^{-1})^{-1}\, N^{-1} + M_2\, \varepsilon\, N^2\ \ \Delta\ . \tag{23}$$

The estimate (23) is fully consistent with the estimate (14c) from Theorem 1.

Thus, the numerical results in Tables 3 and 4 are in good agreement with the theoretical results.

6 Conclusions

In the case of the Dirichlet problem for a singularly perturbed ordinary differential convection-diffusion equation with perturbation parameter ε ($\varepsilon \in (0, 1]$), the standard difference scheme (classical scheme on a uniform grid) is considered in the presence of perturbations. Such a difference scheme is not ε-uniformly stable to perturbations in the data. Perturbations of grid solutions generated by *perturbations in the data of the grid problem* and *computer perturbations* are discussed. The data of numerical experiments to study the influence of computer perturbations on the grid solutions are presented. Results of numerical experiments are consistent with the theoretical results.

Acknowledgements This research was supported by the Russian Foundation for Basic Research under grant No.13-01-00618.

References

1. Samarskii, A.A.: Theory of Difference Schemes. Marcel Dekker Inc., New York (2001)
2. Shishkin, G.I., Shishkina, L.P.: Difference Methods for Singular Perturbation Problems. Monographs and Surveys in Pure and Applied Mathematics. Chapman and Hall/CRC, Boca Raton (2009)
3. Shishkin, G.I.: Conditioning of a difference scheme of the solution decomposition method for a singularly perturbed convection-diffusion equation. Trudy IMM UrO RAN **18**(2), 291–304 (2012) (in Russian)
4. Shishkin, G.I.: Stability of a Standard finite difference scheme for a singularly perturbed convection-diffusion equation. Doklady Math. **87**(1), 107–109 (2013)
5. Shishkin, G.I.: Data perturbation stability of difference schemes on uniform grids for a singularly perturbed convection-diffusion equation. Russian J. Numer. Anal. Math. Model. **28**(4), 381–417 (2013)
6. Shishkin, G.I.: Stability of difference schemes on uniform grids for a singularly perturbed convection-diffusion equation. In: Cangiani, A., Davidchack, R.L., Georgoulis, E., Gorban, A.N., Levesley, J., Tretyakov, M.V. (eds.) Numerical Mathematics and Advanced Applications 2011: Proceedings of ENUMATH 2011, the 9th European Conference on Numerical Mathematics and Advanced Applications, Leicester, September 2011, pp. 293–302. Springer, Berlin (2013)
7. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted numerical methods for singular perturbation problems. Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions, revised edn. World Scientific, Singapore (2012)
8. Bakhvalov, N.S., Zidkov, N.P., Kobelikov, G.M.: Numerical Methods. Laboratory of Basic Knowledge, Moscow (2001) (in Russian)
9. 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, New York (2000)

A Higher Order Immersed Discontinuous Galerkin Finite Element Method for the Acoustic Interface Problem

S. Adjerid and K. Moon

Abstract We present an interface discontinuous Galerkin finite element method on non-fitted meshes for solving acoustic wave propagation problems in nonhomogeneous media. The proposed method uses the standard discontinuous Galerkin finite element formulation with polynomial approximation on elements that contain one material while on interface elements containing multiple materials it uses a specially build piecewise polynomial shape functions that satisfy the interface jump conditions. We present several computational results that suggest that the proposed method has optimal convergence rates.

Keywords Immersed method • Discontinuous Galerkin • Acoustic problem

1 Introduction

Simulation of wave propagation in nonhomogeneous media arises in many applications in science and engineering such as geophysics, acoustics, and electromagnetism and leads to systems of partial differential equations with discontinuous coefficients. These problems present several challenges to scientists as they involve large time integration and wave propagation through bodies that are thousands of wavelengths in size which requires the solution of large problems with complex geometries.

The discontinuous Galerkin (DG) formulation is a natural choice for first-order linear hyperbolic systems leading to compact high-order schemes with low-dispersion and low-dissipation errors. The DG formulation can easily enforce boundary conditions on complex geometries and can handle discontinuous solutions

S. Adjerid (✉) • K. Moon
Department of Mathematics Virginia Tech, Blacksburg, VA 24061, USA
e-mail: adjerids@vt.edu

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_6,

up to element boundaries. There exist mainly two DG formulations to solve transient linear hyperbolic systems in the literature: (i) semi-discrete DG formulations using a method-of-lines approach [1–3] where the system is first discretized in space using the DG formulation combined with low-storage Runge–Kutta time integrators that have relatively large stability limits and low-dissipation and low-dispersion errors [4–7] and (ii) space–time explicit DG methods by Falk and Richter [8] on properly designed space–time meshes. Motivated by the work in [8] a space–time DG formulation has also been used for linear transient interface symmetric hyperbolic systems [9] where the problem is marched in time by solving sets of small problems in space–time.

Interface problems have been considered for a long time and many numerical methods have been developed. Both finite difference and finite element approaches can be employed (see [10–16] and references therein). Piraux and Lombard [17, 18] developed explicit interface method to solve acoustic wave propagation in nonhomogeneous media on non-fitted meshes. On the other hand higher order immersed finite element (IFE) methods have been developed for diffusive problems [19–25]. In this manuscript we present a higher order interface discontinuous Galerkin finite element (IDGFE) method for solving the acoustic problem in nonhomogeneous media on non-fitted meshes.

This manuscript is organized as follows. In Sect. 2 we state the problem and interface conditions. In Sect. 3 we construct the IFE shape functions and the immersed discontinuous Galerkin formulation. In Sect. 4 we present several numerical results and conclude in Sect. 5.

2 Problem Statement

Let u and p, respectively, be the velocity and pressure defined on the interval $I = (a, b)$ which is split into $I_1 = (a, \alpha)$ and $I_2 = (\alpha, b)$ such that $\bar{I} = \overline{I_1 \cup I_2}$.

Now we consider the acoustic interface problem on I where $\mathbf{U} = [u, p]^T$ satisfies

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0 \qquad x \in I \setminus \{\alpha\}, \qquad t > t_0, \tag{1a}$$

$$\mathbf{A}|_{I_i} = \mathbf{A}_i = \begin{pmatrix} 0 & \frac{1}{\rho_i} \\ \rho_i c_i^2 & 0 \end{pmatrix}, \qquad i = 1, 2, \tag{1b}$$

with ρ_i, c_i, respectively, being the densities and sound speeds in I_i. The matrix $\mathbf{A}$ can be split as $\mathbf{A} = \mathbf{A}^+ + \mathbf{A}^-$ such that $\mathbf{A}^+$ and A^-, respectively, have nonnegative and nonpositive eigenvalues, i.e., if $A = X \Lambda X^{-1}$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, then $\Lambda^{\pm} = \frac{\Lambda \pm |\Lambda|}{2}$ and $A^{\pm} = X\Lambda^{\pm}X^{-1}$.

This problem is subject to the inflow boundary conditions

$$\mathbf{A}_1^+\mathbf{U}|_{x=a} = \mathbf{g}_a, \qquad \mathbf{A}_2^-\mathbf{U}|_{x=b} = \mathbf{g}_b, \tag{1c}$$

initial conditions

$$\mathbf{U}(x,t_0) = \mathbf{U}_0(x),\ x \in I, \tag{1d}$$

and physical interface conditions

$$\begin{cases} [u]|_{x=\alpha} = 0, \\ [p]|_{x=\alpha} = 0, \end{cases} \tag{1e}$$

where $[u]|_{x=\alpha} = u(\alpha^+,t) - u(\alpha^-,t)$ is the jump of u across the interface $x = \alpha$.

3 Discontinuous Galerkin Discretization

In order to apply the discontinuous Galerkin method to the acoustic problem (1) we partition the interval I into N subintervals $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ and use polynomials on non-interface elements containing one material and specially designed piecewise polynomial shape functions on interface elements containing more than one material. We start by showing how to construct interface polynomial spaces and shape functions.

3.1 Interface Shape Functions

Applying the physical interface conditions (1e) all immersed shape functions must be continuous at the interface. To uniquely define high-degree immersed shape functions we need additional jump conditions (referred to as "extended jump conditions") derived in [17] from the acoustic equations (1a):

$$\begin{aligned} &\frac{\partial^{2k}}{\partial t^{2k}}\mathbf{U} = c^{2k}\frac{\partial^{2k}}{\partial x^{2k}}\mathbf{U}, \\ &\frac{\partial^{2k+1}}{\partial t^{2k+1}}\mathbf{U} = -c^{2k}\mathbf{A}\frac{\partial^{2k+1}}{\partial x^{2k+1}}\mathbf{U}, \qquad \text{for} \quad k \geq 0. \end{aligned}$$

The continuity of $\mathbf{U}$ and its time derivatives at the interface yields the jump conditions

$$\frac{\partial^{2k}}{\partial x^{2k}}\mathbf{U}(\alpha^+,t) = \begin{pmatrix} \left(\frac{c_1}{c_2}\right)^{2k} & 0 \\ 0 & \left(\frac{c_1}{c_2}\right)^{2k} \end{pmatrix} \frac{\partial^{2k}}{\partial x^{2k}}\mathbf{U}(\alpha^-,t), \tag{2a}$$

$$\frac{\partial^{2k+1}}{\partial x^{2k+1}}\mathbf{U}(\alpha^+,t) = \begin{pmatrix} \left(\frac{\rho_1}{\rho_2}\right)\left(\frac{c_1}{c_2}\right)^{2k+2} & 0 \\ 0 & \left(\frac{\rho_2}{\rho_1}\right)\left(\frac{c_1}{c_2}\right)^{2k} \end{pmatrix} \frac{\partial^{2k+1}}{\partial x^{2k+1}}\mathbf{U}(\alpha^-,t), \quad \text{(2b)}$$

for $k \geq 0$.

Now let us consider the reference interface element $[-1,1]$ containing an interface $\xi = \hat{\alpha}$ such that $-1 = \xi_0 < \xi_1 < \cdots < \xi_{i-1} < \hat{\alpha} < \xi_i < \cdots < \xi_q = 1$ and let $L_j,\ j = 0,1,\ldots,q$ be the standard Lagrange polynomials such that $L_j(\xi_l) = \delta_{jl}$. Without loss of generality we assume that $\xi_j \neq \hat{\alpha}$ and let $\mathcal{I}_1 = \{0,1,\ldots,i-1\}$ and $\mathcal{I}_2 = \{i,i+1,\ldots,q\}$ and $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. We express the qth-degree immersed shape functions in terms of L_j as

$$\phi_i^{(q)} = \begin{cases} \phi_i^{(q),1} & = L_i + \sum\limits_{j\in\mathcal{I}\backslash\mathcal{I}_s} c_j L_j, \quad \text{on } (-1,\hat{\alpha}), \\ \phi_i^{(q),2} & = \sum\limits_{j\in\mathcal{I}_s} c_j L_j, \quad \text{on } (\hat{\alpha},1), \end{cases} \qquad i \in \mathcal{I}_s, s = 1,2. \quad (3)$$

We note that these immersed shape functions are of Lagrange type, i.e., $\phi_i^{(q)}$ $(\xi_j) = \delta_{ij}$.

The constants c_j for the velocity are determined by the jump conditions

$$\phi_i^{(q),1}(\hat{\alpha}^-) = \phi_i^{(q),2}(\hat{\alpha}^+) \quad (4)$$

$$\frac{\partial^k \phi_i^{(q),1}(\hat{\alpha}^-)}{\partial \xi^k} = r_k \frac{\partial^k \phi_i^{(q),2}(\hat{\alpha}^+)}{\partial \xi^k}, \qquad k = 1,2,\ldots,q \quad (5)$$

where for $k = 2l$(even) and $k = 2l+1$(odd)

$$r_{2l} = \left(\frac{c_1}{c_2}\right)^{2l}, \qquad r_{2l+1} = \left(\frac{\rho_1}{\rho_2}\right)\left(\frac{c_1}{c_2}\right)^{2l+2}. \quad (6)$$

The immersed shape functions for the pressure are obtained in a similar manner with r_{2l+1} replaced by

$$r_{2l+1} = \left(\frac{\rho_2}{\rho_1}\right)\left(\frac{c_1}{c_2}\right)^{2l}. \quad (7)$$

We present qth-degree immersed shape functions on $[-1,1]$, for $\hat{\alpha} = 0.4$, $c_1 = 1$, $\rho_1 = 2$, $c_2 = 2$, $\rho_2 = 4$ and $q = 1,2,3,4$, in Fig. 1.

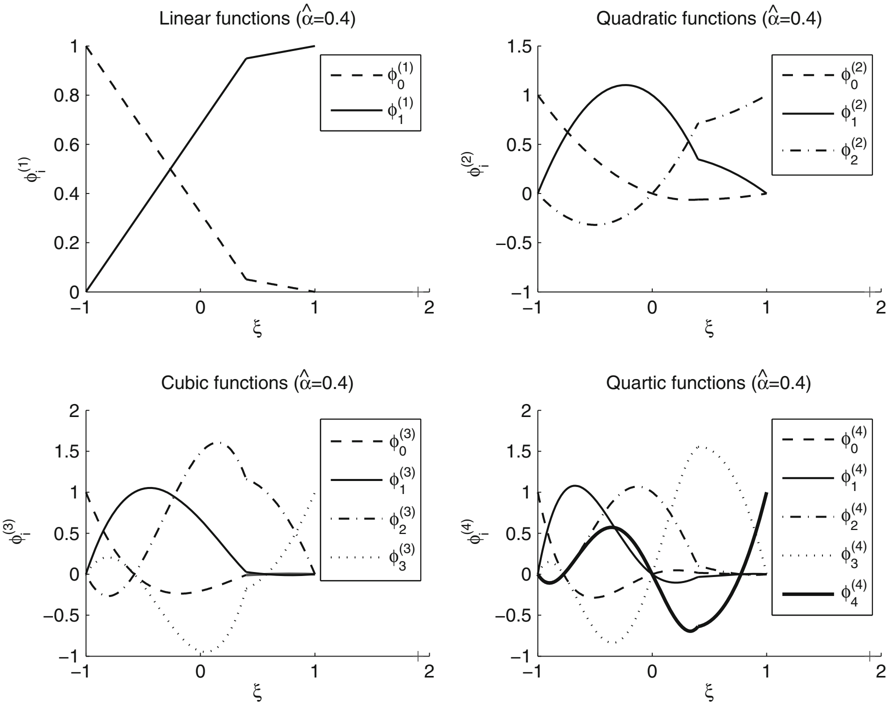


Fig. 1 Immersed shape functions for the velocity

3.2 DG Formulation

We use the standard DG weak formulation on a non-interface element (x_l, x_{l+1}) by multiplying (1a) by a test function $\mathbf{V}$, integrating over the element, and integrating by parts to obtain

$$\int_{x_l}^{x_{l+1}} \mathbf{V}^T \frac{\partial \mathbf{U}}{\partial t} dx + \mathbf{V}^T \mathbf{A}\mathbf{U} \mid_{x_l}^{x_{l+1}} - \int_{x_l}^{x_{l+1}} \frac{\partial \mathbf{V}^T}{\partial x} \mathbf{A}\mathbf{U} dx = 0. \qquad (8)$$

Apply flux splitting $\mathbf{A} = \mathbf{A}^- + \mathbf{A}^+$ to define the DG formulation to find $\mathbf{U}_h \in \mathcal{P}_q$, the space of polynomials of degree not exceeding q, such that

$$\begin{aligned} &\int_{x_l}^{x_{l+1}} \mathbf{V}_h^T \frac{\partial \mathbf{U}_h}{\partial t} dx + \mathbf{V}_h^{-T} \mathbf{A}^+ \mathbf{U}_h^- \mid_{x=x_{l+1}} + \mathbf{V}_h^{-T} \mathbf{A}^- \mathbf{U}_h^+ \mid_{x=x_{l+1}} \\ &\quad - \mathbf{V}_h^{+T} \mathbf{A}^+ \mathbf{U}_h^- \mid_{x=x_l} - \mathbf{V}_h^{+T} \mathbf{A}^- \mathbf{U}_h^+ \mid_{x=x_l} \\ &\quad - \int_{x_l}^{x_{l+1}} \frac{\partial \mathbf{V}_h^T}{\partial x} \mathbf{A}\mathbf{U}_h dx = 0, \ \forall \ \mathbf{V}_h \in \mathcal{P}_q. \end{aligned} \qquad (9)$$

On an interface element (x_l, x_{l+1}) containing an interface point $x = \alpha$, let $\mathcal{V}_{h,\text{IFE}}$ be the space spanned by the shape functions $\psi_i^{(q)}$, $i = 0, 1, \ldots, q$ obtained from $\phi_i^{(q)}$ through the affine mapping from $[-1, 1] \rightarrow [x_l, x_{l+1}]$. We multiply (1a) by a test function $\mathbf{V}_h \in \mathcal{V}_{h,\text{IFE}}$ and integrate over the element. We integrate by parts on (x_l, α) and (α, x_{l+1}). The discrete DG formulation is obtained by approximating $\mathbf{U}$ on (x_l, x_{l+1}) by $\mathbf{U}_h \in \mathcal{V}_{h,\text{IFE}}$ and the boundary terms at x_l and x_{l+1} are approximated using numerical fluxes. We note that the resulting IDGFE formulation contains a penalty term at the interface $x = \alpha$.

Multiply (1a) by $\mathbf{V}$ and integrate over (x_l, α) to obtain

$$\int_{x_l}^{\alpha} \mathbf{V}_1^T \frac{\partial \mathbf{U}_1}{\partial x} dx + \mathbf{V}_1^T \mathbf{A}_1 \mathbf{U}_1 \mid_{x=\alpha} - \mathbf{V}_1^T \mathbf{A}_1 \mathbf{U}_1 \mid_{x=x_l} - \int_{x_l}^{\alpha} \frac{\partial \mathbf{V}_1}{\partial x}^T \mathbf{A}_1 \mathbf{U}_1 dx = 0.$$

Similarly, integrate over (α, x_{l+1}) and integrate by parts to obtain

$$\int_{\alpha}^{x_{l+1}} \mathbf{V}_2^T \frac{\partial \mathbf{U}_2}{\partial t} dx + \mathbf{V}_2^T \mathbf{A}_2 \mathbf{U}_2 \mid_{x=x_{l+1}} - \mathbf{V}_2^T \mathbf{A}_2 \mathbf{U}_2 \mid_{x=\alpha} - \int_{\alpha}^{x_{l+1}} \frac{\partial \mathbf{V}_2}{\partial x}^T \mathbf{A}_2 \mathbf{U}_2 dx = 0.$$

Combining the previous two equations and applying the flux splitting at $x = x_l$ and $x = x_{l+1}$ yield the DG formulation on the interface element (x_l, x_{l+1}) which consists of determining $\mathbf{U}_h \in \mathcal{V}_{h,\text{IFE}}$ such that

$$\begin{aligned}
&\int_{x_l}^{\alpha} \mathbf{V}_{1,h}^T \frac{\partial \mathbf{U}_{1,h}}{\partial t} dx + \int_{\alpha}^{x_{l+1}} \mathbf{V}_{2,h}^T \frac{\partial \mathbf{U}_{2,h}}{\partial t} dx + {\mathbf{V}_{2,h}^-}^T \mathbf{A}_2^+ \mathbf{U}_{2,h}^- \mid_{x=x_{l+1}} \\
&+ {\mathbf{V}_{2,h}^-}^T \mathbf{A}_2^- \mathbf{U}_{2,h}^+ \mid_{x=x_{l+1}} - {\mathbf{V}_{1,h}^+}^T \mathbf{A}_1^+ \mathbf{U}_{1,h}^- \mid_{x=x_l} - {\mathbf{V}_{1,h}^+}^T \mathbf{A}_1^- \mathbf{U}_{1,h}^+ \mid_{x=x_l} \\
&+ \mathbf{V}_{1,h}^T \mathbf{A}_1 \mathbf{U}_{1,h} \mid_{x=\alpha} - \mathbf{V}_{2,h}^T \mathbf{A}_2 \mathbf{U}_{2,h} \mid_{x=\alpha} - \int_{x_l}^{\alpha} \frac{\partial \mathbf{V}_{1,h}^T}{\partial x} \mathbf{A}_1 \mathbf{U}_{1,h} dx \\
&- \int_{\alpha}^{x_{l+1}} \frac{\partial \mathbf{V}_{2,h}^T}{\partial x} \mathbf{A}_2 \mathbf{U}_{2,h} dx = 0, \qquad \forall\, \mathbf{V}_h \in \mathcal{V}_{h,\text{IFE}},
\end{aligned} \tag{10}$$

where $\mathbf{V}_{i,h} = \mathbf{V}_h \mid_{I_i \cap (x_l, x_{l+1})}$ and $\mathbf{U}_{i,h} = \mathbf{U}_h \mid_{I_i \cap (x_l, x_{l+1})}$ for $i = 1, 2$.

Combining the previous equation for the interface element and the DG formulation on non-interface elements we obtain the IDGFE method for solving the acoustic problem on non-fitted meshes.

4 Computational Examples

Consider the acoustic interface problem (1) with an incident wave

$$u^i(x,t) = u_0(x - c_1 t), \qquad p^i(x,t) = p_0(x - c_1 t), \tag{11a}$$

with the initial pulse

$$\mathbf{U}_0(x) = -f_0\left(t_0 - \frac{x}{c_1}\right)\begin{bmatrix}\frac{1}{c_1}\\ \rho_1\end{bmatrix} = \begin{bmatrix}u_0\\ p_0\end{bmatrix}, \tag{11b}$$

where

$$f_0(\xi) = \begin{cases}\sin(w_c\xi) - \frac{21}{32}\sin(2w_c\xi) + \frac{63}{768}\sin(4w_c\xi) - \frac{1}{512}\sin(8w_c\xi), & \text{if } 0 < \xi < \frac{1}{f_c},\\ 0, & \text{elsewhere.}\end{cases} \tag{11c}$$

When the incident wave hits the interface $x = \alpha$ it results in a reflected wave

$$\begin{aligned} u^r(x,t) &= \frac{c_2\rho_2 - c_1\rho_1}{c_1\rho_1 + c_2\rho_2} u_0(x + c_1 t - 2(\alpha - \delta)),\\ p^r(x,t) &= \frac{c_1\rho_1 - c_2\rho_2}{c_1\rho_1 + c_2\rho_2} p_0(x + c_1 t - 2(\alpha - \delta)), \end{aligned} \tag{11d}$$

with $\delta = c_1(t_0 - \frac{1}{2f_c})$, and a transmitted wave

$$\begin{aligned} u^t(x,t) &= \frac{2c_1\rho_1}{c_1\rho_1 + c_2\rho_2} u_0\left(\frac{c_1}{c_2}(x - c_2 t - \alpha) + \alpha\right),\\ p^t(x,t) &= \frac{2c_2\rho_2}{c_1\rho_1 + c_2\rho_2} p_0\left(\frac{c_1}{c_2}(x - c_2 t - \alpha) + \alpha\right), \end{aligned} \tag{11e}$$

Thus, the true solution of (1) can be written as

$$\mathbf{U}(x,t) = \begin{cases}\mathbf{U}^i(x,t) + \mathbf{U}^r(x,t) & \text{if } x < \alpha,\\ \mathbf{U}^t(x,t) & \text{if } x > \alpha,\end{cases} \tag{11f}$$

where $\mathbf{U}^i = [u^i, p^i]^T$, $\mathbf{U}^r = [u^r, p^r]^T$ and $\mathbf{U}^t = [u^t, p^t]^T$.

Example 4.1. As a test problem we select the physical parameters $c_1 = 1\,\text{m/s}$, $\rho_1 = 2\,\text{kg/m}^3$, $c_2 = 2\,\text{m/s}$, $\rho_2 = 4\,\text{kg/m}^3$, $\alpha = 10^{-4}\,\text{m}$, $t_0 = 0\,\text{s}$, $w_c = 2\pi f_c$, and $f_c = 0.5\,\text{Hz}$ and solve the acoustic problem (1) on the interval $[-5, 5]$ using uniform non-fitted meshes having $N = 100, 120, 130, 140, 150$ elements and polynomial spaces of degrees $q = 1, 2, 3, 4$ and integrate in time using the classical fourth-order Runge–Kutta method from $t = 0$ to $t = 2$ with time steps $\Delta t = d\ \Delta x$, $d = 10^{-q}$, $\Delta x = 10/N$ is the mesh size. We present the L^2 errors and their orders of convergence at $t = 2$ in Tables 1, 2, 3, and 4 that suggest that the proposed IDGFE yields optimal convergence rates. We plot the true and numerical solutions for $N = 150$ and $q = 4$ at $t = 0, 2$ in Fig. 2 to show that both solutions coincide.

Table 1 L^2 errors and orders of convergence for Example 4.1 with $q = 1$ at $t = 2$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
100	$2.9207e-2$		$2.6743e-2$	
110	$2.3162e-2$	2.4331	$2.1103e-2$	2.4849
120	$1.8634e-2$	2.4999	$1.6907e-2$	2.5482
130	$1.5200e-2$	2.5448	$1.3735e-2$	2.5957
140	$1.2563e-2$	2.5712	$1.1302e-2$	2.6311
150	$1.0512e-2$	2.5830	$9.4087e-3$	2.6571

Table 2 L^2 errors and orders of convergence for Example 4.1 with $q = 2$ at $t = 2$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
100	$1.1669e-3$		$7.0329e-4$	
110	$8.3352e-4$	3.5301	$4.7916e-4$	4.0261
120	$6.1460e-4$	3.5016	$3.3702e-4$	4.0443
130	$4.6598e-4$	3.4586	$2.4414e-4$	4.0276
140	$3.6196e-4$	3.4087	$1.8174e-4$	3.9835
150	$2.8710e-4$	3.3581	$1.3866e-4$	3.9207

Table 3 L^2 errors and orders of convergence for Example 4.1 with $q = 3$ at $t = 2$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
100	$5.6914e-5$		$2.5968e-5$	
110	$4.0138e-4$	−20.49	$4.0096e-4$	−28.71
120	$2.5217e-5$	31.805	$1.0286e-5$	42.098
130	$1.7874e-5$	4.3001	$7.0391e-6$	4.7393
140	$1.3064e-5$	4.2298	$5.0186e-6$	4.5655
150	$9.7916e-6$	4.1792	$3.6965e-6$	4.4316

Table 4 L^2 errors and orders of convergence for Example 4.1 with $q = 4$ at $t = 2$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
100	$4.4295e-5$		$4.4102e-5$	
110	$1.4661e-6$	35.76	$5.7277e-7$	45.57
120	$9.3950e-7$	5.1148	$3.6126e-7$	5.2969
130	$6.2588e-7$	5.0746	$2.3848e-7$	5.1885
140	$4.3041e-7$	5.0524	$1.6308e-7$	5.1284
150	$3.0400e-7$	5.0398	$1.1475e-7$	5.0949

Example 4.2. Let us consider the acoustic problem where medium 1 is water and medium 2 is air with $c_1 = 1,450\,\text{m/s}$, $\rho_1 = 1,000\,\text{kg/m}^3$, $c_2 = 340\,\text{m/s}$, $\rho_2 = 1.3\,\text{kg/m}^3$, $t_0 = 0.051\,\text{s}$, $w_c = 2\pi f_c$ and $f_c = 50\,\text{Hz}$ and the true solution is given by (11).

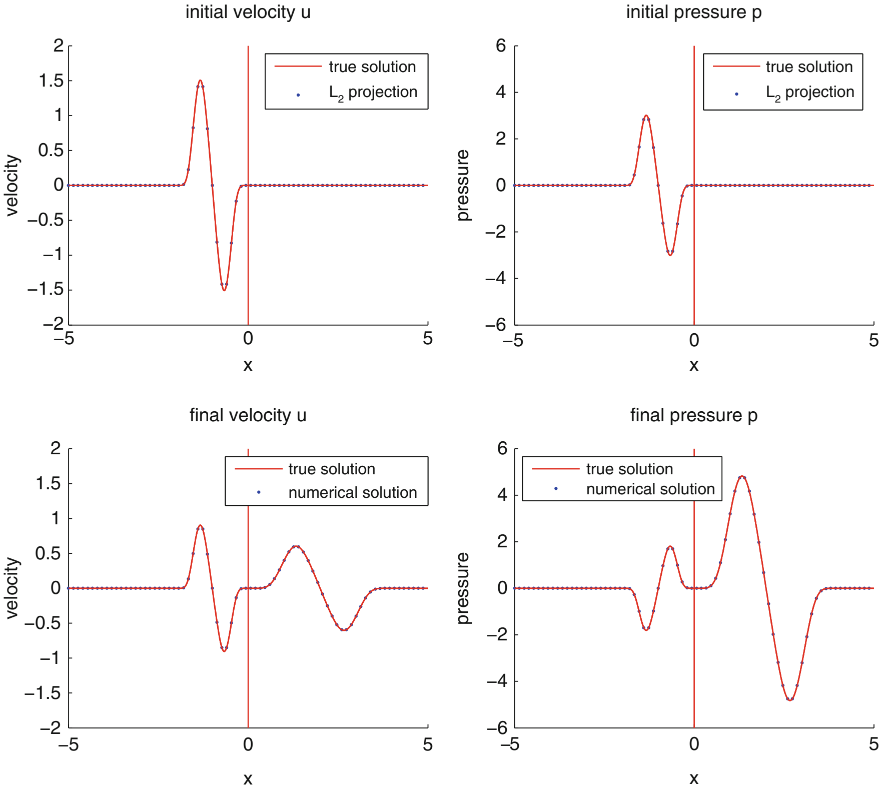


Fig. 2 True and numerical velocity and pressure for Example 4.1 using $N = 150$ elements and $q = 4$ at $t = 0$ (*top*) and $t = 2$ (*bottom*)

We solve this problem on $[40, 140]$ with the interface point $\alpha = 96.3\,\text{m}$ using uniform meshes having $N = 105, 121, 208, 224, 240$ elements (having one interface element) for degrees $q = 1, 2, 3, 4$ and integrate from $t = 0.051\,\text{s}$ to $t = 0.091\,\text{s}$ using the classical fourth-order Runge–Kutta method with time step sizes $\Delta t = d\ \Delta x$, $d = 10^{-(q+3)}$ and $\Delta x = 100/N$. The IDGFE solution for $q = 1$ exhibits large dispersion errors and thus optimal convergence rates are not observed in this case. As predicted in [26] high-degree approximations greatly reduce the dispersion errors and yields optimal $O(h^{q+1})$ convergence rates as shown in Tables 5, 6, and 7. The true and numerical velocity and pressure coincide as shown in Fig. 3. In this problem the transmitted IDGFE pressure is much smaller than the incident and reflected pressures and is plotted separately in Fig. 4 which coincides with the true pressure.

Example 4.3. Our IDGFE method has been extended to the two-dimensional acoustic problem on $[0, 20]^2$ and $-2 < t < 3$ where a planar wave hits the linear

Table 5 L^2 errors and orders of convergence for Example 4.2 with $q = 2$ at $t = 0.091$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
105	$1.8842e{-1}$		$6.8117e{-3}$	
121	$1.3195e{-1}$	2.5120	$4.3000e{-3}$	3.2434
208	$3.3138e{-2}$	2.5505	$6.9107e{-4}$	3.3745
224	$2.6941e{-2}$	2.7939	$5.5910e{-4}$	2.8596
240	$2.2105e{-2}$	2.8673	$4.5885e{-4}$	2.8640

Table 6 L^2 errors and orders of convergence for Example 4.2 with $q = 3$ at $t = 0.091$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
105	$1.5431e{-2}$		$6.3989e{-4}$	
121	$8.1492e{-3}$	4.5014	$3.9177e{-4}$	3.4593
208	$1.4417e{-3}$	3.1972	$5.2776e{-5}$	3.7003
224	$9.3264e{-4}$	5.8776	$3.9786e{-5}$	3.8126
240	$6.0543e{-4}$	6.2628	$3.0393e{-5}$	3.9035

Table 7 L^2 errors and orders of convergence for Example 4.2 with $q = 4$ at $t = 0.091$

N	$\frac{\|u-u_h\|_{L_2}}{\|u\|_{L_2}}$	Order	$\frac{\|p-p_h\|_{L_2}}{\|p\|_{L_2}}$	Order
105	$1.0735e{-3}$		$1.0388e{-4}$	
121	$5.0048e{-4}$	5.3802	$5.4430e{-5}$	4.5567
208	$2.5220e{-5}$	5.5154	$4.5425e{-6}$	4.5841
224	$1.1362e{-5}$	1.0760	$3.1243e{-6}$	5.0501
240	$5.9673e{-6}$	9.3337	$2.1905e{-6}$	5.1467

oblique interface $y = -5x + 70.05555$. The true and numerical solutions at $t = 3$ shown in Fig. 5 obtained on a 100×100 uniform Cartesian mesh for piecewise bilinear approximations are in full agreement. More details on the two-dimensional implementation of IDGFE method will be discussed in a forthcoming paper [27].

5 Conclusion

A higher order immersed discontinuous Galerkin finite element method is developed for acoustic wave propagation in a nonhomogeneous medium. On non-interface elements we use the standard DG formulation for hyperbolic systems with polynomial approximations while on an interface element containing two materials we construct piecewise polynomial IFE shape functions satisfying appropriate jump conditions across the interface. The new immersed shape functions are combined with a DG formulation that leads to a very efficient and conservative higher order finite element

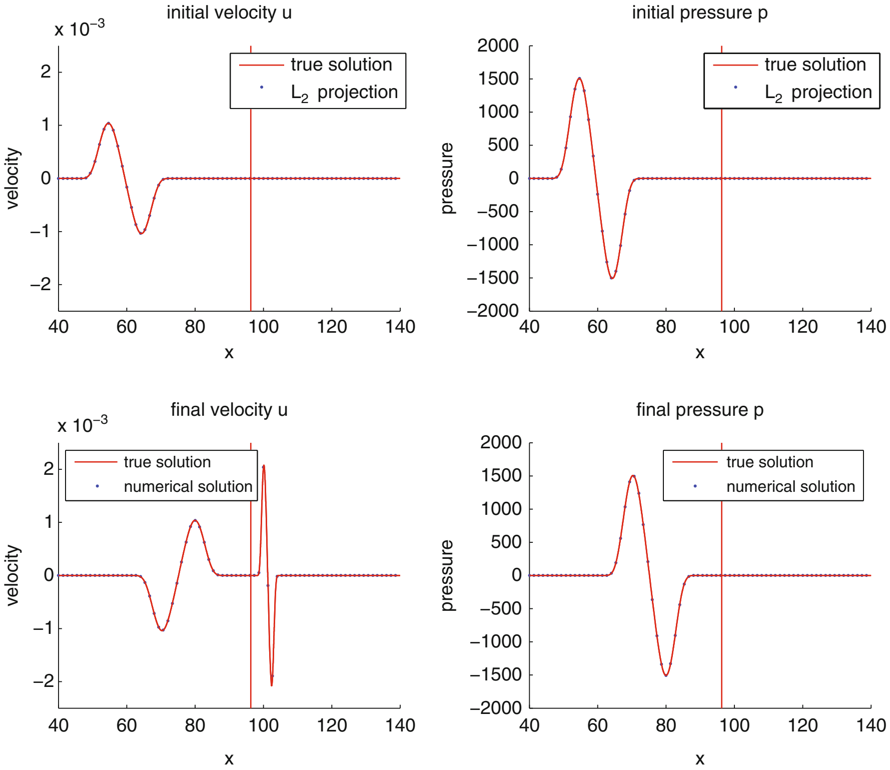


Fig. 3 True and numerical velocity and pressure for Example 4.2 with $N = 300$ and $q = 4$ at $t = 0.051$ (*top*) and $t = 0.091$ (*bottom*)

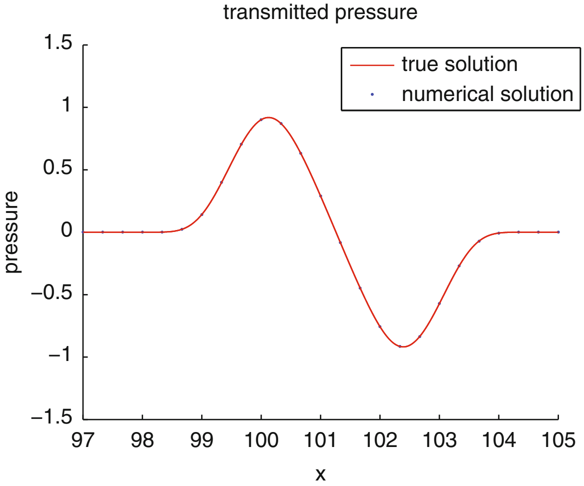


Fig. 4 True and numerical transmitted pressure for Example 4.2 with $N = 300$, $q = 4$ at $t = 0.091$

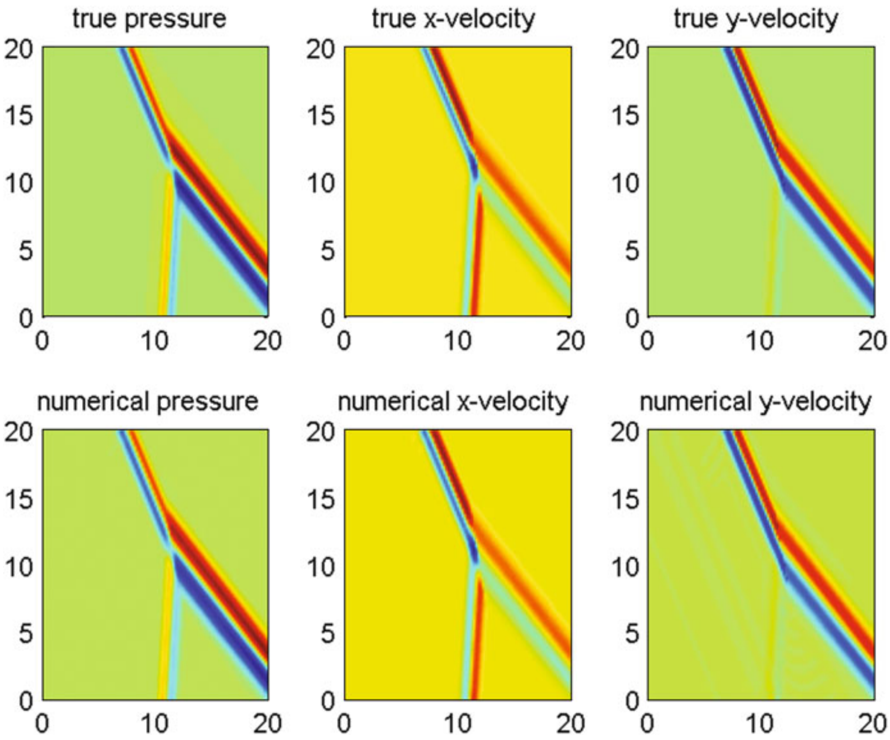


Fig. 5 True (*top*) and IDGFE (*bottom*) solutions for Example 4.3 at $t = 3$

method for wave propagation in a nonhomogeneous medium which yields optimal convergence rates for both the velocity and pressure. Several challenges such as solving two-dimensional wave propagation problems with curved and/or moving interfaces remain to be addressed.

Acknowledgements This research was partially supported by the National Science Foundation (Grant Number DMS 1016313).

References

1. Hesthaven, J., Warburton, T.: Nodal high-order methods on unstructured grids-I. Time-domain solution of Maxwell's equations. J. Comput. Phys. **181**, 186–221 (2002)
2. Minoli, C.A.A., Kopriva, D.A.: Discontinuous Galerkin spectral element approximations on moving meshes. J. Comput. Phys. **230**, 1876–1902 (2011)
3. Wilcox, L.C., Stadler, G., Burstedde, C., Ghattas, O.: A high-order discontinuous Galerkin method for wave propagation through coupled elastic–acoustic media. J. Comput. Phys. **229**(24), 9373–9396 (2010)

4. Carpenter, M.H., Kennedy, C.A.: Fourth order 2N-storage Runge Kutta scheme. Techical Memorandum NASA-TM-109112, NASA Langley Research Center, Hampton (1994)
5. Manthey, J.L., Hu, F.Q., Hussaini, M.Y.: Low-dissipation and low-dispersion Runge-Kutta schemes for computational acoustics. J. Comput. Phys. **124**, 177–191 (1996)
6. Hu, F.Q., Hussaini, M.Y., Rasetarinera, P.: An analysis of the discontinuous Galerkin method for wave propagation problems. J. Comput. Phys. **151**, 921–946 (1999)
7. Williamson, J.H.: Low storage Runge Kutta schemes. J. Comput. Phys. **35**, 48–56 (1980)
8. Falk, R., Richter, G.: Explicit finite element methods for symmetric hyperbolic equations. SIAM J. Numer. Anal. **36**, 935–952 (1999)
9. Monk, P., Richter, G.R.: A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media. J. Sci. Comput. **22–23**, 443–477 (2005)
10. Babuška, I.: The finite element method for elliptic equations with discontinuous coefficients. Computing **5**, 207–213 (1970)
11. Bramble, J.H., King, J.T.: A finite element method for interface problems in domains with smooth boundary and interfaces. Adv. Comput. Math. **6**, 109–138 (1996)
12. Chen, Z., Zou, J.: Finite element methods and their convergence for elliptic and parabolic interface problems. Numer. Math. **79**, 175–202 (1998)
13. Heinrich, B.: Finite Difference Methods on Irregular Networks. International Series of Numerical Mathematics, vol. 82. Birkhäuser, Boston (1987)
14. Li, Z., Ito, K.: The Immersed Interface Method: Numerical Solutions of PDEs Involving Interfaces and Irregular Domains. Frontiers in Applied Mathematics, vol. 33. Cambridge University Press, SIAM, Cambridge (2006)
15. Samarskiĭ, A.A., Andreev, V.B.: Méthodes aux Différences pour Équations Elliptiques. Mir, Moscow (1978)
16. Xu, J.: Estimate of the convergence rate of the finite element solutions to elliptic equation of second order with discontinuous coefficients. Nat. Sci. J. Xiangtan Univ. **1**, 1–5 (1982)
17. Lombard, B., Piraux, J.: A new interface method for hyperbolic problems with discontinuous coefficients: one-dimensional acoustic example. J. Comput. Phys. **1168**, 227–248 (2001)
18. Lombard, B., Piraux, J.: Numerical treatment of two dimensional interfaces for acoustic and elastic waves. J. Comput. Phys. **195**, 90–116 (2004)
19. Adjerid, S., Ben-Romdhane, M., Lin, T.: Higher degree immersed finite element methods for elliptic interface problems. Int. J. Numer. Anal. Model. **11**(3), 541–566 (2014)
20. Adjerid, S., Lin, T.: Higher-order immersed discontinuous Galerkin methods. Int. J. Inform. Syst. Sci. **3**(4), 555–568 (2007)
21. Adjerid, S., Lin, T.: p-th degree immersed finite element for boundary value problems with discontinuous coefficients. Appl. Numer. Math. **59**(6),1303–1321 (2009)
22. He, X., Lin, T., Lin Y.: Approximation capability of a bilinear immersed finite element space. Numer. Meth. Part. Differ. Equat. **24**, 1265–1300 (2008)
23. He, X., Lin, T., Lin, Y.: Immersed finite element methods for elliptic interface problems with non-homogeneous jump conditions. Int. J. Numer. Anal. Model. **8**(2), 284–301 (2011)
24. He, X., Lin, T., Lin, Y., Zhang, X.: Immersed finite element methods for parabolic equations with moving interface. Numer. Meth. Part. Differ. Equat. **29**(2), 619–646 (2013)
25. Kafafy, R., Lin, T., Lin, Y., Wang, J.: Three-dimensional immersed finite element methods for electric field simulation in composite materials. Int. J. Numer. Meth. Eng. **64**, 904–972 (2005)
26. Ainsworth, M.: Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. J. Comput. Phys. **198**, 106–130 (2004)
27. Adjerid, S., Moon, K.: A high order immersed discontinuous Galerkin method for the two dimensional acoustic problem (2014, in preparation)

A Parameter-Uniform Numerical Method for a Boundary Value Problem for a Singularly Perturbed Delay Differential Equation

M. Manikandan, N. Shivaranjani, J.J.H. Miller, and S. Valarmathi

Abstract In this paper, a boundary value problem for a second-order singularly perturbed delay differential equation is considered. The solution of this problem exhibits boundary layers at $x = 0$ and $x = 2$ and interior layers at $x = 1$. A numerical method composed of a classical finite difference scheme applied on a piecewise-uniform Shishkin mesh is suggested to solve the problem. The method is proved to be first-order convergent in the maximum norm uniformly in the perturbation parameter. Numerical illustrations support the theory.

Keywords Singular perturbation problems • Boundary layers • Delay differential equations • Finite difference scheme • Shishkin mesh • Parameter-uniform convergence

1 Introduction

Delay differential equations play an important role in the mathematical modeling of various practical phenomena in the subjects of bioscience and control theory. Singularly perturbed delay differential equations arise frequently in the modeling of human pupil-light reflex, the study of bistable devices and variational problems in control theory.

In [1], Lange and Miura give asymptotic expansion approximations for the solutions of singularly perturbed second-order delay differential equations with

M. Manikandan (✉) • N. Shivaranjani • S. Valarmathi
Department of Mathematics, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India
e-mail: manimaths89@yahoo.com; valarmathi07@gmail.com

J.J.H. Miller
Institute for Numerical Computation and Analysis, Dublin, Ireland
e-mail: jm@incaireland.org

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_7,

small delay. In [2], a numerical method is suggested for the initial value problem for a class of delay differential equations. The method uses a hybrid difference scheme on a fitted mesh and is proved to be second-order convergent under the assumption that $C\varepsilon \leq N^{-1}$. In [3], a fitted operator scheme on a uniform mesh is suggested to solve an initial value problem for a class of linear and hence a class of semilinear first-order delay differential equations. Further, the method is proved to be first-order parameter-uniform convergent. In [4], under the assumption of analyticity of the input data, the hp version of the finite element method on an appropriate mesh is proved to have an exponential rate of convergence. In [5], a numerical method known as initial value technique is suggested to solve singularly perturbed boundary value problems for second-order delay differential equations of reaction-diffusion type. In [6], the same technique is applied to solve a system of second-order delay differential equations of reaction-diffusion type.

In this paper we consider a boundary value problem for a singularly perturbed delay differential equation of reaction-diffusion type. We construct a numerical method using a classical finite difference scheme on an appropriate Shishkin mesh, which resolves not only the usual boundary layers but also the interior layers arising from the delay term. More precisely, the singularly perturbed boundary value problem is

$$Lu(x) \;=\; -\varepsilon u''(x) + a(x)u(x) + b(x)u(x-1) = f(x) \;\text{ on } (0,2), \tag{1}$$

with

$$u \;=\; \phi \;\text{ on }\; [-1,0] \;\text{ and }\; u(2) = l, \tag{2}$$

where ϕ is sufficiently smooth on $[-1,0]$. For all $x \in [0,2]$, it is assumed that $a(x)$ and $b(x)$ satisfy

$$a(x) + b(x) > 2\alpha \tag{3}$$

and

$$b(x) < 0, \tag{4}$$

for some real number $\alpha > 0$. Furthermore, the functions $a(x), b(x)$, and $f(x)$ are assumed to be in $C^3([0,2])$.

The above assumptions ensure that $u \in \mathcal{C} = C^0([0,2]) \cap C^1((0,2)) \cap C^2((0,1) \cup (1,2))$.

The problem (1) and (2) can be rewritten as

$$L_1u(x) = -\varepsilon u''(x) + a(x)u(x) = f(x) - b(x)\phi(x-1) = g(x) \;\text{ on }\; (0,1) \tag{5}$$

$$L_2u(x) = -\varepsilon u''(x) + a(x)u(x) + b(x)u(x-1) = f(x) \;\text{ on }\; (1,2) \tag{6}$$

$$u = \phi \;\text{ on }\; [-1,0],\; u(1-) = u(1+),\; u'(1-) = u'(1+) \text{ and } u(2) = l. \tag{7}$$

The reduced problem corresponding to (1) and (2) is defined by

$$a(x)u_0(x) = g(x) \text{ on } (0,1) \tag{8}$$

$$a(x)u_0(x) + b(x)u_0(x-1) = f(x) \text{ on } (1,2). \tag{9}$$

In general as $u_0(x)$ need not satisfy $u_0(0) = u(0)$ and $u_0(2) = u(2)$, the solution $u(x)$ exhibits boundary layers at $x = 0$ and $x = 2$. In addition to that, at $x = 1$, $u_0(1-) = [f(1) - b(1)\phi(0-)]/a(1), u_0(1+) = [f(1) - b(1)u_0(0+)]/a(1)$, and as $u_0(1-)$ need not be equal to $u_0(1+)$, the solution $u(x)$ exhibits interior layers at $x = 1$.

For any function y on a domain D the following norm is introduced: $\parallel y \parallel_D = \sup_{x \in D} |y(x)|$. If $D = \overline{\Omega}$, the subscript is dropped. Throughout the paper C denotes a generic positive constant, which is independent of x and singular perturbation and discretization parameters.

The plan of the paper is as follows. In Sect. 2, the analytical results of the solution are presented. Improved estimates are presented in Sect. 3. In Sect. 4, piecewise-uniform Shishkin meshes are introduced and in Sect. 5, the discrete problem is defined and the discrete maximum principle and the discrete stability properties are established. In Sect. 6, numerical analysis is presented and the error bounds are established. In Sect. 7, numerical illustrations are presented.

2 Analytical Results

The operator L satisfies the following maximum principle.

Lemma 1. *Let $a(x)$ and $b(x)$ satisfy (3) and (4). Let ψ be in $\mathcal{C}$ such that $\psi(0) \geq 0$, $\psi(2) \geq 0$, $L\psi \geq 0$ on $(0,2)$ then $\psi \geq 0$ on $[0,2]$.*

Proof. Let x^* be such that $\psi(x^*) = \min_{x \in [0,2]} \psi(x)$. If $\psi(x^*) \geq 0$, there is nothing to prove. Suppose therefore that $\psi(x^*) < 0$. Then $x^* \notin \{0,2\}$. As $\psi''(x^*) \geq 0$,

$$\begin{aligned} L\psi(x^*) &= -\varepsilon\psi''(x^*) + a(x^*)\psi(x^*) + b(x^*)\psi(x^*-1) \\ &\leq -\varepsilon\psi''(x^*) + (a(x^*) + b(x^*))\psi(x^*) < 0, \text{ as } \psi(x^*-1) \geq \psi(x^*), \end{aligned}$$

which is a contradiction. This completes the proof. □

As a consequence of the maximum principle, there is established the stability result for the problem (1) and (2) in the following:

Lemma 2. *Let conditions (3) and (4) hold. If ψ is any function in $\mathcal{C}$, then for all $x \in [0,2]$,*

$$|\psi(x)| \leq \max\left\{|\psi(0)|, |\psi(2)|, \frac{1}{\alpha} \parallel L\psi \parallel\right\}.$$

Proof. Define the two functions:

$$\theta^{\pm}(x) \;=\; \max\left\{|\psi(0)|, |\psi(2)|, \frac{1}{\alpha} \parallel L\psi \parallel\right\} \;\pm\; \psi(x).$$

Using the properties of $a(x)$ and $b(x)$, it is not hard to verify that $\theta^{\pm}(x) \geq 0$ for $x \in \{0, 2\}$ and $L\theta^{\pm} \geq 0$ on $(0, 2)$. It follows from Lemma 1 that $\theta^{\pm} \geq 0$ on $[0, 2]$.

Standard estimates of the solution of (1) and (2) and its derivatives are contained in the following lemma. The proof is by the method of steps. □

Lemma 3. *Let conditions* (3) *and* (4) *hold and let u be the solution of* (1) *and* (2)*. Then, for all* $x \in [0, 2]$,

$$|u^{(k)}(x)| \leq C\,\varepsilon^{-\frac{k}{2}}(||u|| + ||f||),\ \text{for } k = 0, 1$$

and

$$|u^{(k)}(x)| \leq C\,\varepsilon^{-\frac{k}{2}}(||u|| + ||f|| + \varepsilon^{\frac{(k-2)}{2}}||f^{\,(k-2)}||),\ \ \text{for}\ \ k = 2, 3, 4.$$

Proof. The bound on u is an immediate consequence of Lemma 2 and the differential equation (1).

Rewriting the differential equation (1) gives

$$u''(x) = \varepsilon^{-1}(a(x)u(x) + b(x)u(x-1) - f(x)) \tag{10}$$

and it is not hard to see that the bound on u'' follow from (10).

To bound $u'(x)$, on the interval $(0, 1)$, consider an interval $N = [a, a + \sqrt{\varepsilon}] \subset [0, 1]$. By the mean value theorem, for some $y \in N$,

$$u'(y) = \frac{u(a + \sqrt{\varepsilon}) - u(a)}{\sqrt{\varepsilon}}$$

and it follows that $|u'(y)| \leq 2\varepsilon^{-\frac{1}{2}}||u||$. Now, for any $x \in N$

$$u'(x) = u'(y) + \int_y^x u''(s)ds = u'(y) + \varepsilon^{-1}\int_y^x (-f(s) + a(s)u(s) + b(s)\phi(s-1))ds$$

and so

$$|u'(x)| \leq |u'(y)| + C\varepsilon^{-1}(||f|| + ||u||)\int_y^x ds \leq C\varepsilon^{-\frac{1}{2}}(||f|| + ||u||)$$

from which the required bound follows.

To bound $u'(x)$, on the interval $(1,2)$, consider an interval $N = [a, a+\sqrt{\varepsilon}] \subset (1,2]$. By the mean value theorem, for some $y \in N$,

$$u'(y) = \frac{u(a+\sqrt{\varepsilon}) - u(a)}{\sqrt{\varepsilon}}$$

and it follows that $|u'(y)| \le 2\varepsilon^{-\frac{1}{2}}||u||$. Now, for any $x \in N$,

$$u'(x) = u'(y) + \int_y^x u''(s)ds = u'(y) + \varepsilon^{-1}\int_y^x (-f(s) + a(s)u(s)) + b(s)u(s-1))ds$$

and so

$$|u'(x)| \le |u'(y)| + C\varepsilon^{-1}(||f|| + ||u||)\int_y^x ds \le C\varepsilon^{-\frac{1}{2}}(||f|| + ||u||)$$

from which the required bound follows.

Differentiating (10) once and twice give $u^{(3)}(x) = \varepsilon^{-1}(a(x)u'(x) + a'(x)u(x) + b(x)u'(x-1) + b'(x)u(x-1) - f'(x))$, $u^{(4)}(x) = \varepsilon^{-1}(a(x)u''(x) + 2a'(x)u'(x) + a''(x)u(x) + b(x)u''(x-1) + 2b'(x)u'(x-1) + b''(x)u(x-1) - f''(x))$ and the bounds on $u^{(3)}$ and $u^{(4)}$ follow from those on u' and u''.

The Shishkin decomposition of the solution u of (1) and (2) is $u = v + w$ where the smooth component v is the solution of

$$L_1 v = g \;\text{ on } (0, 1-),\;\; v(0) = u_0(0),\;\; v(1-0) = (a(1))^{-1}(f(1) - b(1)\phi(0)) \quad (11)$$

$$L_2 v = f \;\text{ on } (1+, 2),\; v(1+0) = (a(1))^{-1}(f(1) - b(1)u_0(0)),\; v(2) = u_0(2) \quad (12)$$

and the singular component w is the solution of

$$\begin{gathered} L_1 w = 0 \;\text{ on } (0,1),\;\; L_2 w = 0 \;\text{ on } (1,2) \\ \text{with } w(0) = u(0) - v(0),\; [w](1) = -[v](1),\; [w'](1) = -[v'](1),\; w(2) = u(2) - v(2). \end{gathered} \quad (13)$$

The singular component is given a further decomposition

$$w(x) = w^L(x) + w^R(x) \quad (14)$$

with

$$w^L(x) = w(0)w_1^L(x) + Aw_2^L(x), \quad (15)$$

$$\text{satisfying } L_1 w_1^L(x) = 0, x \in (0,1) \text{ with } w_1^L(0) = 1, w_1^L(1) = 0, \quad (16)$$

$$w_1^L(x) = 0 \text{ on } (1,2], \quad (17)$$

$$L_2 w_2^L(x) = 0, x \in (1,2) \;\text{ with }\; w_2^L(1) = 1, w_2^L(2) = 0, \quad (18)$$

$$w_2^L(x) = 0 \text{ on } [0,1), \tag{19}$$

$$\text{and } \; w^R(x) = Bw_1^R(x) + w(2)w_2^R(x), \tag{20}$$

$$\text{satisfying } \; L_1 w_1^R(x) = 0, x \in (0,1) \text{ with } w_1^R(0) = 0, w_1^R(1) = 1, \tag{21}$$

$$w_1^R(x) = 0 \text{ on } (1,2], \tag{22}$$

$$L_2 w_2^R(x) = 0, x \in (1,2) \; \text{ with } \; w_2^R(1) = 0, w_2^R(2) = 1, \tag{23}$$

$$w_2^R(x) = 0 \text{ on } [0,1). \tag{24}$$

Here, A and B are constants to be chosen in such a way that the jump conditions at $x = 1$ are satisfied.

Bounds on the smooth component and its derivatives are contained in the following lemma. □

Lemma 4. *Let conditions* (3) *and* (4) *hold. Then the smooth component* v *and its derivatives satisfy, for all* $x \in [0,2]$,

$$|v^{(k)}(x)| \le C, \quad \text{for } \; k = 0,1,2$$

and

$$|v^{(k)}(x)| \le C(1 + \varepsilon^{1-\frac{k}{2}}), \quad \text{for } \; k = 3,4.$$

Proof. Decomposing the smooth component v as $v = u_0 + \varepsilon v_1$, it is not hard to see that v_1 satisfies a problem given by

$$L_1 v_1 = u_0'' \text{ on } (0,1-), \; v_1(0) = 0, \; v_1(1-0) = 0,$$

$$L_2 v_1 = u_0'' \text{ on } (1+,2), \; v_1(1+0) = 0, \; v_1(2) = 0.$$

We proceed by the method of steps. First consider $(0,1)$. On $(0,1), v_1$ satisfies a problem similar to (P_ε) of Chap. 6 in [7]. Hence on $(0,1), |v_1^{(k)}| \le C\,\varepsilon^{\frac{-k}{2}}$, for $k = 0,1,2,3$. From (8), $|u_0^{(k)}| \le C$, for $k = 0,1,2,3$. Hence on $(0,1), |v^{(k)}| \le C\,(1 + \varepsilon^{1-\frac{k}{2}})$, for $k = 0,1,2,3$.

The arguments used to bound v and its derivatives in the interval $(1,2)$ are given below.

The bound on v is an immediate consequence of the defining equation (12) for v and Lemma 2.

The bounds on v' and v'' are found as follows. Differentiating twice the equation (12) for v, it is not hard to see that v'' satisfies

$$\text{Ł}_1 v'' = h, \tag{25}$$

where

$$h(x) = f''(x) - 2a'(x)v'(x) - a''(x)v(x)$$
$$-2b'(x)v'(x-1) - b''(x)v(x-1) - b(x)v''(x-1).$$

Also the defining equation (12) for v yields

$$v''(1+) = 0, \quad v''(2) = 0. \tag{26}$$

Applying Lemma 1 on p. 39 of [7] to v'' and using the bounds of v, v' and v'' on $[0, 1]$ gives

$$||v''||_{[1,2]} \leq C(1 + ||v'||_{[1,2]}). \tag{27}$$

Choosing $x^* \in [1, 2]$ such that

$$v'(x^*) = ||v'||_{[1,2]} \tag{28}$$

and using a Taylor expansion it follows that for any $y \in [1 - x^*, 2 - x^*]$ and some η, such that $x^* < \eta < x^* + y$,

$$v(x^* + y) = v(x^*) + y\ v'(x^*) + \frac{y^2}{2}\ v''(\eta). \tag{29}$$

Rearranging (29) yields

$$v'(x^*) = \frac{v(x^* + y) - v(x^*)}{y} - \frac{y}{2}v''(\eta) \tag{30}$$

and so, from (28) and (30),

$$||v'||_{[1,2]} \leq \frac{2}{y}||v||_{[1,2]} + \frac{y}{2}||v''||_{[1,2]}. \tag{31}$$

Using (31) and (27) and the bound on v yields

$$\left(1 - \frac{Cy}{2}\right)||v''||_{[1,2]} \leq C\left(1 + \frac{2}{y}\right). \tag{32}$$

Choosing $y = \min(\frac{1}{C}, 2 - x^*)$, (32) then gives $||v''||_{[1,2]} \leq C$ and (31) gives $||v'||_{[1,2]} \leq C$ as required. The bounds on $v^{(3)}$, $v^{(4)}$ are derived by similar arguments.

The layer functions $B_1^L, B_1^R, B_2^L, B_2^R, B_1, B_2$, associated with the solution u, are defined by

$$B_1^L(x) = e^{-x\sqrt{\alpha}/\sqrt{\varepsilon}},\ B_1^R(x) = e^{-(1-x)\sqrt{\alpha}/\sqrt{\varepsilon}},\ B_1(x) = B_1^L(x) + B_1^R(x),\ \text{on } [0, 1],$$

$$B_2^L(x) = e^{-(x-1)\sqrt{\alpha}/\sqrt{\varepsilon}},\ B_2^R(x) = e^{-(2-x)\sqrt{\alpha}/\sqrt{\varepsilon}},\ B_2(x) = B_2^L(x) + B_2^R(x),\ \text{on } [1, 2].$$

Bounds on the singular components w^L and w^R of u and their derivatives are contained in the following lemma. □

Lemma 5. *Let conditions* (3) *and* (4) *hold. Then there exists a constant* C, *such that, for* $x \in [0, 1]$,

$$\left|w^{L,(k)}(x)\right| \leq C\,\frac{B_1^L(x)}{\varepsilon^{k/2}}, \text{ for } k = 0, 1, 2, 3$$

and, for $x \in [1, 2]$,

$$\left|w^{L,(k)}(x)\right| \leq C\,\frac{B_2^L(x)}{\varepsilon^{k/2}}, \text{ for } k = 0, 1, 2, 3.$$

Analogous results hold for w^R *and its derivatives.*

Proof. First we derive the bound on w^L on $(0, 1)$. On $[0, 1], w^L$ satisfies $L_1 w^L = 0$, $w^L(0) = w(0), w^L(1) = A$. Then following the procedure adapted in [7] it is not hard to find that

$$|w^{L,(k)}(x)| \leq C\varepsilon^{-\frac{k}{2}} B_1^L(x), x \in [0, 1].$$

We now derive the bound on w^L on $[1, 2]$. From the defining equation for w^L, we have

$$L_2 w^L(x) = -\varepsilon\, w^{L,\prime\prime}(x) + a(x)w^L(x) + b(x)w^L(x-1) = 0$$

or

$$\begin{aligned} L_1 w^L(x) &= -\varepsilon\, w^{L,\prime\prime}(x) + a(x)w^L(x) = -b(x)w^L(x-1) \\ &\Rightarrow |L_1 w^L(x)| \leq C\, B_1^L(x-1) = C B_2^L(x). \end{aligned} \tag{33}$$

Also, $w^L(1) = A$, $w^L(2) = 0$. Hence by using Lemma 1 on p. 39 of [7] for the operator L_1, on $[1, 2]$ leads to the required bound on w^L.

Consider the differential equation

$$L_2 w^L = 0, x \in (1, 2).$$

Then $w^{L,\prime\prime}(x) = \varepsilon^{-1}(a(x)w^L(x) + b(x)w^L(x-1))$. Hence

$$\begin{aligned} |w^{L,\prime\prime}(x)| &\leq C\varepsilon^{-1}(B_2^L(x) + B_1^L(x-1)) \\ &= C\varepsilon^{-1} B_2^L(x) \text{ since } B_1^L(x-1) = B_2^L(x) \text{ for } x \in [1, 2]. \end{aligned}$$

Using the mean value theorem and the bound of $w^L(x)$, arguments similar to those used to bound $u'(x)$ lead to the bound of $w^{L,\prime}(x)$. The bounds on $w^{L,\prime\prime\prime}(x)$ and $w^{L,(4)}(x)$ are derived similarly.

Analogous arguments lead to the estimates of w^R and its derivatives. □

3 Improved Estimates

In the following lemma, sharper estimates of the smooth component are presented.

Lemma 6. *Let conditions* (3) *and* (4) *hold. Then the smooth component v of the solution u of* (1) *and* (2) *satisfies, for $x \in [0,1)$,*

$$|v^{(k)}(x)| \leq C\,(1+B_1(x)) \ \text{ for } \ k=0,1,2 \ \text{ and } \ |v'''(x)| \ \leq \ C\ \left(1+\frac{B_1(x)}{\sqrt{\varepsilon}}\right)$$

and, for $x \in (1,2]$,

$$|v^{(k)}(x)| \leq C\,(1+B_2(x)), \ \text{ for } \ k=0,1,2 \ \text{ and } \ |v'''(x)| \ \leq \ C\ \left(1+\frac{B_2(x)}{\sqrt{\varepsilon}}\right).$$

Proof. Define barrier functions

$$\psi^{\pm}(x) \ = \ C(1+B_1(x)) \ \pm \ v^{(k)}(x), \ \ k=0,1,2 \ \text{ and } x \in (0,1).$$

Using Lemma 4, we find that $L\psi^{\pm} \ \geq \ 0$ on $(0,1)$ and $\psi^{\pm}(x) \ \geq \ 0$ at the points $x=0$ and $x=1-0$, for a proper choice of the constant C.

By using the maximum principle in [7] for the operator $L_1, \psi^{\pm} \ \geq \ 0$ on $[0,1)$. Thus, we conclude that for $k=0,1,2,$

$$|v^{(k)}(x)| \ \leq \ C(1+B_1(x)), \ \ x \in [0,1). \tag{34}$$

Consider (25) and (26), satisfied by v'' and note that $\parallel h' \parallel \leq C$, from Lemma 4.

For convenience let p denote v'' and then

$$L_1 p = h \ \text{ on } \ (0,1), \ \ p(x)=0 \ \text{ at } \ x=0 \ \text{ and } \ x=1-0. \tag{35}$$

Let z and r be the smooth and singular components of p satisfying

$$L_1 z = h \ \text{ on } \ (0,1), \ \ z(x)=h(x)/a(x) \ \text{ at } \ x=0 \ \text{ and } \ x=1-0.$$

and

$$L_1 r = 0 \ \text{ on } \ (0,1), \ \ r=-z \ \text{ at } \ x=0 \ \text{ and } \ x=1-0.$$

Using Lemmas 4 and 5 we have, for $x \in \ [0,1)$,

$$\begin{aligned} &|z'(x)| \leq C, \\ &|r'(x)| \leq C\ \tfrac{B_1(x)}{\sqrt{\varepsilon}}. \end{aligned}$$

Hence, for $x \in [0, 1)$,

$$|v'''(x)| = |p'(x)| \leq C \, \frac{(1 + B_1(x))}{\sqrt{\varepsilon}}. \tag{36}$$

From (34) and (36), for $k = 0, 1, 2, 3$ and $x \in [0, 1)$, the required results follow.

The bounds on v and its derivatives are similarly derived when $x \in (1, 2]$. □

4 The Shishkin Mesh

A piecewise-uniform Shishkin mesh with N mesh intervals is now constructed on $\overline{\Omega} = [0, 2]$ as follows. Let $\Omega^N = {\Omega_1}^N \cup {\Omega_2}^N$ where ${\Omega_1}^N = \{x_j\}_{j=1}^{\frac{N}{2}-1}$, ${\Omega_2}^N = \{x_j\}_{j=\frac{N}{2}+1}^{N-1}$ and $x_{\frac{N}{2}} = 1$. Then $\overline{\Omega_1}^N = \{x_j\}_{j=0}^{\frac{N}{2}}$, $\overline{\Omega_2}^N = \{x_j\}_{j=\frac{N}{2}}^{N}$, $\overline{\Omega_1}^N \cup \overline{\Omega_2}^N = \overline{\Omega}^N = \{x_j\}_{j=0}^{N}$ and $\Gamma^N = \{0, 2\}$. The interval $[0, 1]$ is divided into three subintervals as follows:

$$[0, \tau] \cup (\tau, 1 - \tau] \cup (1 - \tau, 1].$$

The parameter τ, which determine the points separating the uniform meshes, is defined by

$$\tau = \min \left\{ \frac{1}{4}, \sqrt{\varepsilon/\alpha} \ln N \right\}. \tag{37}$$

Then, on the subinterval $(\tau, 1 - \tau]$, a uniform mesh with $\frac{N}{4}$ mesh points is placed and on each of the subintervals $[0, \tau]$ and $(1 - \tau, 1]$, a uniform mesh of $\frac{N}{8}$ mesh points is placed.

Similarly, the interval $(1, 2]$ is also divided into 3 subintervals $(1, 1 + \tau]$, $(1 + \tau, 2 - \tau]$ $,and$ $(2 - \tau, 2]$, using the same parameter τ. In particular, when the parameter τ takes on its left-hand value, the Shishkin mesh $\overline{\Omega}^N$ becomes a classical uniform mesh throughout from 0 to 2.

In practice, it is convenient to take

$$N = 8k, \;\; k \geq 2. \tag{38}$$

From the above construction of $\overline{\Omega_1}^N$, it is clear that the transition points $\{\tau, 1 - \tau\}$ are the only points at which the mesh size can change and that it does not necessarily change at each of these points. The following notations are introduced: $h_j = x_j - x_{j-1}, h_{j+1} = x_{j+1} - x_j$, and if $x_j = \tau$, then $h_j^- = x_j - x_{j-1}$, $h_j^+ = x_{j+1} - x_j$, $J = \{x_j : h_j^+ \neq h_j^-\}$. For each point x_j in the mesh intervals $[0, \tau]$ and $(1 - \tau, 1]$,

$$x_j - x_{j-1} = 8\, N^{-1}\tau, \tag{39}$$

and for $x_j \in (\tau, 1 - \tau]$, $\;x_j - x_{j-1} = 4N^{-1}(1 - 2\tau)$.

5 The Discrete Problem

In this section, a classical finite difference operator with an appropriate Shishkin mesh is used to construct a numerical method for (1) and (2) which is shown later to be essentially first-order parameter-uniform convergent.

The discrete two-point boundary value problem is now defined to be

$$\begin{aligned} L^N U(x_j) &= -\varepsilon\delta^2 U(x_j) + a(x_j)U(x_j) + b(x_j)U(x_j-1) = f(x_j) \ \text{ on } \Omega^N, \\ U &= u \ \text{ on } \Gamma^N. \end{aligned} \tag{40}$$

The problem (40) can be rewritten as

$$\begin{aligned} L_1^N U(x_j) &= -\varepsilon\delta^2 U(x_j) + a(x_j)U(x_j) = g(x_j) \ \text{ on } \Omega^{-N}, \\ L_2^N U(x_j) &= -\varepsilon\delta^2 U(x_j) + a(x_j)U(x_j) + b(x_j)U(x_j-1) = f(x_j) \ \text{ on } \Omega^{+N}, \\ U &= u \ \text{ on } \Gamma^N, \\ D^- U(x_{N/2}) &= D^+ U(x_{N/2}). \end{aligned} \tag{41}$$

This is used to compute numerical approximations to the solution of (1) and (2). The following discrete results are analogous to those for the continuous case.

Lemma 7. *Let conditions* (3) *and* (4) *hold. Then, for any mesh function* Ψ*, the inequalities* $\Psi \geq 0$ *on* Γ^N, $L_1^N\Psi \geq 0$ *on* Ω_1^N, $L_2^N\Psi \geq 0$ *on* Ω_2^N*, and* $D^+\Psi(x_{N/2}) - D^-\Psi(x_{N/2}) \leq 0$ *imply that* $\Psi \geq 0$ *on* $\overline{\Omega}^N$.

Proof. Let j^* be such that $\Psi(x_{j^*}) = \min\limits_j \Psi(x_j)$ and assume that the lemma is false. Then $\Psi(x_{j^*}) < 0$. From the hypotheses we have $j^* \neq 0,\ N$.

Suppose $x_{j^*} \in {\Omega_1}^N$. $\Psi(x_{j^*}) - \Psi(x_{j^*-1}) \leq 0,\ \Psi(x_{j^*+1}) - \Psi(x_{j^*}) \geq 0$, so $\delta^2\Psi(x_{j^*}) \geq 0$. It follows that

$$L_1^N\Psi(x_{j^*}) \ = \ -\varepsilon\delta^2\Psi(x_{j^*}) + a(x_{j^*})\Psi(x_{j^*}) < 0,$$

which is a contradiction. If $x_{j^*} \in {\Omega_2}^N$, a similar argument shows that

$$L_2^N\Psi(x_{j^*}) = -\varepsilon\delta^2\Psi(x_{j^*}) + a(x_{j^*})\Psi(x_{j^*}) + b(x_{j^*})\Psi(x_{j^*}-1) < 0,$$

which is a contradiction. Finally if $x_{j^*} = x_{N/2}$, then

$$D^-\Psi(x_{N/2}) \leq 0 \leq D^+\Psi(x_{N/2}) \leq D^-\Psi(x_{N/2}), \text{ by the hypothesis}$$

and so

$$\Psi(x_{\frac{N}{2}-1}) = \Psi(x_{N/2}) = \Psi(x_{\frac{N}{2}+1}) < 0.$$

Then $L_1^N\Psi(x_{\frac{N}{2}-1}) < 0$, a contradiction. This concludes the proof of the lemma.

An immediate consequence of this is the following discrete stability result. □

Lemma 8. *Let conditions* (3) *and* (4) *hold. Then, for any mesh function* Ψ*,*

$$|\Psi(x_j)| \leq \max\left\{|\Psi(x_0)|, |\Psi(x_N)|, \frac{1}{\alpha}||L_1^N\Psi||_{\Omega_1^N}, \frac{1}{\alpha}||L_2^N\Psi||_{\Omega_2^N}\right\}, \quad 0 \leq j \leq N.$$

Proof. Consider the barrier functions

$$\Theta^{\pm}(x_j) = \max\{|\Psi(x_0)|, |\Psi(x_N)|, \frac{1}{\alpha}||L_1^N\Psi||_{\Omega_1^N}, \frac{1}{\alpha}||L_2^N\Psi||_{\Omega_2^N}\}\pm\Psi(x_j),\ 0 \leq j \leq N.$$

Using the properties of $a(x)$ and $b(x)$, it is not hard to find that $\Theta^{\pm}(x_j) \geq 0$ for $j = 0, N$,$L_1^N\Theta^{\pm}(x_j) \geq 0$ for $x_j \in \Omega_1^N$, and $L_2^N\Theta^{\pm}(x_j) \geq 0$ for $x_j \in \Omega_2^N$. At $j = \frac{N}{2}$,

$$D^+\Theta^{\pm}(x_{N/2}) - D^-\Theta^{\pm}(x_{N/2}) = D^+\Psi(x_{N/2}) - D^-\Psi(x_{N/2}) = 0.$$

Hence by Lemma 7, $\Theta^{\pm} \geq 0$ on $\overline{\Omega}^N$, which leads to the required result. □

6 Error Estimate

Analogous to the continuous case, the discrete solution U can be decomposed into V and W which are defined to be the solutions of the following discrete problems:

$$L_1^N V(x_j) = g(x_j),\ x_j \in {\Omega_1}^N,\ V(0) = v(0),\ V(x_{N/2-1}) = v(1-),$$

$$L_2^N V(x_j) = f(x_j),\ x_j \in {\Omega_2}^N,\ V(x_{N/2+1}) = v(1+),\ V(2) = v(2)$$

and

$$L_1^N W(x_j) = 0,\ x_j \in {\Omega_1}^N,\ W(0) = w(0),$$

$$L_2^N W(x_j) = 0,\ x_j \in {\Omega_2}^N,\ W(2) = w(2),$$

$$V(x_{N/2+1}) + W(x_{N/2+1}) = V(x_{N/2-1}) + W(x_{N/2-1}),$$

$$D^-W(x_{N/2}) + D^-V(x_{N/2}) = D^+W(x_{N/2}) + D^+V(x_{N/2}).$$

The error at each point $x_j \in \overline{\Omega}^N$ is denoted by $e(x_j) = U(x_j) - u(x_j)$. Then the local truncation error $L^N e(x_j)$, for $j \neq N/2$, has the decomposition

$$L^N e(x_j) = L^N (V - v)(x_j) + L^N (W - w)(x_j).$$

The errors in the smooth and singular components are bounded in the following theorem.

Theorem 1. *Let conditions* (3) *and* (4) *hold. If v denotes the smooth component of the solution of* (1) *and* (2) *and V the smooth component of the solution of the problem* (41)*, then, for* $j \neq N/2$,

$$|L_1^N (V - v)(x_j)| \leq C\, N^{-1}, \quad 0 \leq j \leq \frac{N}{2} - 1, \tag{42}$$

$$|L_2^N (V - v)(x_j)| \leq C\, N^{-1}, \quad \frac{N}{2} + 1 \leq j \leq N. \tag{43}$$

If w denotes the singular component of the solution of (1) *and* (2) *and W the singular component of the solution of the problem* (41)*, then, for* $j \neq N/2$,

$$|L_1^N (W - w)(x_j)| \leq C\, N^{-1} \ln N, \quad 0 \leq j \leq \frac{N}{2} - 1, \tag{44}$$

$$|L_2^N (W - w)(x_j)| \leq C\, N^{-1} \ln N, \quad \frac{N}{2} + 1 \leq j \leq N. \tag{45}$$

Proof. As the expression derived for the local truncation error in V and W and estimates for the derivatives of the smooth and singular components are exactly in the form found in Chap. 6 of [7], the required bounds hold good.

At the point $x_j = x_{N/2}$,

$$\begin{aligned}(D^+ - D^-)e(x_{N/2}) &= (D^+ - D^-)(U - u)(x_{N/2}) \\ &= (D^+ - D^-)U(x_{N/2}) - (D^+ - D^-)u(x_{N/2}).\end{aligned}$$

Recall that $(D^+ - D^-)U(x_{N/2}) = 0$. Let $h^* = h^-_{N/2} = h^+_{N/2}$, where $h^-_{N/2} = x_{N/2} - x_{N/2-1}$ and $h^+_{N/2} = x_{N/2+1} - x_{N/2}$.

Then

$$\begin{aligned}|(D^+ - D^-)e(x_{N/2})| &= |(D^+ - D^-)u(x_{N/2})| \\ &\leq \left|\left(D^+ - \frac{d}{dx}\right)u(x_{N/2})\right| + \left|\left(D^- - \frac{d}{dx}\right)u(x_{N/2})\right| \\ &\leq \frac{1}{2}h^+_{N/2} \max_{\eta_1 \in (1,2)} |u''(\eta_1)| + \frac{1}{2}h^-_{N/2} \max_{\eta_2 \in (0,1)} |u''(\eta_2)| \\ &\leq C\, h^* \max_{x \in (0,1) \cup (1,2)} \left|u''(x)\right|.\end{aligned}$$

Therefore,

$$|(D^+ - D^-)e(x_{N/2})| \ \le \ C \ \frac{h^*}{\varepsilon}. \tag{46}$$

Define a set of discrete barrier functions on $\overline{\Omega}^N$ by

$$\omega(x_j) = \begin{cases} \dfrac{\Pi_{k=1}^{j}(1 + \sqrt{\alpha/\varepsilon}\, h_k)}{\Pi_{k=1}^{N/2}(1 + \sqrt{\alpha/\varepsilon}\, h_k)}, & 0 \le j \le N/2 \\ \\ \dfrac{\Pi_{k=j}^{N-1}(1 + \sqrt{\alpha/\varepsilon}\, h_{k+1})}{\Pi_{k=N/2}^{N-1}(1 + \sqrt{\alpha/\varepsilon}\, h_{k+1})}, & N/2 \le j \le N. \end{cases} \tag{47}$$

Note that

$$\omega(0) = 0, \ \omega(1) = 1, \ \omega(2) = 0 \tag{48}$$

and from (47), for $0 \le j \le N$,

$$0 \le \omega(x_j) \le 1. \tag{49}$$

For $x_j \in \overline{\Omega_1}^N$,

$$\begin{aligned} D^+\omega(x_j) &= \frac{\omega(x_{j+1}) - \omega(x_j)}{h_{j+1}} \\ &= \frac{1}{h_{j+1}} \ \frac{\Pi_{k=1}^{j}(1 + \sqrt{\alpha/\varepsilon}\, h_k)}{\Pi_{k=1}^{N/2}(1 + \sqrt{\alpha/\varepsilon}\, h_k)} \ \left(1 + \sqrt{\alpha/\varepsilon}\, h_{j+1} - 1\right) \\ &= \sqrt{\alpha/\varepsilon}\omega(x_j). \end{aligned}$$

Therefore,

$$D^+\omega(x_j) = \sqrt{\alpha/\varepsilon}\, \omega(x_j). \tag{50}$$

$$\begin{aligned} D^-\omega(x_j) &= \frac{\omega(x_j) - \omega(x_{j-1})}{h_j} \\ &= \frac{1}{h_j} \ \frac{\Pi_{k=1}^{j}(1 + \sqrt{\alpha/\varepsilon}\, h_k)}{\Pi_{k=1}^{N/2}(1 + \sqrt{\alpha/\varepsilon}\, h_k)} \ \left(1 - \frac{1}{1 + \sqrt{\alpha/\varepsilon}\, h_j}\right) \\ &= \sqrt{\alpha/\varepsilon}\frac{1}{(1 + \sqrt{\alpha/\varepsilon}\, h_j)} \ \omega(x_j). \end{aligned}$$

Therefore,

$$D^{-}\omega(x_j) = \sqrt{\alpha/\varepsilon}\frac{1}{(1+\sqrt{\alpha/\varepsilon}\,h_j)}\,\omega(x_j). \tag{51}$$

$$\begin{aligned}
\delta^2\omega(x_j) &= \frac{D^{+}\omega(x_j) - D^{-}\omega(x_j)}{(h_j + h_{j+1})/2}\\
&= \frac{1}{(h_j+h_{j+1})/2}\left(\sqrt{\frac{\alpha}{\varepsilon}} - \sqrt{\frac{\alpha}{\varepsilon}}\frac{1}{(1+\sqrt{\alpha/\varepsilon}\,h_j)}\right)\omega(x_j)\\
&= \frac{2\alpha}{\varepsilon}\frac{h_j}{(h_j+h_{j+1})}\left(\frac{1}{1+\sqrt{\alpha/\varepsilon}\,h_j}\right)\omega(x_j)\\
&\le \frac{2\alpha}{\varepsilon}\,\omega(x_j).
\end{aligned}$$

Therefore,

$$\delta^2\omega(x_j) \;\le\; \frac{2\alpha}{\varepsilon}\,\omega(x_j). \tag{52}$$

Similarly, for $x_j \in \overline{\Omega_2}^N$,

$$D^{+}\omega(x_j) = -\sqrt{\alpha/\varepsilon}\frac{1}{(1+\sqrt{\alpha/\varepsilon}\,h_{j+1})}\,\omega(x_j), \qquad D^{-}\omega(x_j) = -\sqrt{\alpha/\varepsilon}\omega(x_j)$$

$$\text{and} \qquad \delta^2\omega(x_j) \;\le\; \frac{2\alpha}{\varepsilon}\,\omega(x_j). \tag{53}$$

In particular, at $x_j = x_{N/2}$, using (53), (51), and (48),

$$\begin{aligned}
(D^{+} - D^{-})\omega(x_j) &= -\sqrt{\alpha/\varepsilon}\frac{1}{(1+\sqrt{\alpha/\varepsilon}\,h^{+}_{N/2})} - \sqrt{\alpha/\varepsilon}\frac{1}{(1+\sqrt{\alpha/\varepsilon}\,h^{-}_{N/2})}\\
&\le -\frac{C}{\sqrt{\varepsilon}}.
\end{aligned} \tag{54}$$

From (52) and (53),

$$-\varepsilon\,\delta^2\omega(x_j) \;\ge\; -2\alpha\;\omega(x_j).$$

Therefore

$$\begin{aligned} L_1^N \omega(x_j) &= -\varepsilon\, \delta^2 \omega(x_j) + a(x_j)\omega(x_j) \\ &\geq -2\alpha\ \omega(x_j)\ + a(x_j)\omega(x_j) \\ &= (a(x_j) - 2\alpha)\ \omega(x_j). \end{aligned} \tag{55}$$

and

$$\begin{aligned} L_2^N \omega(x_j) &= -\varepsilon\, \delta^2 \omega(x_j) + a(x_j)\omega(x_j) + b(x_j)\omega(x_j - 1) \\ &\geq -2\alpha\ \omega(x_j)\ + a(x_j)\omega(x_j) + b(x_j) \\ &= (a(x_j) - 2\alpha)\ \omega(x_j) + b(x_j). \end{aligned} \tag{56}$$

We now state and prove the main theoretical result of this paper. □

Theorem 2. *Let $u(x_j)$ be the solution of the problem* (1) *and* (2) *and $U(x_j)$ be the solution of the problem* (41). *Then, for* $0 \leq j \leq N$,

$$|(U - u)(x_j)| \leq C\, N^{-1} \ln N.$$

Proof. Consider the mesh function Ψ given by

$$\Psi(x_j) = C_1 N^{-1} \ln N + C_2 \sqrt{\alpha/\varepsilon}\ h^* \omega(x_j) \pm e(x_j), \quad 0 \leq j \leq N,$$

where C_1 and C_2 are constants. Then,

$$L_1^N \Psi(x_j)\ =\ C_1 a(x_j) N^{-1} \ln N + C_2 \sqrt{\alpha/\varepsilon}\ h^* L_1^N \omega(x_j) \pm L_1^N e(x_j). \tag{57}$$

Using (55) in (57) and Theorem 1,

$$L_1^N \Psi(x_j) \geq C_1 a(x_j) N^{-1} \ln N + C_2 \sqrt{\alpha/\varepsilon}\ h^* (a(x_j) - 2\alpha)\omega(x_j) \pm C\, N^{-1} \ln N \geq 0,$$

for appropriate choices of C_1 and C_2. For $x_j \in {\Omega_2}^N$,

$$L_2^N \Psi(x_j)\ =\ C_1(a(x_j) + b(x_j)) N^{-1} \ln N + C_2 \sqrt{\alpha/\varepsilon}\ h^* L_2^N \omega(x_j) \pm L_2^N e(x_j). \tag{58}$$

Using (56) in (58),

$$\begin{aligned} L_2^N \Psi(x_j) \geq & C_1 (a(x_j) + b(x_j)) N^{-1} \ln N \\ & + C_2 \sqrt{\alpha/\varepsilon}\ h^* ((a(x_j) - 2\alpha)\ \omega(x_j) + b(x_j)) \pm C\, N^{-1} \ln N. \end{aligned}$$

Let $\lambda(x_j) = (a(x_j) - 2\alpha)\omega(x_j) + b(x_j))$. Then choosing $C_1 > \dfrac{C_2 ||\lambda||}{2\alpha} + C$, and Theorem 1, $L_2^N \Psi(x_j) \geq 0$.

Further,

$$\begin{aligned} D^{+}\Psi(1) - D^{-}\Psi(1) &\leq -C_2\frac{Ch^*}{\varepsilon} \pm C\frac{h^*}{\varepsilon}, \text{ using (46) and (54)} \\ &\leq 0, \text{ for proper choice of } C_2. \end{aligned}$$

Also, using (48), $\Psi(0) = C_1 N^{-1}\ln N \geq 0$, $\Psi(2) = C_1 N^{-1}\ln N \geq 0$.

Therefore, using Lemma 7 for Ψ, it follows that $\Psi(x_j) \geq 0$ for all $0 \leq j \leq N$. As, from (49), $\omega(x_j) \leq 1$ for $0 \leq j \leq N$,

$$|(U - u)(x_j)| \leq CN^{-1}\ln N,$$

which completes the proof. □

7 Numerical Illustrations

The ε-uniform convergence of the numerical method proposed in this paper is illustrated through two examples presented in this section.

Example 1. Consider the BVP

$$-\varepsilon u''(x) + 2u(x) - u(x-1) = 0, \text{ for } x \in (0,1) \cup (1,2),$$

$$u(x) = 1 \text{ for } x \in [-1,0],\ u(2) = 1.$$

The maximum pointwise errors and the rate of convergence for this BVP are presented in Table 1.

Table 1 Values of D_ε^N, D^N, p^N, p^* and $C_{p^*}^N$ for $\alpha = 0.9$

	Number of mesh points N				
ε	64	128	256	512	1,024
2^0	0.252E−02	0.128E−02	0.644E−03	0.323E−03	0.162E−03
2^{-3}	0.741E−02	0.385E−02	0.195E−02	0.976E−03	0.488E−03
2^{-6}	0.880E−02	0.577E−02	0.348E−02	0.201E−02	0.114E−02
2^{-9}	0.881E−02	0.577E−02	0.348E−02	0.201E−02	0.113E−02
2^{-12}	0.881E−02	0.577E−02	0.348E−02	0.201E−02	0.113E−02
D^N	0.881E−02	0.577E−02	0.348E−02	0.201E−02	0.114E−02
p^N	0.610E+00	0.732E+00	0.792E+00	0.817E+00	
C_p^N	0.323E+00	0.323E+00	0.297E+00	0.261E+00	0.226E+00

Computed order of ε-uniform convergence, $p^* = 0.6097$
Computed ε-uniform error constant, $C_{p^*}^N = 0.3227$

Table 2 Values of D_ε^N, D^N, p^N, p^*, and $C_{p^*}^N$ for $\alpha = 0.9$

	Number of mesh points N				
ε	64	128	256	512	1,024
2^0	0.202E−02	0.102E−02	0.511E−03	0.256E−03	0.128E−03
2^{-3}	0.577E−02	0.297E−02	0.149E−02	0.743E−03	0.371E−03
2^{-6}	0.679E−02	0.442E−02	0.264E−02	0.151E−02	0.854E−03
2^{-9}	0.673E−02	0.438E−02	0.261E−02	0.150E−02	0.842E−03
2^{-12}	0.670E−02	0.437E−02	0.261E−02	0.149E−02	0.839E−03
D^N	0.679E−02	0.442E−02	0.264E−02	0.151E−02	0.854E−03
p^N	0.618E+00	0.745E+00	0.803E+00	0.824E+00	
C_p^N	0.255E+00	0.255E+00	0.233E+00	0.205E+00	0.178E+00

Computed order of ε-uniform convergence, $p^* = 0.6184$
Computed ε-uniform error constant, $C_{p^*}^N = 0.2548$

Example 2. Consider the BVP

$$-\varepsilon u''(x) + (2+x)u(x) - u(x-1) = 0, \text{ for } x \in (0,1) \cup (1,2),$$

$$u(x) = 1 \text{ for } x \in [-1,0],\ u(2) = 1.$$

The maximum pointwise errors and the rate of convergence for this BVP are presented in Table 2.

Acknowledgements The authors express their sincere thanks to Dr. N. Ramanujam, UGC Emeritus Fellow, Bharathidasan University, Tiruchirappalli, for the useful discussions on delay differential equations.

References

1. Lange, C.G., Miura, R.M.: Singular perturbation analysis of boundary-value problems for differential - difference equations. SIAM J. Appl. Math. **42**(3), 502–530 (1982)
2. Cen, Z.: A hybrid finite difference scheme for a class of singularly perturbed delay differential equations. Neural, Parallel Sci. Comput. **16**, 303–308 (2008)
3. Hongjiong, T.: Numerical methods for singularly perturbed delay differential equations. In: Proceedings of the International Conference on Boundary and Interior Layers, Computational and Asymptotic Methods - BAIL 2004. ONERA, Toulouse (2004)
4. Nicaise, S., Xenophontos, C.: Robust approximation of singularly perturbed delay differential equations by the hp finite element method. Comput. Meth. Appl. Math. **13**(1), 21–37 (2013)
5. Subburayan, V., Ramanujam, N.: An initial value technique for singularly perturbed reaction diffusion problems with a negative shift. Novi Sad J. Math. **43**(2), 67–80 (2013)
6. Subburayan, V., Ramanujam, N.: An initial value technique for singularly perturbed system of reaction-diffusion type delay differential equations. J. KSIAM. **17**(4), 221–237 (2013)
7. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems. World Scientific Publishing Co., Singapore, New Jersey, London, Hong Kong (1996)

Optimal L^∞-Error Estimate for a System of Elliptic Quasi-Variational Inequalities with Noncoercive Operators

M. Boulbrachene

Abstract This paper deals with the standard finite element approximation of a noncoercive system of quasi-variational inequalities (QVIs) arising in stochastic control problems. We improve a result obtained in Boulbrachene (Comput. Math. Appl. 45, 983–989, 2003) and establish the optimal L^∞ convergence order making use of the concepts of subsolutions and discrete regularity.

Keywords System of quasi-variational inequalities • Finite elements • Subsolution • Discrete regularity • L^∞-error estimate

1 Introduction

We are concerned with the standard finite element approximation in the L^∞ norm of the noncoercive problem associated with the system of quasi-variational inequalities (QVIs): find $U = (u^1, \dots, u^M) \in (H_0^1(\Omega))^M$ such that

$$\begin{cases} a^i(u^i, v - u^i) \geqq (f^i, v - u^i) \ \forall v \in H_0^1(\Omega) \\ u^i \leq k + u^{i+1}, \ v \leq k + u^{i+1} \\ u^{M+1} = u^1 \end{cases} . \tag{1}$$

This system appears in stochastic control problems with switching [1, 2]. Here Ω is a bounded domain of $\mathbb{R}^N$, $N \geq 1$, with smooth boundary Γ, $(.,.)$ denotes the

This paper is not in final form and a more detailed version of it will be submitted for publication elsewhere.

M. Boulbrachene (✉)
Department of Mathematics and Statistics, Sultan Qaboos University,
P.O. Box 36, Muscat 123, Sultanate of Oman
e-mail: boulbrac@squ.edu.om

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_8,

standard inner product in $L^2(\Omega)$, f^i are positive functions in $W^{2,\infty}(\Omega)$, k is a positive number, and $a^i(u,v)$ are M bilinear forms

$$a^i(u,v) = \int_\Omega \left(\sum_{1\leq j,k\leq N} a^i_{jk}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_k} + \sum_{k=1}^{N} b^i_k(x)\frac{\partial u}{\partial x_k}v + a^i_0(x)uv \right) dx$$

assumed to be noncoercive. The coefficients $a^i_{jk}(x)$, $b^i_k(x), a^i_0(x)$ are in $C^2(\bar{\Omega})$, $x \in \bar{\Omega}$, and satisfy $a^i_0(x) \geqq c_0 \geq 0$, $(x \in \bar{\Omega};\ c_0 > 0)$, and $\sum_{1\leq j,k\leq N} a^i_{jk}(x)\xi_j\xi_k \geqq \alpha\,|\xi|^2$; $(x \in \bar{\Omega},\ \xi \in R^N,\ \alpha > 0)$.

Componentwise, u^i can be regarded as the solution of the variational inequalities (VI) with source term f^i and obstacle $k + u^{i+1}$. Let us then adopt the notation $u^i = \sigma(f^i, k + u^{i+1})$, $i = 1, 2\ldots, M$ for system (1).

Now, let Ω be decomposed into triangles and let τ_h denote the set of all those elements; $h > 0$ is the mesh size.We assume that the family τ_h is regular and quasi-uniform. Let also $\mathbb{V}_h$ be the finite element space consisting of continuous piecewise linear functions vanishing on Γ, and $\{\varphi_s\}$, $s = 1, 2, \ldots, m(h)$ the basis of $\mathbb{V}_h$. The discrete counterpart of (1) consists of seeking $(u^1_h, \ldots, u^M_h) \in (\mathbb{V}_h)^M$ such that

$$\begin{cases} a^i(u^i_h, v - u^i_h) \geqq (f^i, v - u^i_h)\ \forall v \in \mathbb{V}_h \\ v \leq k + u^{i+1}_h,\ u^i_h \leq k + u^{i+1}_h \\ u^{M+1}_h = u^1_h \end{cases}, \tag{2}$$

where, componentwise, $u^i_h = \sigma_h(f^i, k + u^{i+1}_h)$ denotes the solution of the discrete VI with source term f^i and obstacle $k + u^{i+1}_h$.

The finite element approximation of the coercive problem was carried out in [3] and optimal error estimate in the L^∞ norm was obtained. Regarding the noncoercive problem, a quasi-optimal error estimate has been derived in [4], that is,

$$\max_{1\leq i\leq M} \| u^i - u^i_h \|_\infty \leq Ch^2\,|\log h|^3\,.$$

In the present paper, we get rid of the "extra-$|\log h|$ factor" and establish the optimal convergence order:

$$\max_{1\leq i\leq M} \| u^i - u^i_h \|_\infty \leq Ch^2\,|\log h|^2\,.$$

For that, we shall employ the concepts of subsolutions and "discrete regularity." More precisely, we use the characterization the continuous solution (resp. the discrete solution) as the maximum elements of the set of continuous subsolutions (resp. the set of discrete subsolutions). The so-called discrete regularity plays a crucial role in deriving the optimal order as it permits to replace the nonsmooth obstacle functions $k + u^{i+1}_h$ appearing in problem (2) with functions in $W^{2,p}(\Omega)$.

2 Background

In order to deal with the noncoercive problem, we consider the equivalent formulation: find $(u^1,\ldots,u^M) \in (H_0^1(\Omega))^M$ such that

$$\begin{cases} b^i(u^i, v - u^i) \geqq (f^i + \lambda u^i, v - u^i) \; \forall v \in H_0^1(\Omega) \\ u^i \leq k + u^{i+1}, \; v \leq k + u^{i+1} \\ u^{M+1} = u^1 \end{cases}, \tag{3}$$

where $b^i(u, v) = a^i(u, v) + \lambda(u, v)$, and $\lambda > 0$ is large enough such that $b^i(.,.)$ are strongly coercive on $H^1(\Omega)$, i.e, $b^i(v, v) \geq \gamma \, \|v\|^2_{H^1(\Omega)}$, $\gamma > 0$.

System (1) or (3) has a unique solution which belongs to $(W^{2,p}(\Omega))^M$, $1 \leq p < \infty$ [2]. Below, we give some useful qualitative properties enjoyed by the solutions of system (1) and (3) respectively. These properties are greatly needed in the proof of the main result.

Notation 1. *Let $k, \tilde{k}$ be two positive constants and let $(f^1,\ldots,f^M)$, $(\tilde{f}^1,\ldots,f^{\tilde{M}})$ be two families of source terms.We denote by $u^i = \sigma(f^i, k + u^{i+1})$ and $\tilde{u}^i = \sigma(\tilde{f}^i, k + \tilde{u}^{i+1})$, $i = 1, 2\ldots, M$, the corresponding solutions to system (1).*

For the sake of simplicity, we shall adopt, as in [3], the notation $u^i = \sigma(f^i, k)$ instead of $u^i = \sigma(f^i, k + u^{i+1})$.

Theorem 1 (Continuous Lipschitz dependence). *Let C be a constant such that $Ca_0^i(x) > 1$. Then, we have*

$$\max_{1 \leq i \leq M} \| \, u^i - \tilde{u}^i \, \|_{\infty} \leq C \left(\left| k - \tilde{k} \right| + \left\| f^i - \tilde{f}^i \right\|_{\infty} \right).$$

Proof. We adapt [3]. □

Definition 1 (Continuous subsolution). $(w^1,\ldots,w^M) \in (H_0^1(\Omega))^M$ is said to be a subsolution for the system of QVIs (1) if

$$\begin{cases} b^i(w^i, v) \; \leq \; (f^i + \lambda w^i, v) \;\; \forall v \in H_0^1(\Omega), \; v \geq 0 \\ w^i \leq k + w^{i+1}, \; v \leq k + w^{i+1} \\ w^{M+1} = w^1 \end{cases}. \tag{4}$$

Theorem 2 ([4]). *Let $\mathbb{X}$ denote the set of such subsolutions. Then, the solution of system of QVIs (1) is the maximum element of the set $\mathbb{X}$.*

As in the continuous situation, one can tackle the discrete problem by considering the equivalent formulation

$$\begin{cases} b^i(u_h^i, v - u_h^i) \geqq (f^i + \lambda u_h^i, v - u_h^i) \; \forall v \in \mathbb{V}_h \\ v \leq k + u_h^{i+1}, \; u_h^i \leq k + u_h^{i+1} \\ u_h^{M+1} = u_h^1 \end{cases}. \tag{5}$$

For that, we need the discrete maximum principle *(d.m.p)*. In other words, we assume that the matrices with generic coefficients $b^i\ (\varphi_l, \varphi_s)\ \forall i = 1, 2, \ldots, M$, are M-Matrices [5,6]. Under this d.mp, system (2) or (5) has a unique solution [4]. Furthermore, as in the continuous case, we have the discrete analogs of Theorems 1 and 2.

Theorem 3 (Discrete Lipschitz dependence). *Let* $u_h^i = \sigma_h(f^i, k)$, $\tilde{u}_h^i = \sigma_h(\tilde{f}^i, \tilde{k})$*, and let* C *be a constant such that* $Ca_0^i(x) > 1$*. Then, under the d.m.p, we have*

$$\max_{1 \le i \le M} \parallel u_h^i - \tilde{u}_h^i \parallel_\infty \le C\left(\left|k - \tilde{k}\right| + \left\|f - \tilde{f}\right\|_\infty\right).$$

Definition 2 (Discrete subsolution). $\left(w_h^i, \ldots, w_h^M\right) \in (\mathbb{V}_h)^M$ is said to be a *subsolution* for the system (5) if

$$\begin{cases} b^i(w_h^i, \varphi_s) \le (f^i + \lambda w_h^i, \varphi_s)\ \ \forall \varphi_s;\ s = 1, \ldots, m(h) \\ w_h^i \le k + w_h^{i+1} \\ w_h^{M+1} = w_h^1 \end{cases}. \tag{6}$$

Theorem 4 ([4]). *Let* $\mathbb{X}_h$ *denote the set of such subsolutions. Then, under the d.m.p, the solution of system of QVIs (5) is the maximum element of the set* $\mathbb{X}_h$.

3 The Main Result

Theorem 5. *There exists a constant* C *independent of both* h *and* k *such that*

$$\max_{1 \le i \le M} \parallel u^i - u_h^i \parallel_\infty \le Ch^2 \left|\log h\right|^2.$$

The optimal order requires a special care in the smoothing of the obstacle functions $k + u_h^{i+1}$. This is ensured by the so-called discrete regularity.

Lemma 1 (Discrete regularity). *There exists a family of right-hand side* $\left\{g^{1(h)}, \ldots, g^{M\ (h)}\right\}_{h>0}$ *and a constant* C *independent of* h *such that* $\left\|g^{i\ (h)}\right\|_\infty \le C$ *and*

$$b^i(u_h^i, v) = (g^{i(h)}, v)\ \forall v \in \mathbb{V}_h. \tag{7}$$

Proof. We adapt [7,8]. □

Let $u^{i,(h)}$ denote the associated continuous solution. Then,

$$\left\|u^{i,(h)}\right\|_{W^{2,p}(\Omega)} \le C \tag{8}$$

and, therefore, thanks to [9], we have

$$\left\| u^{i,(h)} - u_h^i \right\|_{\infty} \leq Ch^2 \left|\log h\right| \ \forall i = 1, 2, \ldots, M. \tag{9}$$

Let us now introduce the following variational inequalities (VIs):

$$\begin{cases} b^i(\bar{u}^i, v - \bar{u}^i) \geqq (f^i + \lambda u^{i,(h)}, v - \bar{u}^i) \, \forall v \in H_0^1(\Omega) \\ \bar{u}^i \leq k + u^{i+1,(h)}, \ v \leq k + u^{i+1(h)} \\ u^{M+1,(h)} = u^{1,(h)} \end{cases}. \tag{10}$$

Let us denote by $\bar{u}^i = \sigma(f^i + \lambda u^{i,(h)}, k + u^{i+1,(h)})$, the solution of the above VI with source term $f^i + \lambda u^{i,(h)}$ and obstacle $k + u^{i+1,(h)}$, $i = 1, 2, \ldots, M$.

Lemma 2.

$$\| \bar{u}^i - u_h^i \|_{\infty} \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M. \tag{11}$$

Proof. Let us denote by $\bar{\omega}_{ih} = \sigma_h(f^i + \lambda u^{i,(h)}, k + u^{i+1,(h)})$ the approximation of $\bar{u}^i = \sigma(f^i + \lambda u^{i,(h)}, k + u^{i+1,(h)})$. Since $u^{i+1,(h)} \in W^{2,p}(\Omega)$, making use of standard results on L^{∞} error estimate for elliptic VIs [10], we have

$$\| \bar{u}^i - \bar{\omega}_{ih} \|_{\infty} \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M.$$

□

On the other hand, since $u_h^i = \sigma_h(f^i + \lambda u_h^i, k + u_h^{i+1})$, combining Lipschitz dependence with respect to both the right-hand side and the obstacle, for elliptic VIs and (9), we get

$$\begin{aligned} &\| \bar{\omega}_{ih} - u_h^i \|_{\infty} \\ &\quad \leq C \left(\left\| (f^i + \lambda u^{i,(h)}) - (f^i + \lambda u_h^i) \right\|_{\infty} + \left\| (k + \ u^{i+1,(h)}) - (k + u_h^{i+1}) \right\|_{\infty} \right) \\ &\quad \leq C\lambda (\| u^{i,(h)} - u_h^{i,} \|_{\infty} + \| u^{i+1(h)} - u_h^{i+1} \|_{\infty}) \leq Ch^2 \left|\log h\right|^2. \end{aligned}$$

Hence

$$\| \bar{u}^i - u_h^i \|_{\infty} \leq \| \bar{u}^i - \bar{\omega}_{ih} \|_{\infty} + \| \bar{\omega}_{ih} - u_h^i \|_{\infty} \leq Ch^2 \left|\log h\right|^2.$$

Lemma 3.

$$\left\| u^{i,(h)} - \bar{u}^i \right\|_{\infty} \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M. \tag{12}$$

Proof. Since

$$\left\| u^{i,(h)} - \bar{u}^i \right\|_{\infty} \leq \left\| u^{i,(h)} - u_h^i \right\|_{\infty} + \left\| u_h^i - \bar{u}^i \right\|_{\infty}.$$

□

Then, making use of (9) and (11), we get

$$\left\| u^{i,(h)} - \bar{u}^i \right\|_\infty \leq Ch^2 \left|\log h\right|^2 .$$

Theorem 6. *There exists* $(\beta^{1,(h)}, \ldots, \beta^{M,(h)})$ *such that*

$$\beta^{i,(h)} \leq u^i \ \ and \ \ \left\| \beta^{i,(h)} - u_h^i \right\|_\infty \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M.$$

Proof. Indeed, $\bar{u}^i$ being the solution of the VI (10), it is also a subsolution for the same VI, that is,

$$\begin{cases} b^i(\bar{u}^i, v) \leq (f^i + \lambda u^{i,(h)}, v) \ \forall v \in H_0^1(\Omega), v > 0 \\ \bar{u}^i \leq k + u^{i+1,(h)}, \ \ v \leq k + u^{i+1(h)} \\ u^{M+1,(h)} = u^{1,(h)} \end{cases} .$$

Then

$$\begin{cases} b^i(\bar{u}^i, v) \leq (f^i + \lambda \left\| u^{i,(h)} - \bar{u}^i \right\|_\infty + \lambda \bar{u}^i, v) \\ \bar{u}^i \leq k + \left\| u^{i+1(h)} - \bar{u}^{i+1} \right\|_\infty + \bar{u}^{i+1} \\ \bar{u}^{M+1,(h)} = \bar{u}^{1,(h)} \end{cases} .$$

So, using (12), we get

$$\begin{cases} b^i(\bar{u}^i, v) \leq (f^i + \lambda Ch^2 \left|\log h\right|^2 + \lambda \bar{u}^i, v) \\ \bar{u}^i \leq k + Ch^2 \left|\log h\right|^2 + \bar{u}^{i+1} \\ \bar{u}^{M+1,(h)} = \bar{u}^{1,(h)} \end{cases}$$

that is, $(\bar{u}^1, \ldots, \bar{u}^M)$ is a subsolution for the system of QVIs (1) with source term $\left(f^1 + \lambda Ch^2 \left|\log h\right|^2, \ldots, f^M + \lambda Ch^2 \left|\log h\right|^2\right)$ and parameter $\tilde{k} = k + Ch^2 \left|\log h\right|^2$. Let us denote by $\bar{U}^i = \sigma(f^i + \lambda Ch^2 \left|\log h\right|^2, k + Ch^2 \left|\log h\right|^2), i = 1, 2 \ldots, M$, the solution of such a system. So, as $u^i = \sigma(f^i, k), i = 1, 2 \ldots, M$, is the solution of system (1), making use of Theorem 1, we have

$$\| u^i - \bar{U}^i \|_\infty \leq C$$
$$\left(\left| k - (k + \lambda Ch^2 \left|\log h\right|^2) \right| + \left\| f^i - (f^i + \lambda Ch^2 \left|\log h\right|^2) \right\|_\infty \right) \leq Ch^2 \left|\log h\right|^2 .$$

Hence, making use of Theorem 2, we obtain

$$\bar{u}^i \leq \bar{U}^i \leq u^i + Ch^2 \left|\log h\right|$$

and, taking

$$\beta^{i,(h)} = \bar{u}^i - Ch^2 \left|\log h\right|^2$$

we clearly have

$$\beta^{i,(h)} \leq u^i .$$

Finally, using (11), we obtain

$$\left\| \beta^{i,(h)} - u_h^i \right\|_\infty \leq \left\| \bar{u}^i - Ch^2 \left|\log h\right|^2 - u_h^i \right\|_\infty \leq Ch^2 \left|\log h\right|^2 .$$

□

Theorem 7. *There exists* $(\alpha_h^1, \ldots, \alpha_h^M)$ *such that*

$$\alpha_h^i \leq u_h^i \text{ and } \left\| \alpha_h^i - u^i \right\|_\infty \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M. \tag{13}$$

Proof. We only sketch the proof which uses the VIs:

$$\begin{cases} b^i(\bar{u}_h^i, v - \bar{u}_h^i) \geqq (f^i + \lambda u^i, v - \bar{u}_h^i) \ \forall v \in \mathbb{V}_h \\ \bar{u}_h^i \leq r_h(k + u^{i+1}),\ v \leq r_h(k + u^{i+1}) \\ u^{M+1} = u^1 \end{cases} \tag{14}$$

and combines the estimate

$$\| u^i - \bar{u}_h^i \|_\infty \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M \tag{15}$$

with Theorems 3 and 4 to get the discrete subsolution

$$\alpha_h^i = \bar{u}_h^i - Ch^2 \left|\log h\right|^2$$

satisfying (13).

Now, combining Theorems 6 and 7, we are in a position to derive the main result.

□

Proof. Indeed, making use of both Theorems 6 and 7, we have

$$\begin{aligned} u_h^i &\leq \beta^{i,(h)} + Ch^2 \left|\log h\right|^2 \\ &\leq u^i + Ch^2 \left|\log h\right|^2 \\ &\leq \alpha_h^i + Ch^2 \left|\log h\right|^2 . \end{aligned}$$

Thus

$$\| u^i - u_h^i \|_\infty \leq Ch^2 \left|\log h\right|^2 \ \forall i = 1, 2, \ldots, M.$$

□

References

1. Evans, L.C., Friedman A.: Optimal stochastic switching and the Dirichlet problem for the Bellman equations. Trans. Am. Math. Soc. **253**, 365–389 (1979)
2. Lions, P.L., Menaldi, J.L.: Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations (part I). SIAM Contr. Optim. **20**, 58–81 (1982)
3. Boulbrachene, M., Cortey-Dumont, P.: Optimal L^{∞} error estimate of a finite element approximation of Hamilton-Jacobi-Bellman. Numer. Funct. Anal. Optim. **41**, 421–435 (2009)
4. Boulbrachene, M.: L^{∞} error estimate for a system of elliptic quasi-variational inequalities with noncoercive operators. Comput. Math. Appl. **45**, 983–989 (2003)
5. Ciarlet, P.G., Raviart, P.A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Eng. **2**, 17–31 (1973)
6. Karatson, J., Korotov, S.: Discrete maximum principle for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. Numer. Math. **99**, 669–698 (2005)
7. Cortey Dumont, P.: Contribution a l' approximation des inequations variationnelles en norme L^{∞}. C. R. Acad. Sci. Paris Ser. I Math. **296**, 17 (1983)
8. Cortey Dumont, P.: Sur l' analyse numerique des equations de Hamilton-Jacobi-Bellman. Math. Methods Appl. Sci. **9**, 198–209 (1987)
9. Nitsche, J.: L^{∞}-convergence of finite element approximations. In: Mathematical Aspects of Finite Element Methods. Lecture Notes in Mathematics, vol. 606. Springer, Berlin (1977)
10. Cortey-Dumont, P.: On the finite element approximation in the L^{∞} norm of variational inequalities with nonlinear operators. Numer. Math. **47**, 45–57 (1985)

Convergence of Finite Element Approximations for Generalized Marguerre–von Kármán Equations

A. Ghezal and D.A. Chacha

Abstract In this work, we establish the convergence of a conforming finite element approximations to the generalized Marguerre–von Kármán equations. More precisely, we consider here the generalized Marguerre–von Kármán equations, which constitute a mathematical model for a nonlinearly elastic shallow shell subjected to boundary conditions of von Kármán's type only on a portion of its lateral face, the remaining portion being free. We first reduce the discrete problem of these equations to a single discrete cubic operator equation, whose unknown is the approximate of vertical displacement of the shallow shell. We next solve this discrete operator equation, by adapting a compactness method due to J.L. Lions and on Brouwer's fixed point theorem (Lions, Quelques méthodes de résolution des problèmes aux limites non linéaires, Dunod, Paris, 1969). Then we establish the convergence of a conforming finite element approximations to these equations.

Keywords Marguerre–von Kármán equations • Finite element method • Compactness method

1 Introduction

The two-dimensional Marguerre–von Kármán equations for nonlinearly elastic shallow shells were originally proposed by Marguerre [1] in 1938 and von Kármán and Tsien [2] in 1939; they generalize the equations of von Kármán for thin elastic plates proposed by von Kármán [3] in 1910.

In 1986, Ciarlet and Paumier [4] justified the classical Marguerre–von Kármán equations by means of a formal asymptotic analysis. Then, in 2002, Gratie [5] has

A. Ghezal (✉) • D.A. Chacha
Laboratoire de mathématiques appliquées, Université Kasdi Merbah,
B.P 511, Ouargla 30000, Algérie
e-mail: ghezal.abderrezak@univ-ouargla.dz; Chacha.dj@univ-ouargla.dz

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_9,

generalized these equations, where only a portion of the lateral face is subjected to boundary conditions of von Kármán's type, the remaining portion being free. She showed that the leading term of the asymptotic expansion is characterized by a two-dimensional boundary value problem called generalized Marguerre–von Kármán equations. In 2006, Ciarlet and Gratie [6] have established an existence theorem for these equations. In the same way but for the dynamical case, we quote the previous works [7, 8], where we recently identified the dynamical equations of generalized Marguerre–von Kármán shallow shells and we established the existence of solutions to these equations using compactness method of Lions [9]. In this direction, we quote also the previous work [10] for justification of the generalized Marguerre–von Kármán equations with Signorini conditions.

For numerical approximations, some studies have been done for the von Kármán equations. Miyoshi [11] studied the mixed finite element method for these equations. Kesavan [12, 13] proposed an iterative finite element method of the bifurcation branches near simple eigenvalues of the linearized problem of von Kármán equations and mixed finite element method for the same problem. Brezzi [14] and Brezzi et al. [15, 16] analyzed a finite element approximations of von Kármán plate bending equations and studied a Hellan-Herrmann-Johnson mixed finite element scheme for the von Kármán equations. Reinhart [17] proposed an approximation of the von Kármán equations using a Hermann-Miyoshi finite element scheme. Ciarlet et al. [18] studied the finite element method for the generalized von Kármán equations.

The objective of this study is to extend the results which studied by Ciarlet et al. [18] to the generalized Marguerre–von Kármán shallow shell.

2 Generalized Marguerre–von Kármán Equations

Let ω be a connected bounded open subset of $\mathbb{R}^2$ with a Lipschitz-continuous boundary γ, ω being locally on a single side of γ, and we assume $0 \in \gamma$ and we denote by $\gamma(y)$ the arc joining 0 to the point $y \in \gamma$. Let γ_1 be a relatively open subset of γ such that length $\gamma_1 > 0$ and length $\gamma_2 > 0$, where $\gamma_2 = \gamma \backslash \gamma_1$. The unit outer normal vector (ν_α) and the unit tangent vector (τ_α) along the boundary γ are related by $\tau_1 = -\nu_2$ and $\tau_2 = \nu_1$. The outer normal and tangential derivative operators $\nu_\alpha \partial_\alpha$ and $\tau_\alpha \partial_\alpha$ along γ are denoted respectively by ∂_ν and ∂_τ. As shown in [6], the generalized Marguerre–von Kármán equations are written as

$$-\partial_{\alpha\beta} m_{\alpha\beta}(\nabla^2 \xi) = [\Phi, \xi + \tilde{\theta}] + f \text{ in } \omega,$$

$$\Delta^2 \Phi = -[\xi, \xi + 2\tilde{\theta}] \text{ in } \omega,$$

$$\xi = \partial_\nu \xi = 0 \text{ on } \gamma_1,$$

$$m_{\alpha\beta}(\nabla^2\xi)\nu_\alpha\nu_\beta = 0 \text{ on } \gamma_2,$$
$$\partial_\alpha m_{\alpha\beta}(\nabla^2\xi)\nu_\beta + \partial_\tau(m_{\alpha\beta}(\nabla^2\xi)\nu_\alpha\tau_\beta) = 0 \text{ on } \gamma_2,$$
$$\Phi = \Phi_0 \text{ and } \partial_\nu\Phi = \Phi_1 \text{ on } \gamma,$$

where

$$m_{\alpha\beta}(\nabla^2\xi) = -\frac{1}{3}\left\{\frac{4\lambda\mu}{\lambda+2\mu}\Delta\xi\delta_{\alpha\beta} + 4\mu\partial_{\alpha\beta}\xi\right\},$$
$$\Phi_0(y) = -\gamma_1\int_{\gamma(y)}\tilde{h}_2 d\gamma + \gamma_2\int_{\gamma(y)}\tilde{h}_1 d\gamma + \int_{\gamma(y)}(x_1\tilde{h}_2 - x_2\tilde{h}_1)d\gamma,\ y \in \gamma,$$
$$\Phi_1(y) = -\nu_1\int_{\gamma(y)}\tilde{h}_2 d\gamma + \nu_2\int_{\gamma(y)}\tilde{h}_1 d\gamma,\ y \in \gamma,$$
$$[\Phi, \xi] = \partial_{11}\Phi\partial_{22}\xi + \partial_{22}\Phi\partial_{11}\xi - 2\partial_{12}\Phi\partial_{12}\xi.$$

The known functions $\tilde{\theta}$ and f are, up to constant factors, the function that defines the middle surface of the shell and the resultant of the vertical forces acting on the shell. The functions Φ_0 and Φ_1 are known functions of the appropriately "scaled" density $(h_\alpha) : \gamma_1 \to \mathbb{R}^2$ of the resultant of the horizontal forces acting on the portion of the lateral face of the shell with γ_1 as its middle line and the functions $\tilde{h}_\alpha \in L^2(\gamma)$ defined by $\tilde{h}_\alpha = h_\alpha$ on γ_1, $\tilde{h}_\alpha = 0$ on γ_2. The constants λ and μ are the Lamé constants of the material. The unknown $\xi : \bar{\omega} \to \mathbb{R}$ is, up to constant factors, the vertical component of the displacement field of the middle surface of the shell and the unknown $\Phi : \bar{\omega} \to \mathbb{R}$ is the Airy function.

3 The Continuous Cubic Operator Equation

Let us briefly recall some of the results obtained in [6] concerning the properties of the continuous cubic operator equation.

Let $\tilde{\chi} \in H^2(\omega)$ denote the unique solution of the boundary value problem:

$$\Delta^2\tilde{\chi} = [\tilde{\theta}, \tilde{\theta}] \text{ in } \omega, \tag{1}$$

$$\tilde{\chi} = \Phi_0 \text{ and } \partial_\nu\tilde{\chi} = \Phi_1 \text{ on } \gamma. \tag{2}$$

Let $F \in V(\omega)$ denote the unique solution of the boundary value problem:

$$-\partial_{\alpha\beta}m_{\alpha\beta}\left(\nabla^2 F\right) = f \text{ in } \omega, \tag{3}$$

$$F = \partial_\nu F = 0 \text{ on } \gamma_1, \tag{4}$$

$$m_{\alpha\beta}\left(\nabla^2 F\right)\nu_\alpha\nu_\beta = 0 \text{ on } \gamma_2, \tag{5}$$

$$\partial_\alpha m_{\alpha\beta}\left(\nabla^2 F\right)\nu_\beta + \partial_\tau\left(m_{\alpha\beta}\left(\nabla^2 F\right)\nu_\alpha\tau_\beta\right) = 0 \text{ on } \gamma_2, \tag{6}$$

where

$$V(\omega) = \left\{\eta \in H^2(\omega); \eta = \partial_\nu\eta = 0 \text{ on } \gamma_1\right\}.$$

Let the bilinear mapping:

$$B : H^2(\omega) \times H^2(\omega) \to H_0^2(\omega),$$

be defined as follows: for each pair $(\xi, \eta) \in H^2(\omega) \times H^2(\omega)$, the function $B(\xi, \eta) \in H_0^2(\omega)$ is the unique solution of the boundary value problem:

$$\Delta^2 B(\xi, \eta) = [\xi, \eta] \text{ in } \omega, \tag{7}$$

$$B\,(\xi, \eta) = \partial_\nu B\,(\xi, \eta) = 0 \text{ on } \gamma. \tag{8}$$

Let the second bilinear mapping:

$$\tilde{B} : H^2(\omega) \times H^2(\omega) \to V(\omega),$$

be defined as follows: for each pair $(\Phi, \xi) \in H^2(\omega) \times H^2(\omega)$, the function $\tilde{B}(\Phi, \xi) \in V(\omega)$ is the unique solution of the boundary value problem:

$$-\partial_{\alpha\beta} m_{\alpha\beta}(\nabla^2 \tilde{B}(\Phi, \xi)) = [\Phi, \xi] \text{ in } \omega, \tag{9}$$

$$\tilde{B}(\Phi, \xi) = \partial_\nu \tilde{B}(\Phi, \xi) = 0 \text{ on } \gamma_1, \tag{10}$$

$$m_{\alpha\beta}(\nabla^2 \tilde{B}(\Phi, \xi))\nu_\alpha\nu_\beta = 0 \text{ on } \gamma_2, \tag{11}$$

$$\partial_\alpha m_{\alpha\beta}(\nabla^2 \tilde{B}(\Phi, \xi))\nu_\beta + \partial_\tau(m_{\alpha\beta}(\nabla^2 \tilde{B}(\Phi, \xi))\nu_\alpha\tau_\beta) = 0 \text{ on } \gamma_2. \tag{12}$$

First, Ciarlet and Gratie [6] have shown that the generalized Marguerre–von Kármán equations are reduced to a cubic operator equation, such that a pair $(\xi, \Phi) \in V(\omega) \times H^2(\omega)$ satisfies the generalized Marguerre–von Kármán equations if and only if the function $\tilde{\xi} = (\tilde{\theta} + \xi) \in V(\omega)$ satisfies the cubic operator equation:

$$\tilde{C}(\tilde{\xi}) + (I - \tilde{L})\tilde{\xi} - \tilde{F} = 0, \tag{13}$$

and the Airy function $\Phi \in H^2(\omega)$ is given by

$$\Phi = \tilde{\chi} - B(\tilde{\xi}, \tilde{\xi}), \tag{14}$$

where the cubic mapping $\tilde{C} : V(\omega) \to V(\omega)$ is defined by $\tilde{C}(\eta) = \tilde{B}(B(\eta, \eta), \eta)$, the linear mapping $\tilde{L} : V(\omega) \to V(\omega)$ is defined by $\tilde{L}\eta = \tilde{B}(\tilde{\chi}, \eta)$, and $\tilde{F} = \tilde{\theta} + F$.

Noting that, finding the solution $\tilde{\xi}$ of the above operator equation (13) is equivalent to solving the following variational problem:

$$(P) \begin{cases} \text{Find } \tilde{\xi} \in V(\omega) \text{ such that,} \\ ((\tilde{C}(\tilde{\xi}) + (I - \tilde{L})\tilde{\xi} - \tilde{F}, \eta)) = 0 \text{ for all } \eta \in V(\omega), \end{cases}$$

where $((.,.))$ is the inner product on $V(\omega)$ defined by $((\zeta, \eta)) = -\int_\omega m_{\alpha\beta}(\nabla^2\zeta) \partial_{\alpha\beta}\eta d\omega$ and let $\|.\|$ denote the norm associated with the inner product $((.,.))$ which is equivalent to the norm $\|.\|_{H^2(\omega)}$ over the space $V(\omega)$.

Next, Ciarlet and Gratie [6] have shown that, under the assumptions (ω is simply connected, the functions $\tilde{h}_\alpha$ satisfy natural compatibility conditions, and the norms $\|h_\alpha\|_{L^2(\gamma_1)}$ are small enough), the generalized Marguerre–von Kármán equations have at least one solution $(\xi, \Phi) \in V(\omega) \times H^2(\omega)$ in the sense of distributions.

The cubic operator equation (13) generalizes an operator equation originally introduced by Berger [19] and Berger and Fife [20], then used by Ciarlet et al. [6,21] for analyzing the generalized von Kármán and Marguerre–von Kármán equations.

4 The Discrete Cubic Operator Equation

We assume that γ is a polygon. Let $W_h \subset H^2(\omega)$, $V_h \subset V(\omega)$, $V_{0h} \subset H_0^2(\omega)$, be standard conforming finite element spaces satisfying the minimal conditions of [22, Theorem 6.1-7]. Strong and weak convergence are noted $\to$ and $\rightharpoonup$ respectively. All convergence are meant to hold as $h \to 0$.

Let $\tilde{\chi}_h \in W_h$ denote standard finite element approximation of $\tilde{\chi} \in H^2(\omega)$, which therefor satisfies $\|\tilde{\chi}_h - \tilde{\chi}\|_{H^2(\omega)} \to 0$.

Let $F_h \in V_h$ denote the unique solution of the variational equation

$$-\int_\omega \partial_{\alpha\beta} m_{\alpha\beta}(\nabla^2 F_h)\eta_h d\omega = \int_\omega f\eta_h d\omega \text{ for all } \eta_h \in V_h,$$

which satisfies $\|F_h - F\|_{H^2(\omega)} \to 0$.

Let the bilinear mapping $B_h : H^2(\omega) \times H^2(\omega) \to V_{0h}$ be defined as follows: for each pair $(\xi, \eta) \in H^2(\omega) \times H^2(\omega)$, the function $B_h(\xi, \eta) \in V_{0h}$ is the unique solution of the variational equation

$$\int_\omega \Delta B_h(\xi, \eta)\Delta\varsigma_h d\omega = \int_\omega [\xi, \eta]\varsigma_h d\omega \text{ for all } \varsigma_h \in V_{0h},$$

hence, for $(\xi, \eta) \in H^2(\omega) \times H^2(\omega)$ fixed, $\|B_h(\xi, \eta) - B(\xi, \eta)\|_{H^2(\omega)} \to 0$.

Finally, let the bilinear mapping $\tilde{B}_h : H^2(\omega) \times H^2(\omega) \to V_h$ be defined as follows: for each pair $(\Phi, \xi) \in H^2(\omega) \times H^2(\omega)$, the function $\tilde{B}_h(\Phi, \xi) \in V_h$ is the unique solution of the variational equation

$$-\int_\omega \partial_{\alpha\beta} m_{\alpha\beta}(\nabla^2 \tilde{B}_h(\Phi, \xi))\eta_h d\omega = \int_\omega [\Phi, \xi]\eta_h d\omega \text{ for all } \eta_h \in V_h,$$

hence, for $(\Phi, \xi) \in H^2(\omega) \times H^2(\omega)$ fixed, $\|\tilde{B}_h(\Phi, \xi) - \tilde{B}(\Phi, \xi)\|_{H^2(\omega)} \to 0$.

For each $h > 0$, the discrete problem is then defined through the following theorem:

Theorem 1. *The discrete problem of generalized Marguerre–von Kármán equations consists in finding $(\tilde{\xi}_h, \Phi_h) \in V_h \times W_h$, such that $\tilde{\xi}_h$ satisfies the discrete operator equation:*

$$\tilde{C}_h(\tilde{\xi}_h) + (I - \tilde{L}_h)\tilde{\xi}_h - \tilde{F}_h = 0 \textit{ in } V_h, \tag{15}$$

and Φ_h is given by

$$\Phi_h = \tilde{\chi}_h - B_h(\tilde{\xi}_h, \tilde{\xi}_h) \textit{ in } W_h, \tag{16}$$

where the discrete cubic mapping $\tilde{C}_h : V_h \to V_h$ is defined by $\tilde{C}_h(\eta_h) = \tilde{B}_h(B_h(\eta_h, \eta_h), \eta_h)$, the linear mapping $\tilde{L}_h : V_h \to V_h$ is defined by $\tilde{L}_h\eta_h = \tilde{B}_h(\tilde{\chi}_h, \eta_h)$, $\tilde{\xi}_h = \tilde{\theta} + \xi_h$ and $\tilde{F}_h = \tilde{\theta} + F_h$.

Proof. The discrete problem of generalized Marguerre–von Kármán equations consists in finding $(\tilde{\xi}_h, \Phi_h) \in V_h \times W_h$, such that $\tilde{\xi}_h$ satisfies the variational equation

$$-\int_\omega \partial_{\alpha\beta} m_{\alpha\beta}(\nabla^2(\tilde{\xi}_h - \tilde{\theta}))\eta_h d\omega = \int_\omega ([\Phi_h, \tilde{\xi}_h] + f)\eta_h d\omega \text{ for all } \eta_h \in V_h, \tag{17}$$

and Φ_h satisfies the variational equation

$$\int_\omega \Delta^2 \Phi_h . \vartheta_h = \int_\omega ([\tilde{\theta}, \tilde{\theta}] - [\tilde{\xi}_h, \tilde{\xi}_h])\vartheta_h d\omega \text{ for all } \vartheta_h \in W_h. \tag{18}$$

By definition of the function $\tilde{\chi}_h$ and the mapping B_h, (18) imply that

$$\Phi_h = \tilde{\chi}_h - B_h(\tilde{\xi}_h, \tilde{\xi}_h) \text{ in } W_h.$$

By definition of the function $\tilde{F}_h$ and the mapping $\tilde{B}_h$, (17) imply that

$$\tilde{\xi}_h - \tilde{F}_h = \tilde{B}_h(\Phi_h, \tilde{\xi}_h) \text{ in } V_h. \tag{19}$$

Eliminating Φ_h between these two operator equations (16) and (19) yields the single operator equation

$$\tilde{B}_h(B_h(\tilde{\xi}_h,\tilde{\xi}_h),\tilde{\xi}_h)+\tilde{\xi}_h-\tilde{B}_h(\tilde{\chi}_h,\tilde{\xi}_h)-\tilde{F}_h=0 \text{ in } V_h.$$

Then, we conclude that $\tilde{\xi}_h \in V_h$ is found by solving the discrete operator equation:

$$\tilde{C}_h(\tilde{\xi}_h)+(I-\tilde{L}_h)\tilde{\xi}_h-\tilde{F}_h=0 \text{ in } V_h.$$

□

5 Convergence

Note that finding $\tilde{\xi}_h$ is equivalent to solving the following discrete variational problem:

$$(P_h)\begin{cases} \text{Find } \tilde{\xi}_h \in V_h \text{ such that,} \\ ((\tilde{C}_h(\tilde{\xi}_h)+(I-\tilde{L}_h)\tilde{\xi}_h-\tilde{F}_h,\tilde{\eta}_h))=0 \ \text{ for all } \ \tilde{\eta}_h \in \ V_h, \end{cases}$$

where $((.,.))$ is the inner product on V_h defined by

$$((\tilde{\zeta}_h,\tilde{\eta}_h))=-\int_\omega m_{\alpha\beta}(\nabla^2\tilde{\zeta}_h)\partial_{\alpha\beta}\tilde{\eta}_h d\omega.$$

In order to show that the discrete variational problem (P_h) has at least one solution in Theorem 4 below, we will need the following two lemmas:

Lemma 2. *The trilinear form*

$$(\zeta,\eta,\varsigma)\in[H^2(\omega)]^3\to\int_\omega[\zeta,\eta]\varsigma d\omega\in\mathbb{R},$$

is continuous; moreover, becomes symmetric form if at least one of the three spaces $H^2(\omega)$ is replaced by the space $H_0^2(\omega)$.

Proof. The proof is detailed in the proof [part (i)] of [23, Theorem 5.8-2] and in the proof [part (i)] of [18, Theorem 4.1]. □

Lemma 3. *The bilinear mapping B_h is sequentially compact, in the sense that, if*

$$(\xi_h,\eta_h)\rightharpoonup(\xi,\eta)\in[H^2(\omega)]^2,$$

then

$$B_h(\xi_h,\eta_h)\to B_h(\xi,\eta)\in H_0^2(\omega).$$

Proof. We define the following inner product on $H_0^2(\omega)$

$$(\zeta,\varsigma)_\Delta = \int_\omega \Delta\zeta\Delta\varsigma d\omega,$$

and let $\|.\|_\Delta$ denote the norm over the space $H_0^2(\omega)$, which corresponds to the inner product $(.,.)_\Delta$.

From the definition of the mapping B_h, we get

$$(B_h(\xi,\eta),\varsigma)_\Delta = \int_\omega [\xi,\eta]\varsigma d\omega,$$

for all $(\xi,\eta,\varsigma) \in [H^2(\omega)]^2 \times H_0^2(\omega)$.

Then there exists a constant c_1 such that

$$\|B_h(\xi,\eta)\|_\Delta \leq c_1\|\xi\|_{W^{1,4}(\omega)}\|\eta\|_{W^{1,4}(\omega)}, \tag{20}$$

for all $(\xi,\eta) \in [H^2(\omega)]^2$.

Let $(\xi_h,\eta_h) \rightharpoonup (\xi,\eta) \in [H^2(\omega)]^2$, and using the bilinearity of B_h, we have

$$B_h(\xi_h,\eta_h) - B_h(\xi,\eta) = B_h(\xi_h-\xi,\eta) + B_h(\xi,\eta_h-\eta) + B_h(\xi_h-\xi,\eta_h-\eta).$$

From (20), it follows that there exists a constant c_2 such that

$$\|B_h(\xi_h,\eta_h)-B_h(\xi,\eta)\|_\Delta \leq c_2(\|\xi_h-\xi\|_{W^{1,4}(\omega)}\|\eta\|_{W^{1,4}(\omega)}+\|\xi\|_{W^{1,4}(\omega)}\|\eta_h-\eta\|_{W^{1,4}(\omega)} \\ +\|\xi_h-\xi\|_{W^{1,4}(\omega)}\|\eta_h-\eta\|_{W^{1,4}(\omega)}).$$

The compact imbedding of $H^2(\omega)$ into $W^{1,4}(\omega)$ implies that $B_h(\xi_h,\eta_h) \to B_h(\xi,\eta) \in H_0^2(\omega)$, for more details see the proof [part (iv)] of [23, Theorem 5.8-2]. □

Theorem 4. *Assume that ω is simply connected, the functions $h_\alpha : \gamma_1 \to \mathbb{R}$ satisfy natural compatibility conditions, and their norms $\|h_\alpha\|_{L^2(\gamma_1)}$ are small enough. Then*

(a) There exists a constant M such that, for each $h > 0$, the discrete variational problem (P_h) has at least one solution $\tilde{\xi}_h \in V_h$ that satisfies $\|\tilde{\xi}_h\| \leq M$.

(b) Let $(\tilde{\xi}_h)_{h>0}$ be any subsequence that weakly converges in $H^2(\omega)$, let $\xi \in V(\omega)$ denote its limit, and let the associated subsequence $(\Phi_h)_{h>0}$ be defined by (16). Then ξ is a solution of the variational problem (P), and

$$(\tilde{\xi}_h,\Phi_h) \to (\tilde{\xi},\Phi) \text{ in } H^2(\omega)\times H^2(\omega),$$

where Φ is defined by (14).

6 Conclusion and Commentary

This study is concerned with finite element method for approximating solutions to the generalized Marguerre–von Kármán equations, solving these equations amounts to solving a single discrete cubic operator equation. Then we establish the convergence of a conforming finite element approximation to these equations, using weak regularity on solutions, but in order to get an error estimate it needs more regularity.

Note that, in the case $\theta \equiv 0$ in $\bar{\omega}$, we recover the generalized von Kármán equations.

As future work, we will extend these results to the dynamical case.

References

1. Marguerre, K.: Zur Theorie der gekrummten Platte grosser Formanderung. In: Proceedings of the 5th International Congress for Applied Mechanics, pp. 93–101 (1938)
2. von Kármán, T., Tsien, H.S.: The buckling of spherical shells by external pressure. J. Aero. Sci. **7**, 43–50 (1939)
3. von Kármán, T.: Festigkeitesprobleme in Maschinenbau. Encyklopadie der Mathematischen Wissenschaften, vol. IV/4, pp. 311–385. Taubner, Leipzig (1910)
4. Ciarlet, P.G., Paumier, J.C.: A justification of the Marguerre-von Kármán equations. Comput. Mech. **1**, 177–202 (1986)
5. Gratie, L.: Generalized Marguerre-von Kármán equations of a nonlinearly elastic shallow shell. Appl. Anal. **81**, 1107–1126 (2002)
6. Ciarlet, P.G., Gratie, L.: On the existence of solutions to the generalized Marguerre-von Kármán equations. Math. Mech. Solids **11**, 83–100 (2006)
7. Chacha, D.A., Ghezal, A., Bensayah, A.: Modélisation asymptotique d'une coque peu-profonde de Marguerre-von Kármán généralisée dans le cas dynamique. J. ARIMA **13**, 63–76 (2010)
8. Chacha, D.A., Ghezal, A., Bensayah, A.: Existence result for a dynamical equations of generalized Marguerre-von Kármán shallow shells. J. Elast. **111**, 265–283 (2013)
9. Lions, J.L.: Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires. Dunod, Paris (1969)
10. Bensayah, A., Chacha, D.A., Ghezal, A.: Asymptotic modelling of a signorini problem of generalized Marguerre-von Kármán shallow shells. Appl. Anal. **92**(9), 1848–1862 (2013)
11. Miyoshi, T.: A mixed finite element method for the solution of von Kármán equations. Numer. Math. **26**, 255–269 (1976)
12. Kesavan, S.: La méthode de Kikuchi appliquée aux équations de von Kármán. Numer. Math. **32**, 209–232 (1979)
13. Kesavan, S.: Une méthode d'éléments finis mixte pour les équations de von Kármán. RAIRO Anal. Numér. **14**(2), 149–173 (1980)
14. Brezzi, F.: Finite element approximations of the von Kármán equations. RAIRO Anal. Numér. **12**(4), 303–312 (1978)
15. Brezzi, F., Rappaz, J., Raviart, P.A.: Finite dimensional approximation of nonlinear problems, Part I: branches of nonsingular solutions. Numer. Math. **36**, 1–25 (1980)
16. Brezzi, F., Rappaz, J., Raviart, P.A.: Finite dimensional approximation of nonlinear problems, Part III: simple bifurcation points. Numer. Math. **38**, 1–30 (1981)

17. Reinhart, L.: On the numerical analysis of the von Kármán equations: mixed finite element approximation and continuation techniques. Numer. Math. **39**, 371–404 (1982)
18. Ciarlet, P.G., Gratie, L., Kesavan, S.: On the generalized von Kármán equations and their approximation. J. Math. Models Methods Appl. Sci. **4**(17), 617–633 (2007)
19. Berger, M.S.: On von Kármán equations and the buckling of a thin elastic plate. I. The clamped plate. Commun. Pure Appl. Math. **20**, 687–719 (1967)
20. Berger, M.S., Fife, P.C.: On von Kármán equations and the buckling of a thin elastic plate. II. Plate with general edge conditions. Commun. Pure Appl. Math. **21**, 227–241, (1968)
21. Ciarlet, P.G., Gratie, L., Sabu, N.: An existence theorem for generalized von Kármán equations. J. Elast. **62**, 239–248 (2001)
22. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
23. Ciarlet, P.G.: Mathematical Elasticity, Theory of Plates, vol. II. North-Holland, Amsterdam (1997)

The Maple Program Procedures at Solution Systems of Differential Equation with Taylor Collocation Method

S. Servi, Y. Keskin, and G. Oturanç

Abstract In this paper, a maple algorithm Taylor collocation method has been presented for numerically solving the systems of differential equation with variable coefficients under the mixed conditions. The solution is obtained in terms of Taylor polynomials. This method is based on taking the truncated Taylor series of the function in equations and then substituting their matrix forms in the given equation. Hence, the result of matrix equation can be solved and the unknown Taylor coefficients can be found approximately. The results obtained by Taylor collocation method will be compared with the results of differential transform method and Adomian decomposition method.

Keywords Taylor collocation method • Maple program

1 Introduction

Numerical methods which are based on algorithm and given solutions fastly, are come into prominence for solution of differential equations are encountered in applied mathematics and some of engineering problems, don't have analytical solutions or have so difficult and time-consuming solutions. One of these methods is Taylor collocation method. Taylor collocation method, which is given for the solution of systems of linear differential equations [1], is developed to find the approximate solutions of high-order systems of linear differential equations with variable coefficients.

S. Servi (✉) • Y. Keskin • G. Oturanç
Department of Mathematics, Science Faculty, Selçuk University, Konya 42075, Turkey
e-mail: sservi@selcuk.edu.tr

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_10,

2 Taylor Collocation Method

The Taylor method is developed to find an approximate solution of high-order linear differential–difference equations, integro differential equations with variable coefficients under the mixed conditions [2]. The solution is obtained in terms of Taylor polynomials. Firstly, this method is based on taking the truncated Taylor series of the function in equations and then substituting their matrix forms in the given equation. Hence, the result of the matrix equation can be solved and the unknown Taylor coefficients can be found approximately [1, 3–6].

mth-order linear differential equation with variable coefficients

$$\sum_{k=0}^{m} P_k(x)y^{(k)}(x) = f(x),\ a \le x \le b \tag{1}$$

with the mixed conditions

$$\sum_{j=0}^{m-1} [a_{ij}y^{(j)}(a) + b_{ij}y^{(j)}(b) + c_{ij}y^{(j)}(c)] = \lambda_i, i = 0, 1, \dots, m-1;\ a \le c \le b \tag{2}$$

then we can write the Eq. (1)

$$P_m(x)y^{(m)}(x) + \cdots + P_1(x)y^{(1)}(x) + P_0(x)y(x) = f(x),\ a \le x \le b \tag{3}$$

and the approximate solution is expressed in the truncated Taylor series,

$$y(x) = \sum_{n=0}^{N} \frac{y^{(n)}(c)}{n!}(x-c)^n,\ a \le x, c \le b,\ N \ge m. \tag{4}$$

Here $P_k(x)$ $(k = 0, 1, \dots, m)$ $f(x)$ are functions defined on $a \le c \le b$; the real coefficients $a_{i,j}, b_{i,j}, c_{i,j}, \lambda_i$ are appropriate constants. N number shows till which term of the series it will be expansion and $y^{(n)}(c)$ are the Taylor coefficients to be determined. We use collocation points at defined interval of the problem to find the Taylor coefficients.

$$a = x_0 < x_1 < \cdots < x_N = b$$

and the collocation points,

$$x_i = a + i\frac{b-a}{N}, i = 0, 1, 2, \dots, N$$

then we can put series (3) in the matrix form

$$[y(x)] = XM_0A, \tag{5}$$

where

$$X = \left[\, 1 \ (x-c) \ (x-c)^2 \cdots (x-c)^n \,\right]$$

$$A = \left[\, y^{(0)}(c) \ y^{(1)}(c) \ y^{(2)}(c) \cdots y^{(n)}(c) \,\right]^t$$

$$M_0 = \begin{bmatrix} \frac{1}{0!} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{1!} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{2!} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{N!} \end{bmatrix}.$$

Firstly, we substitute x_i Taylor collocation points in Eq. (5):

$$[y(x_i)] = X_i M_0 A; \ i = 0, 1, \ldots, N \tag{6}$$

$$X_i = \left[\, 1 \ (x_i - c) \ (x_i - c)^2 \cdots (x_i - c)^n \,\right]$$

and

$$\begin{aligned} [y(x_0)] &= X_0 M_0 A \\ [y(x_1)] &= X_1 M_0 A \\ &\vdots \\ [y(x_N)] &= X_N M_0 A \end{aligned}$$

$$Y^{(0)} = CM_0 A, \tag{7}$$

where the matrixes form $Y^{(0)}$ and C:

$$Y^{(0)} = \left[\, y(x_0) \ y(x_1) \ y(x_2) \cdots y(x_N) \,\right]^t$$

$$C = \left[\, x_0 \ x_1 \cdots x_N \,\right]^t = \begin{bmatrix} 1 & (x_0 - c) & (x_0 - c)^2 & \cdots & (x_0 - c)^N \\ 1 & (x_1 - c) & (x_1 - c)^2 & \cdots & (x_1 - c)^N \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & (x_N - c) & (x_N - c)^2 & \cdots & (x_N - c)^N \end{bmatrix}.$$

$y^{(k)}(c)$ are the matrix forms of the derivates functions

$$\left[y^{(k)}(x)\right] = XM_k A, k = 0, 1, \ldots, m \le N. \tag{8}$$

We substitute x_i Taylor collocation points in Eq. (8),

$$Y^{(k)} = CM_k A, \ k = 0, 1, \ldots, \ m \le N, \tag{9}$$

where matrix $Y^{(k)}$

$$Y^{(k)} = \left[\, y^{(k)}(x_0)\ y^{(k)}(x_1)\ y^{(k)}(x_2) \ldots y^{(k)}(x_N) \right]^t .$$

We substitute x_i Taylor collocation points in Eq. (2),

$$P_0Y^{(0)} + P_1Y^{(1)} + \cdots + P_mY^{(m)} = F \text{ or } \sum_{k=0}^{m} P_kY^{(k)} = F, \tag{10}$$

where matrixesP_k and F for $k = 0, 1, \ldots, m \leq N$,

$$P_k = \begin{bmatrix} P_k(x_0) & 0 & \cdots & 0 \\ 0 & P_k(x_1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & P_k(x_N) \end{bmatrix}_{(N+1)\times(N+1)}, \quad F = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix}_{(N+1)\times 1},$$

We substitute $Y^{(k)}$ Taylor collocation points in Eq. (9),

$$\left\{ \sum_{k=0}^{m} P_kCM_k \right\} A = F. \tag{11}$$

As the abovementioned matrixes aren't easy to calculate, we can show the matrixes by calculating via Maple. The procedures of these matrixes in Maple and an example can be written as [5]

Procedure 1

```
Mmatrix:= proc (N,m)
local i,j,k,f,M;
for k from 0 to 5 do
f[k]:=(i,j) -> piecewise(i=j-k,1/(i-1)!):
M[k]:=matrix(N,N,f[k]);
od:
eval(M[m]):
end:
where,N is dimension and m is subscript.
```

Procedure 2

```
Pmatrix:= proc (N,a,b,p)
local i,j,k,M,P,g,h;
with(linalg):
for k from 0 to N do
h[k]:=a+k*(b-a)/(N-1):od:
```

```
for k from 1 to N do
g[k]:=(i,j) -> piecewise(i=j,subs(x=h[i-1],p)):
P:=matrix(N,N,g[k]):od:
eval(P):end:
where, N is dimension and p is a P(x) polynomial.
```

Procedure 3

```
Cmatrix:= proc (N,a,b)
local i,j,k,f,M,x,h,g;
with(linalg):
for k from 1 to N do
x[k]:=a+(k-1)*(b-a)/(N-1):od:
for k from 1 to N do
f[k]:=(i,j) -> simplify((x[k]-a)^(j-1)):
h[k]:= matrix(1,N,f[k]):
od:
g[1]:=h[1]:
for k from 1 to N-1 do
g[k+1]:=linalg[stackmatrix](g[k],h[k+1]):
od:
Eval(g[N]):end:
where, N is dimension.
```

Procedure 4

```
Hmatrix:= proc (N,a,b)
local i,j,k,f,M,x,h,g;
f:=(i,j) -> simplify((b-a)^(j-1)):
h:= matrix(1,N,f):
eval(h):
end:
where, N is dimension
```

Procedure 5

```
Fmatrix:= proc (N,a,b,f)
local i,j,k,h,g,F;
for k from 1 to N do
h[k]:=a+(k-1)*(b-a)/(N-1):od:
for k from 1 to N do
g[k]:=(i,j) -> simplify(subs(x=h[i],f)):
F:=matrix(N,1,g[k]):
where, N is dimension and f is f(x) function.
```

Procedure 6

```
Answer:= proc (N,A::matrix)
local i,j,k,f,T,C;
```

```
f:=(i,j) -> x^(j-1)/(j-1)!:
T:=matrix(1,N+1,f);C:=multiply(T,A);eval(C):
end:
where, N is dimension.
```

It calculates equalence of A matrix in the Taylor series which was calculated before.

3 Application [1,5,7]

$$\begin{aligned} y_1' &= y_3 - \cos x \\ y_2' &= y_3 - e^x \\ y_3' &= y_1 - y_2 \end{aligned} \tag{12}$$

$$y_1(0) = 1, y_2(0) = 0 \ ve \ y_3(0) = 2.$$

The exact solution of equation is $y_1 = e^x, y_2 = \sin x, y_3 = e^x + \cos x$. Now, let's solve this problem with the mentioned method and Maple procedure.

$$\begin{aligned} y_1' - y_3 &= -\cos x \\ y_2' - y_3 &= -e^x \\ y_3' - y_1 + y_2 &= 0 \end{aligned}$$

$$y_1(0) = 1, y_2(0) = 0 \ ve \ y_3(0) = 2 \quad 0 \leq x \leq 1.$$

We write Taylor collocation points for $N = 2$

$$x_0 = -1, x_1 = 0, x_2 = 1$$

and we can write the system (12) in the matrix form

$$\sum_{i=0}^{m} P_i(t) y^i(t) = f(t)$$

and we have the Maple procedure of the system (12),

$$\begin{aligned} y_1 &= 1 + t + 0.583408022075000088t^2 \\ y_2 &= t + 0.410985157374999965t^2 \\ y_3 &= 2 + t - 0.0862114323500000612t^2. \end{aligned}$$

The results obtained by Taylor collocation method compared with the results of differential transform method [8] and Adomian decomposition method [7, 9] (Fig. 1).

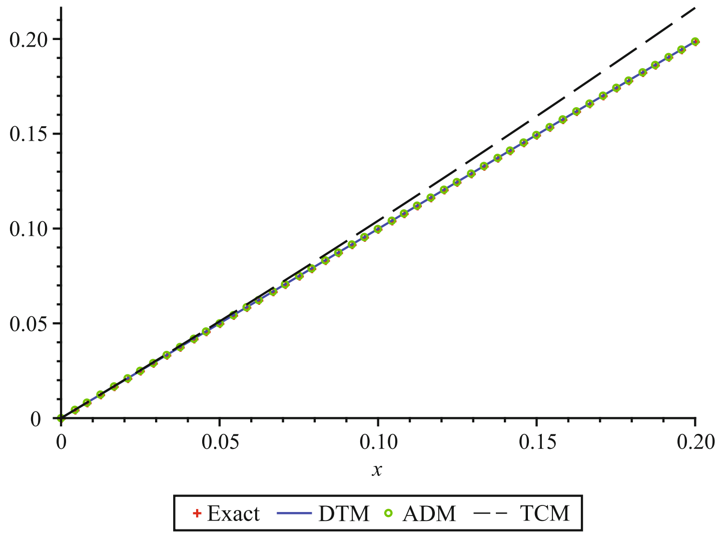


Fig. 1 DTM with 11 steps, ADM 16 steps, TCM 3 steps

4 Conclusion

This study is about the systems of differential equations which don't have analytical solutions or have so difficult and time-consuming solutions. Firstly, we obtain matrix form depending on the values in collocation points of the familiar coefficient functions and unknown function and its derivatives in differential equations, finite Taylor series expansion. Then, the equation is converted a matrix equation with Taylor coefficient by substituting this matrix form. Taylor collocation method can solve only the result matrix equations which corresponding the linear algebraic system. So the solution cannot find for nonlinear systems. But this method gives to approach result to analytical solution in linear equations and it can easily solve with Maple procedures.

References

1. Sezer M., Karamete A., Gulsu M.: Taylor polynomial solutions of systems of linear differential equations with variable coeffiencients. Int. J. Comput. Math. **82**(6), 755–764 (2005)
2. Gulsu M., Sezer M., Güney Z.: Approximate solution of general high-order linear nonhomogeneous difference equations by means of Taylor collocation method. Appl. Math. Comput. **173**, 683–693 (2006)
3. Gulsu, M., Sezer M.: A Taylor polynomial approach for solving differential-difference equations. J. Comput. Appl. Math. **186**, 349–364 (2005)
4. Maple14. www.maplesoft.com
5. Servi S.: On the different appoach numerical solutions for differential equations. M.Sc thesis, Selcuk University (2008, in Turkish)

6. Keskin, Y., Karaoglu, O., Servi, S., Oturanc, G.: The approximate solution of high-order linear fractional differential equations with variable coefficients in terms of generalized taylor polynoms. Math. Comput. Appl. **16**(3), 617–629 (2011)
7. Biazar, J., Babolian, E., Islam, R.: Solution of the system of ordinary differential equations by adomian decomposition method. Appl. Math. Comput. **147**, 713–719 (2004)
8. Kurnaz, A., Oturanç, G.: The differential transform approximation for the system of ordinary differential equation. Int. J. Comput. Math. **82**(6), 709–719 (2005)
9. Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publishers, Boston (1994)

Numerical Study of Convective Heat and Mass Transfer Flow in Channels

G. Sreedevi, D.R.V. Prasada Rao, and R. Raghavendra Rao

Abstract We analyze the combined influence of thermodiffusion and diffusion heat transfer flow of a chemically reacting viscous fluid through a porous medium in a vertical channel under the influence of the transverse magnetic field. The nonlinear coupled equations governing the flow of the heat and mass transfers have been called by using Galerkin finite element analysis with a quadratic approximation function. The velocity, temperature, concentration, and rate of heat and mass transfers are analyzed for the different values of G, M, D^{-1}, N, Sc, S_0, Du, α, and γ. It is found that an increase in S_0 and Du enhances the velocity, temperature, and concentration. An increase in the chemical reaction in γ depreciates a velocity and concentration and enhances a temperature in the degenerating chemical reaction and in the generating chemical reaction case the velocity, concentrations are enhanced and temperature is depreciation.

Keywords Heat and mass transfers • Porous medium • Soret effect and Dufour effect • Chemical reaction

G. Sreedevi (✉) • R.R. Rao
Department of Mathematics, K L University, Green Fields, Vaddeswaram,
Guntur 522502, Andhra Pradesh, India
e-mail: sreedevihari2007@gmail.com; rrrsvu@kluniversity.in

D.R.V.P. Rao
Department of Mathematics, S.K.University, Anantapuramu, Andhra Pradesh, India
e-mail: drv_atp@yahoo.in

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_11,

1 Introduction

Heat and mass transfer problems have assumed greater significance in recent decades due to widespread use and application in industrial segment. Research in magnetohydrodynamics (MHD) viscous flow is pivotal in technological and geothermal applications.

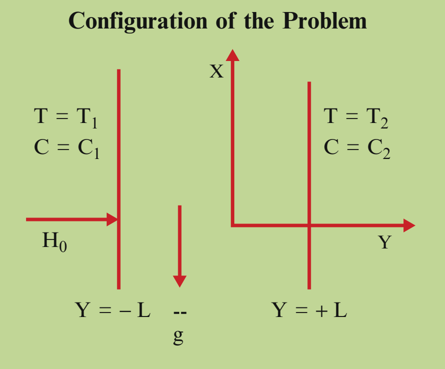
Non-Darcy effects on natural convection in porous media have received a great deal of attention in recent years because of the experiments conducted with several combinations of solids and fluids covering a wide range of governing parameters which indicate that the experimental data for systems other than glass water at low Rayleigh numbers do not agree with theoretical predictions based on the Darcy flow model. This divergence in the heat transfer results has been reviewed in detail by Cheng [1] and Prasad et al. [2] among others. The work of Vafai and Tien [3] was one of the early attempts to account for the boundary and inertia effects in the momentum equation for a porous medium. They found that the momentum boundary layer thickness is of order of $\sqrt{\frac{k}{\varepsilon}}$. Vafai and Thiyagaraja [4] presented analytical solutions for the velocity and temperature fields for the interface region using the Brinkman-Forchheimer-extended Darcy equation. Detailed accounts on non-Darcy convection have been reported in Tien and Hong [5], Prasad et al. [6], and Kalidas and Prasad [7]. Tong and Subramanian [8] and Lauriat and Prasad [9] have studied the viscous effects by using the Brinkman-extended Darcy equations and a numerical study based on the Forchheimer-Brinkman-extended Darcy equation of motion has also been reported by Beckerman et al. [10].

In the above-referred studies, thermal-diffusion and diffusion-thermo effects have been ignored. However, these effects are interesting macroscopically physical phenomenon in fluid mechanics. The heat and mass transfers simultaneously affect each other which creates cross-diffusion. The heat transfer caused by concentration gradient is called the diffusion-thermo or Dufour effect. On the other hand, mass transfer caused by temperature gradients is called Soret or thermal-diffusion effect. Thus Soret effect refers to species differentiation developing in an initial homogeneous mixture submitted to a thermal gradient and the Dufour effect refers to the heat flux produced by a concentration gradient. Most of the studies were based on Soret and Dufour effects on free and mixed convection boundary layer flow in a porous medium [11–14]. Studies were also conducted in non-Darcy convective heat and mass transfer flow through a porous medium [15, 16].

Considering the above, we have attempted to study the combined influence of thermodiffusion, diffusion-thermo, and chemical reaction effects on non-Darcy convective heat and mass transfer flow of a viscous electrically conducting fluid through a porous medium in a vertical channel in the presence of heat sources. The governing equations of flow and heat and mass transfers are solved by using Galerkin finite element technique with quadratic approximation functions.

2 Formulation of the Problem

We consider a fully developed laminar convective heat and mass transfer flow of a viscous, electrically conducting fluid through a porous medium confined in a vertical channel bounded by flat walls. We choose a Cartesian coordinate system $O(x, y, z)$ with x-axis in the vertical direction and y-axis normal to the walls; the walls are taken at $y = \pm L$. The walls are maintained at constant temperature and concentration. A uniform magnetic field of strength H_o is applied normal to the walls. The temperature gradient in the flow field is sufficient to cause natural convection in the flow field. A constant axial pressure gradient is also imposed so that this resultant flow is a mixed convection flow. The porous medium is assumed to be isotropic and homogeneous with constant porosity and effective thermal diffusivity. The thermophysical properties of porous matrix are also assumed to be constant and Boussinesq approximation is invoked by confining the density variation to the buoyancy term. In the absence of any extraneous force, flow is unidirectional along the x-axis which is assumed to be infinite.


Configuration of the Problem

The momentum, energy, and diffusion equations in the scalar form reduce to

$$-\frac{\partial p}{\partial x}+\left(\frac{\mu}{\delta}\right)\frac{\partial^2 u}{\partial y^2}-\left(\frac{\sigma\mu_e^2 H_o^2}{\rho_o}+\frac{\mu}{k}\right)u-\frac{\rho\delta F}{\sqrt{k}}u^2-\rho g=0 \tag{1}$$

$$\rho_0 C_p u\frac{\partial T}{\partial x}=k_f\frac{\partial^2 T}{\partial y^2}+Q\;+\frac{D_m K_T}{T_m}\frac{\partial^2 c}{\partial y^2} \tag{2}$$

$$u\frac{\partial C}{\partial x}=D_1\frac{\partial^2 C}{\partial y^2}-k_1 C\;+\frac{D_m K_T}{C_s C_p}\frac{\partial^2 T}{\partial y^2}. \tag{3}$$

The boundary conditions are

$$\begin{array}{llll} u=0 \quad , & T = T_1 \quad C = C_1 & \text{on} & y = -L \\ u=0 \quad , & T = T_2 \quad C = C_2 & \text{on} & y = +L. \end{array} \tag{4}$$

The axial temperature and concentration gradients $\frac{\partial T}{\partial x}$ and $\frac{\partial C}{\partial x}$ are assumed to be constant, say, A and B respectively. We define the following nondimensional variables as

$$\begin{array}{l} u' = \frac{u}{(\nu/L)}\,,\ (x', y') = (x, y)/L\ ,\quad p' = \frac{p\delta}{(\rho\nu^2/L^2)} \\ \theta = \frac{T-T_2}{T_1-T_2}\quad ,\quad C' = \frac{C-C_2}{C_1-C_2}\quad . \end{array} \tag{5}$$

Introducing these nondimensional variables the governing equations in the dimensionless form are reduced to (on dropping the dashes)

$$\frac{d^2u}{dy^2} = \pi + \delta(M^2 + D^{-1})u - \delta\, G(\theta\ +\ NC) - \delta^2\Delta u^2 \tag{6}$$

$$\frac{d^2\theta}{dy^2} = (PN_T)u - Du\frac{d^2C}{dy^2} \tag{7}$$

$$\frac{d^2C}{dy^2} - \gamma\, C = (Sc\, N_C)u + \frac{ScS_0}{N}\frac{d^2\theta}{dy^2}, \tag{8}$$

where

$\Delta = FD^{-1/2}$ (inertia or Forchheimer parameter)

$G = \frac{\beta g(T_1-T_2)L^3}{\nu^2}$ (Grashof number)

$M^2 = \frac{\sigma\mu_e^2 H_o^2 L^2}{\nu^2}$ (Hartmann number)

$Sc = \frac{\nu}{D_1}$ (Schmidt number)

$N = \frac{\beta^{\bullet}(C_1-C_2)}{\beta(T_1-T_2)}$ (buoyancy ratio)

$P = \frac{\mu C_p}{k_f}$ (Prandtl number)

$\alpha = \frac{QL^2}{\Delta T k_f}$ (heat source parameter)

$\gamma = \frac{k_1 L^2}{D_1}$ (chemical reaction parameter)

$N_T = \frac{AL}{(T_1-T_2)}$ (nondimensional temperature gradient)

$N_c = \frac{BL}{(C_1-C_2)}$ (nondimensional concentration gradient)

$S_0 = \frac{K_T \Delta T}{Tm\Delta C}$ (Soret parameter)

$Du = \frac{D_m K_T \Delta C}{C_s k_f \Delta T}$ (Dufour parameter)

The corresponding boundary conditions are

$$\begin{array}{l} u=0\ ,\quad \theta = 1\ ,\ C = 1\ \text{ on } y = -1 \\ u=0\ ,\quad \theta = 0\ ,\ C = 0\ \text{ on } y = +1. \end{array} \tag{9}$$

3 Method of Solution

Using finite element technique, these differential equations are solved with the corresponding boundary conditions and we assume that if u^i, c^i, θ^i are the approximations of u, C, and θ we define the errors (residual) E_u^i, E_c^i, E_θ^i as

$$E_u^i = \frac{d}{d\eta}\left(\frac{du^i}{d\eta}\right) - M_1^2 u^i + \delta^2 \mathrm{A}(u^i)^2 - \delta G(\theta^i + NC^i) \tag{10}$$

$$E_c^i = \frac{d}{dy}\left(\frac{dC^i}{dy}\right) - \gamma C^i + \frac{ScS_0}{N}\frac{d}{dy}\left(\frac{d\theta^i}{dy}\right) - ScN_c u^i \tag{11}$$

$$E_\theta^i = \frac{d}{dy}\left(\frac{d\theta^i}{dy}\right) - P_1 N_T u^i + DuN_2 \frac{d}{dy}\left(\frac{dC^i}{dy}\right) \tag{12}$$

$$\text{where } u^i = \sum_{k=1}^{3} u_k \psi_k ; C^i = \sum_{k=1}^{3} C_k \psi_k ; \theta^i = \sum_{k=1}^{3} \theta_k \psi_k . \tag{13}$$

In order to predict the heat and mass transfer behavior in the porous medium, Eqs. (5)–(8) are solved. A simple 3-noded line element is considered. u, C, and θ vary inside the element and can be expressed as

$$u = u_1 N_1 + u_2 N_2 + u_3 N_3$$
$$C = C_1 N_1 + C_2 N_2 + C_3 N_3$$
$$\theta = \theta_1 N_1 + \theta_2 N_2 + \theta_3 N_3 .$$

Galerkin's method is used to convert the partial differential equations (10)–(13) into matrix form of equations that results into 3 × 3 local stiffness matrices. All these local matrices are assembled in a global matrix by substituting the global nodal values of order I.

4 Discussion of Results

We investigate the combined influence of thermodiffusion and diffusion-thermo on hydromagnetic convective heat and mass transfer flow chemically reacting viscous fluid flow in a porous medium in a vertical channel. Cool walls are maintained at concentric and constant temperature and concentration, in the presence of temperature-dependent heat sources. The nonlinear coupled equations governing flow and heat and mass transfer have been solved by employing Galerkin finite element technique with quadratic approximation function. The analysis has been carried out with Prandtl number 0.71.

Fig. 1 Variation of u with S_0. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $Du = 0.1$, $\alpha = 0.51$, $\gamma = 0.5$

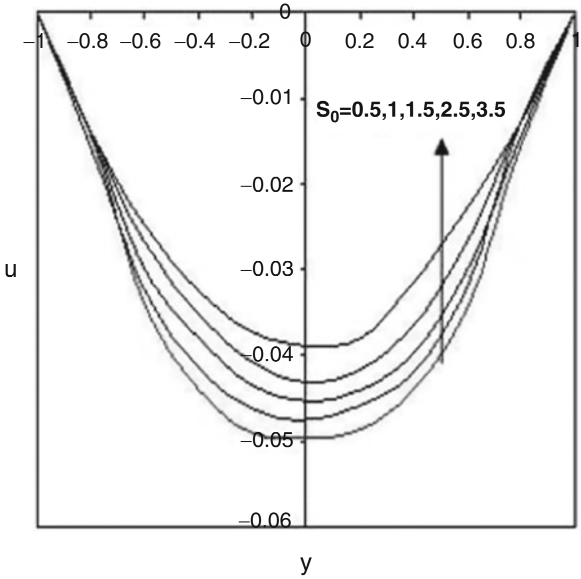
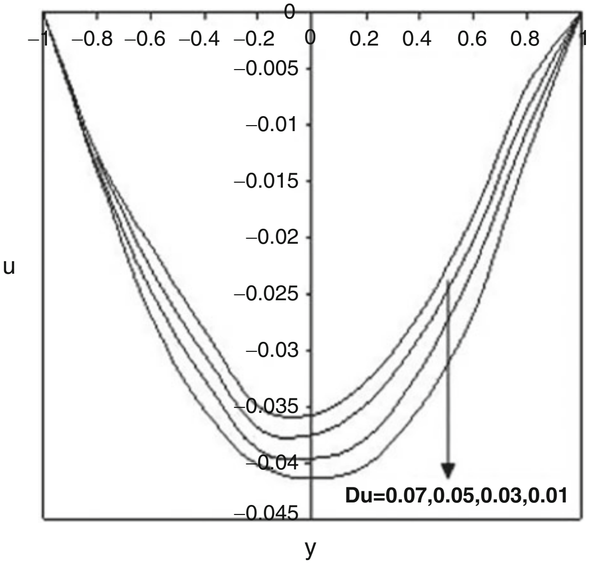


Fig. 2 Variation of u with Du. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $\alpha = 0.51$, $\gamma = 0.5$



The velocity distribution u is shown in Figs. 1, 2, and 3 for different values of S_0, Du, and γ. Its actual axial flow is in the vertically downward direction and hence $u > 0$ represents the reversal flow. The effect of thermodiffusion on u is observed in Fig. 1. It can be seen from the profiles that $|u|$ experiences an enhancement with increase in S_0.

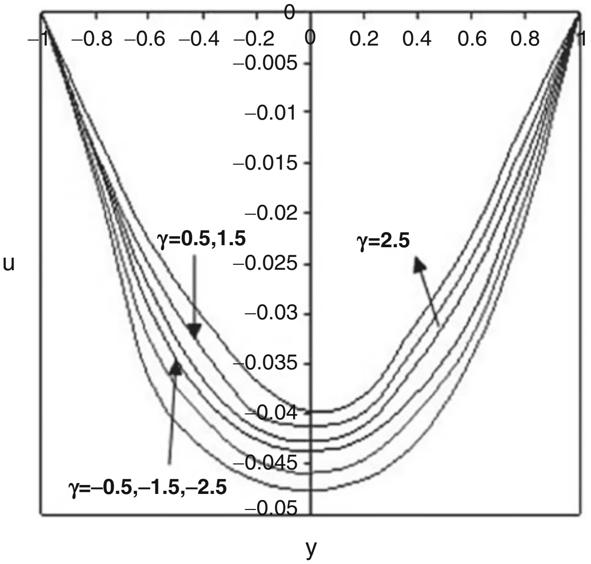


Fig. 3 Variation of u with γ. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $Du = 0.1$, $\alpha = 0.51$

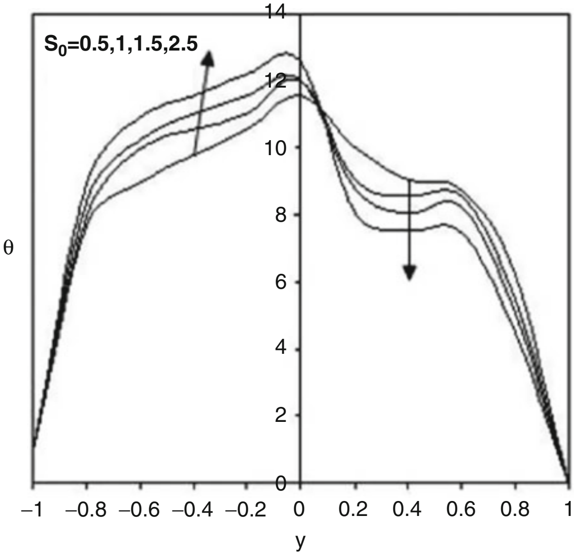


Fig. 4 Variation of θ with S_0. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $Du = 0.1$, $\alpha = 0.51$, $\gamma = 0.5$

Figure 2 represents u with Dufour parameter. It is found that $|u|$decreases with increase in Du in the entire flow region. Thus the higher the diffusion-thermo effects, the smaller the magnitude of u. It is found that $|u|$enhances with the increase in the chemical reaction parameter γ <1.5 and enhances with higher γ >= 2.5, while it enhances with $|\gamma|$decreasing at 1.5 (Fig. 3).

It is observed from the profiles that the higher the diffusion effects (S_0), the larger the actual temperature in the flow region (Fig. 4). Figure 5 represents the Dufour

Fig. 5 Variation of θ with Du. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $\alpha = 0.51$, $\gamma = 0.5$

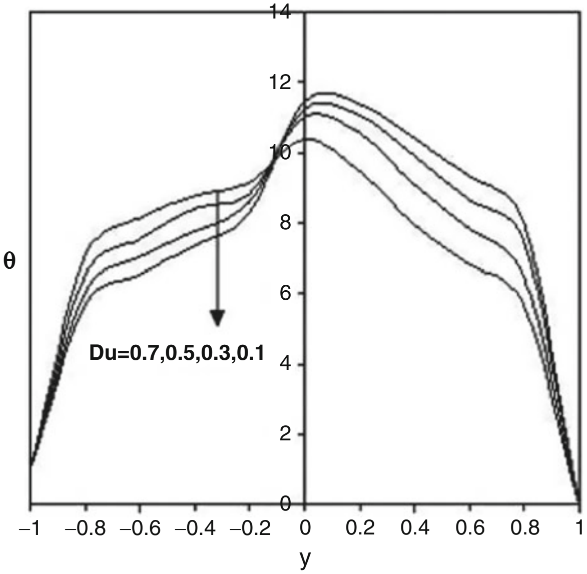
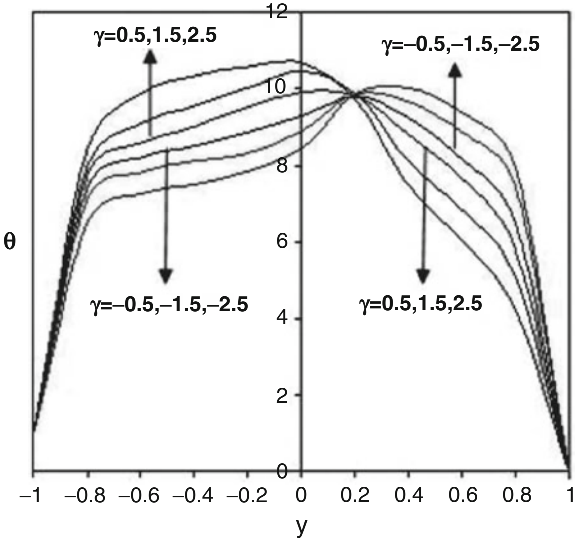


Fig. 6 Variation of θ with γ. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $Du = 0.1$, $\alpha = 0.51$



parameter represents θ. We notice a depreciation in the actual temperature with increase in Du; thus the higher the diffusion-thermo effects, the larger the actual temperature.

Figure 6 represents θ with chemical reaction parameter γ. It is observed that the actual temperature is enhanced in the degenerating chemical reaction case (γ >0) and depreciates in the generating chemical reaction case (γ <0).

Fig. 7 Variation of C with S_0. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $Du = 0.1$, $\alpha = 0.51$, $\gamma = 0.5$

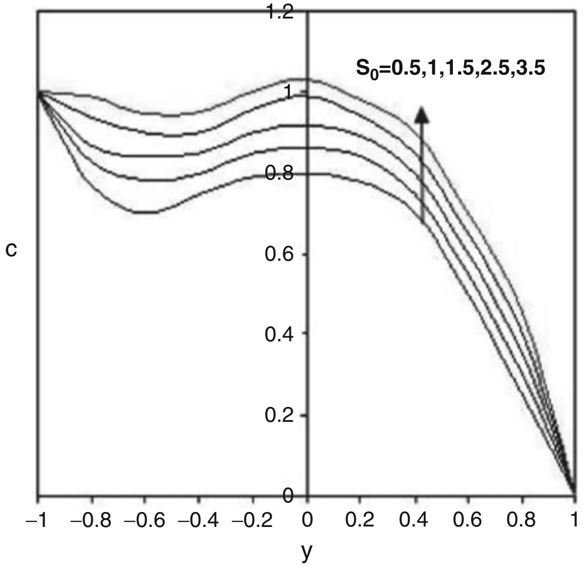
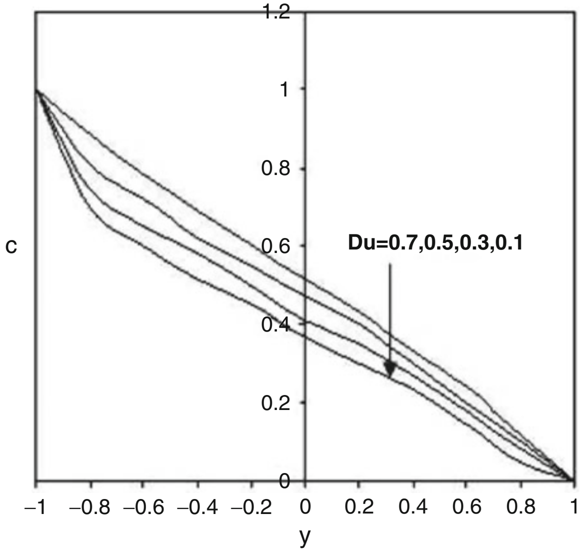


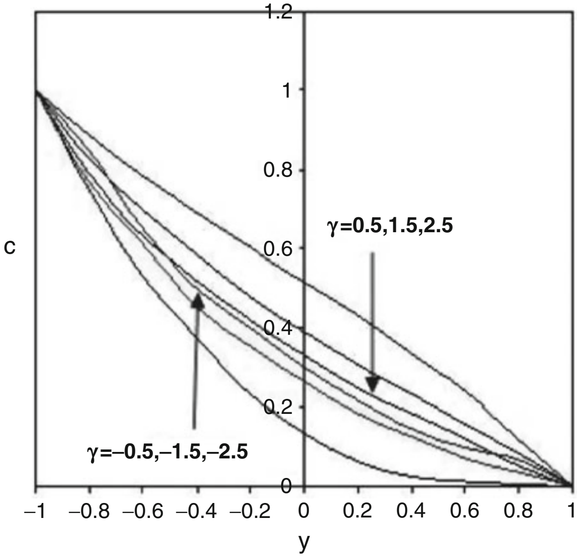
Fig. 8 Variation of C with Du. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $\alpha = 0.51$, $\gamma = 0.5$



The variations of "C" with Soret parameter S_0 show that the actual concentration is enhanced with $S_0 \leq 1.0$ and depreciates with $S_0 \geq 1.5$. We notice an enhancement in "C" (Fig. 7). A variation of "C" with Dufour parameter Du is exhibited in Fig. 8.

As chemical reaction parameter γ depreciates, the actual concentration "C" decreases and as the γ enhances, the variation of concentration "C" enhances (Fig. 9).

Fig. 9 Variation of C with γ. $G = 10^3$, $M = 2$, $N = 1$, $D^{-1} = 10^2$, $Sc = 1.3$, $S_0 = 0.01$, $Du = 0.1$, $\alpha = 0.51$



5 Conclusions

Numerical evaluations were performed and graphical results were obtained to illustrate the details of the flow and heat and mass transfer characteristics and their dependence on some physical parameters. The key findings are summarized below:

- The variation of u with Soret parameter S_0 shows $|u|$ experiences an enhancement with increase in (S_0). The higher the (S_0), the larger the actual temperature in the flow region. The variations of "C" with Soret parameter S_0 show the actual concentration enhances with increase in S_0.
- With increase in chemical reaction parameter γ, $|u|$gets enhanced. The actual temperature enhances in the degenerating chemical reaction case (γ >0) and depreciates in the generating chemical reaction case (γ <0), while the "C" depreciates, with increase in chemical reaction parameter γ.
- The higher the diffusion-thermo effects Du, the smaller the $|u|$and the larger the actual temperature. With enhancement in "Du," the concentration increases.

References

1. Cheng, P.: Heat transfer in geothermal systems. In: Advances in Heat Transfer, vol. 4, pp. 1–105. Academic, New York (1978)
2. Prasad, V., Tuntomo, A.: Inertia effects on natural convection in a vertical porous cavity. Numer. Heat Tran. **11**, 295–320 (1987)

3. Vafai K., Tien, C.L.: Boundary and inertia effects on flow and heat transfer in porous media. Int. J. Heat Mass Tran. **24**, 195–203 (1981)
4. Vafai, K., Thyagaraju, R.: Analysis of flow and heat transfer at the interface region of a porous medium. Int. J. Heat Mass Tran. **30**, 1391–1405 (1987)
5. Tien, C.L., Hong, J.T.: Natural convection in porous media under non-Darcian and non-uniform permeability conditions. In: Kakac et al. (eds.) Natural Convection. Hemisphere, Washington, DC (1985)
6. Prasad, V., Kulacki, F.A., Keyhani, M.: Natural convection in a porous medium. J. Fluid Mech. **150**, 89–119 (1985)
7. Kalidas, N., Prasad Benard V.: International symposium of convection in porous media, non-Darcy effects. In: Proceedings of 25th National Heat Transfer Conference, vol. 1, pp. 593–604 (1988)
8. Tong, T.L., Subramanian, E.: A boundary layer analysis for natural correction in porous enclosures: use of the Brinkman-extended Darcy model. Int. J. Heat Mass Tran. **28**, 563–571 (1985)
9. Laurait, G., Prasad, V.: Natural convection in a vertical porous cavity: a numerical study of Brinkman-extended Darcy formulation. J. Heat Tran. **109**, 688–696 (1987)
10. Beckermann, C., Viskanta, R., Ramadhyani, S.: A numerical study of non-Darcian natural convection in a vertical enclosure filled with a porous medium. Numer. Heat Tran. **10**, 557–570 (1986)
11. Kafoussias, N.G., Williams, E.M.: Thermal-diffusion and diffusion-thermo effects on free convective and mass transfer boundary layer flow with temperature dependent viscosity. Int. J. Eng. Sci. **33**, 1369–1376 (1995)
12. Anghel, M., Takhar, H.S., Pop, I.: Dufour and Soret effects on free convection boundary layer over a vertical surface embedded in a porous medium. J. Heat Mass Tran. **43**, 1265–1274 (2000)
13. Postelnicu, A.: Influence of a magnetic field on heat and mass transfer by natural convection from vertical surfaces in porous media considering Soret and Dufour effects. Int. J. Heat Mass Tran. **47**, 1467–1472 (2004)
14. Alam, M.S., Rahman, M.M.: Dufour and Soret effects on mixed convection flow past a vertical porous flat plate with variable suction. Nonlinear Anal. Model. Control **11**, 3–12 (2006)
15. Nagaleelakumari, S.: Combined influence of chemical reaction and Soret effect on hydrodynamic non-Darcy convective heat and mass transfer flow through a porous medium in a vertical channel with heat generating sources. Ph.D thesis, S P Mahila University, Tirupati (2012)
16. Muralidhar, P.: Effect of chemical reaction and thermo diffusion on convective heat and mass transfer flow of viscous fluid through a porous medium in a vertical channel. Ph.D thesis, Andhra University, Vishakapatnam (2012)

A Parameter Uniform Method for an Initial Value Problem for a System of Singularly Perturbed Delay Differential Equations

Shivaranjani Nagarajan, Ramanujam Narasimhan, J.J.H. Miller, and Valarmathi Sigamani

Abstract In this paper an initial value problem for a coupled system of two singularly perturbed first-order delay differential equations is considered on the interval (0,2]. The components of the solution of this system exhibit initial layers at 0 and interior layers at 1. A numerical method composed of a classical finite difference scheme on a piecewise uniform Shishkin mesh is suggested. This method is proved to be first-order convergent in the maximum norm uniformly in the perturbation parameters. A numerical illustration is provided to support the theory.

Keywords Singular perturbation problems • Boundary layers • Delay differential equations • Finite difference schemes • Shishkin mesh • Parameter uniform convergence

1 Introduction

Singularly perturbed delay differential equations arise in the mathematical modelling of various phenomena of practical importance, for example, in population dynamics, control theory, potential in models for neuron, optical bistable devices,

S. Nagarajan • V. Sigamani
Department of Mathematics, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India
e-mail: shivaranjaninagarajan@gmail.com; valarmathi07@gmail.com

R. Narasimhan
Bharathidasan University, Tiruchirappalli, Tamil Nadu, India
e-mail: matrambdu@gmail.com

J.J.H. Miller (✉)
Institute for Numerical Computation and Analysis, Dublin, Ireland
e-mail: jm@incaireland.org

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_12,

human pupil-light reflex and many other problems in applied mathematics. Models of hospital-acquired infections involving systems of singularly perturbed delay differential equations are described in [1].

Singularly perturbed delay differential equations with small shifts are dealt with in [2]. In [3], a hybrid finite difference scheme is suggested for an initial value problem for scalar delay-differential equation and the method is proved to be second-order convergent. In [4], a numerical method composed of a fitted operator on an equidistant mesh to solve the above problem is suggested. The method is proved to be first-order convergent, uniformly with respect to the perturbation parameter. Related works are found in [5, 6].

On the other hand, very few or no works on systems of singularly perturbed delay differential equations are reported in the literature. In this paper, the following coupled system of two singularly perturbed delay differential equations of first order is considered:

$$\vec{L}\vec{u} \;=\; E\vec{u}'(x) + A(x)\vec{u}(x) + B(x)\vec{u}(x-1) = \vec{f}(x) \;\text{ on }\; (0,2], \tag{1}$$

$$\vec{u} \;= \vec{\phi} \;\text{ on }\; [-1,0]. \tag{2}$$

For all $x \in [0,2]$, $\vec{u}(x) = (u_1(x), u_2(x))^T$ and $\vec{f}(x) = (f_1(x), f_2(x))^T$. E, $A(x)$ and $B(x)$ are 2×2 matrices. $E = \operatorname{diag}(\vec{\varepsilon})$, $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2)$ with $0 < \varepsilon_1 \le \varepsilon_2 \le 1$, $B(x) = \operatorname{diag}(\vec{b})$, $\vec{b} = (b_1(x), b_2(x))$. For all $x \in [0,2]$ it is assumed that the components $a_{ij}(x)$ and $b_i(x)$ of $A(x)$ and $B(x)$, respectively, satisfy

$$b_i,\; a_{ij} \;\le 0 \;\text{ for }\; 1 \le i \ne j \le 2 \;\text{ and }\; a_{ii} > \sum_{i \ne j} |a_{ij}(x) + b_i(x)| \tag{3}$$

and

$$0 < \alpha < \min_{\substack{x\in[0,2]\\ 1\le i\le 2}} \left(\sum_{j=1}^{2} a_{ij}(x) + b_i(x) \right), \text{for some }\; \alpha. \tag{4}$$

Further, the functions $\vec{f}(x), a_{ij}(x), b_i(x), 1 \le i, j \le 2$ are assumed to be in $C^{(2)}([0,2])$. The above assumptions ensure that $\vec{u} \in C^0([0,2]) \cup C^1((0,2])$.

The problem (1) can be rewritten as

$$\vec{L}_1\vec{u} = E\vec{u}'(x) + A(x)\vec{u}(x) = \vec{f}(x) - B(x)\vec{\phi}(x-1) = \vec{g}(x) \text{ on }\; (0,1] \tag{5}$$

$$\vec{L}_2\vec{u} = E\vec{u}'(x) + A(x)\vec{u}(x) + B(x)\vec{u}(x-1) = \vec{f}(x) \text{ on }\; (1,2]. \tag{6}$$

The reduced problem corresponding to (5), (6) is given by

$$A(x)\vec{u}_0(x) = \vec{f}(x) - B(x)\vec{\phi}(x-1) \text{ on }\; (0,1] \tag{7}$$

$$A(x)\vec{u}_0(x) + B(x)\vec{u}_0(x-1) = \vec{f}(x) \text{ on }\; (1,2]. \tag{8}$$

For any vector-valued function $\vec{y}$ on $[0,2]$ the following norms are introduced:

$\| \vec{y}(x) \|= \max_i |y_i(x)|$ and $\| \vec{y} \|= \sup\{\| \vec{y}(x) \|: x \in [1,2]\}$. Throughout the paper C denotes a generic positive constant, which is independent of x and of all singular perturbation and discretisation parameters. Furthermore, inequalities between vectors are understood in the componentwise sense.

The plan of the paper is as follows. In Sect. 2, estimates of the analytical behaviour of the exact solution are presented. In Sects. 3 and 4, the problem is discretised using a Shishkin mesh, which is piecewise uniform, and the numerical analysis is presented. In Sect. 5, the parameter-uniform error of this discretisation is estimated in the maximum norm, and a numerical illustration is presented in Sect. 6.

2 Analytical Results

The operator $\vec{L}$ satisfies the following maximum principle.

Lemma 1. *Let $\vec{\psi}$ be any function in the domain of $\vec{L}$ such that $\vec{\psi}(0) \geq \vec{0}$. Then $\vec{L}\vec{\psi} \geq \vec{0}$ on $(0,2]$ implies $\vec{\psi} \geq \vec{0}$ on $[0,2]$.*

Proof. Let $\psi_{i^*}(x^*) = \min_{i,x}(\psi_i(x))$. Suppose $\psi_{i^*}(x^*) < 0$. Then, $\psi'_{i^*}(x^*) \leq 0$ and

$$\begin{aligned}(\vec{L}\vec{\psi})_{(i^*)}(x^*) &= \varepsilon_{i^*}\psi'_{i^*}(x^*) + \sum_{j=1}^{2} a_{i^*j}\psi_j(x^*) + b_{i^*}\psi_{i^*}(x^*-1)\\ &\leq \sum_{j=1}^{2} a_{i^*j}\psi_j(x^*) + b_{i^*}\psi_{i^*}(x^*)\\ &= (a_{i^*i^*} + b_{i^*})\psi_{i^*}(x^*) + a_{i^*j}\psi_j(x^*) \quad j = 1 \text{ or } 2\\ &\leq (a_{i^*i^*} + a_{i^*j} + b_{i^*})\psi_{i^*}(x^*)\\ &< 0,\end{aligned}$$

which is a contradiction. Hence our assumption is wrong. Therefore, $\psi_{i^*}(x^*) \geq 0$, which proves the lemma. □

Lemma 2. *If $\vec{\psi}$ is any function in the domain of $\vec{L}$, then $||\vec{\psi}|| \leq C \max\{||\vec{\psi}(0)||, \frac{1}{\alpha}||\vec{L}\vec{\psi}||\}$.*

Proof. Consider the barrier functions, $\vec{\theta}^{\pm} = CM \pm \vec{\psi}(x)$, where $M = \max\{||\vec{\psi}(0)||, \frac{1}{\alpha}||\vec{L}\vec{\psi}||\}$. $\vec{\theta}^{\pm}(0) \geq \vec{0}$, for proper choice of C.

$$\begin{aligned}(\vec{L}\vec{\theta}^{\pm})_i(x) &= \left(\sum a_{ij} + b_i\right)(CM) \pm (\vec{L}\vec{\psi})_i(x)\\ &\geq \alpha(\tfrac{1}{\alpha})||\vec{L}\vec{\psi}|| \pm (\vec{L}\vec{\psi})_i(x) \geq \vec{0}.\end{aligned}$$

Hence, $\vec{\theta}^{\pm}(x) \geq 0$.

A Shishkin decomposition of $\vec{u}$ is given by $\vec{u} = \vec{v} + \vec{w}$ where $\vec{v} = (v_1, v_2)^T$ is the solution of

$$\vec{L}_1\vec{v} = E\vec{v}'(x) + A(x)\vec{v}(x) = \vec{f}(x) - B(x)\vec{\phi}(x-1) = \vec{g}(x) \text{ on } (0,1] \quad (9)$$

$$\vec{L}_2\vec{v} = E\vec{v}'(x) + A(x)\vec{v}(x) + B(x)\vec{v}(x-1) = \vec{f}(x) \text{ on } (1,2] \quad (10)$$

$\vec{v}(0) = A^{-1}(0)(\vec{f}(0) - B(0)\phi(-1))$ and $\vec{w} = (w_1, w_2)^T$ satisfies $\vec{L}_1\vec{w} = \vec{0}$ for $x \in (0,1]$ and $\vec{L}_2\vec{w} = \vec{0}$ for $x \in (1,2]$ with $\vec{w}(0) = \vec{u}(0) - \vec{v}(0)$. Here, $\vec{v}$ is called the smooth component of $\vec{u}$ and $\vec{w}$, the singular component of $\vec{u}$. □

Lemma 3. *For $i = 1,2$, there exists a constant C such that $||v_i^{(k)}|| \leq C$for $k = 0,1$ and $||v_i''|| \leq C\varepsilon_i^{-1}$.*

Proof. From (9), it is clear that, for $x \in (0,1]$, the bounds on $\vec{v}$ are the same as in [7]. For $x \in (1,2]$, $\vec{L}_2\vec{v}(x) = E\vec{v}'(x)+A(x)\vec{v}(x)+B(x)\vec{v}(x-1) = \vec{f}(x)$ or $\vec{L}_1\vec{v}(x) = \vec{f}(x) - B(x)\vec{v}(x-1)$. Hence $\parallel (\vec{L}_1\vec{v})(x) \parallel \leq C$ and $\parallel \vec{v}(1) \parallel \leq C$. Hence applying the stability result for $\vec{L}_1$ [7] on the domain $[1,2]$, $||\vec{v}|| \leq C$.

From (10), $\vec{v}'(1) = \vec{f}(1) - \vec{u}_0(1) = \vec{0}$. Differentiating (10) once gives

$$\vec{L}_1\vec{v}' = E(\vec{v}')'(x)+A(x)\vec{v}'(x) = \vec{f}'(x)-A'(x)\vec{v}(x)-B'(x)\vec{v}(x-1)-B(x)\vec{v}'(x-1). \quad (11)$$

Thus, $||\vec{L}_1\vec{v}'|| \leq C$. Hence using stability result for the operator L_1, $||\vec{v}'|| \leq C$ on $[1,2]$. Further, from (11), it is not hard to derive that $\parallel v_i'' \parallel \leq C\varepsilon_i^{-1}, i = 1,2$ on $[1,2]$. Combining with the results for $\vec{v}$ in $[0,1]$, the required bounds of $\vec{v}$ in the whole of $[0,2]$ are obtained. □

From [2] it is seen that the $u_i, i = 1,2$, have exponential layers represented by $e^{\alpha x/\varepsilon_i}$ and $e^{\alpha(x-1)/\varepsilon_i}$. Define functions $B_{1,i}$ and $B_{2,i}, i = 1,2$ by $B_{1,i}(x) = e^{-\alpha x/\varepsilon_i}$, on $[0,2]$ and $B_{2,i}(x) = e^{-(x-1)\alpha/\varepsilon_i}$, on $[1,2]$.

The bounds of the singular component $\vec{w}$ are contained in

Lemma 4. *Let $A(x), B(x)$ satisfy (3) and (4). Then, for each $i = 1,2$, there exists a constant C, such that, for $x \in [0,1]$,*

$$|w_i(x)| \leq CB_{1,2}(x), \quad \left|w_i^{(l)}(x)\right| \leq C\left[\frac{B_{1,1}(x)}{\varepsilon_1^l} + \frac{B_{1,2}(x)}{\varepsilon_2^l}\right], l = 1,2$$

and for $x \in [1,2]$

$$|w_i(x)| \leq CB_{2,2}(x), \quad \left|w_i(x)^{(l)}\right| \leq C\left[\frac{B_{2,1}(x)}{\varepsilon_1^l} + \frac{B_{2,2}(x)}{\varepsilon_2^l}\right], l = 1,2.$$

Proof. For $x \in [0, 1]$, the bounds for $\vec{w}$ are the same as in [7]. For $x \in (1, 2]$,

$$\vec{L}_2\vec{w}(x) := E\vec{w}'(x) + A(x)\vec{w}(x) + B(x)\vec{w}(x-1) = \vec{0}$$

or

$$\vec{L}_1\vec{w}(x) := E\vec{w}'(x) + A(x)\vec{w}(x) = -B(x)\vec{w}(x-1)$$

which implies

$$\parallel \vec{L}_1\vec{w}(x) \parallel \leq CB_{1,2}(x-1).$$

Construct the barrier function

$$\begin{aligned}
\vec{\psi}^{\pm}(x) &= C_1B_{2,2}(x)\vec{e} \pm \vec{w}(x)\\
(\vec{L}_1\vec{\psi}^{\pm}(x))_1 &= C_1\left(\varepsilon_1\left(\frac{-\alpha}{\varepsilon_2}\right)B_{2,2}(x) + a_{11}(x)B_{2,2}(x) + a_{12}(x)B_{2,2}(x)\right) \pm \vec{L}_1\vec{w}(x)\\
&\geq C_1(-\alpha + a_{11}(x) + a_{12}(x))B_{2,2}(x) \pm CB_{1,2}(x-1)\\
&\geq 0, \quad \text{choosing } C_1 \text{ sufficiently large.}
\end{aligned}$$

Similarly, $(\vec{L}_1\vec{\psi}^{\pm}(x))_2 \geq 0$, for sufficiently large C_1. Also, $\vec{\psi}^{\pm}(1) \geq 0$. Hence by maximum principle for the operator $\vec{L}_1$, (in [7]), we have $\vec{\psi}^{\pm}(x) \geq 0, \quad x \in [1, 2]$. Hence for $x \in [1, 2]$, $|\vec{w}(x)| \leq CB_{2,2}(x)$.

From the defining equation for w_1,

$$\begin{aligned}
\varepsilon_1 w_1'(x) &= -\sum_{j=1}^{2} a_{ij}(x)w_j(x) - b_i(x)w_1(x-1)\\
|w_1'(x)| &\leq C\varepsilon_1^{-1}(B_{2,2}(x)) + (B_{1,2}(x-1))\\
&\leq C\varepsilon_1^{-1}(B_{2,2}(x)).
\end{aligned}$$

Thus, $|w_1'(x)| \leq C\varepsilon_1^{-1}(B_{2,2}(x))$. Similarly for the singular component $w_2(x)$, $|w_2'(x)| \leq C\varepsilon_2^{-1}B_{2,2}(x)$.

To obtain parameter uniform convergence of the method, we require sharper estimates for the derivatives of $\vec{w}(x)$.

To find sharper estimates of $w_1'(x)$, consider the equation

$$L_{2,1}w_1(x) = \varepsilon_1 w_1'(x) + a_{11}(x)w_1(x) = a_{12}(x)w_2(x) - b_1(x)w_1(x-1)$$

and hence

$$\begin{aligned}
L_{2,1}w_1'(x) &= -(a_{12}(x)w_2(x))' - (b_1(x)w_1(x-1))' - a_{11}'w_1(x)\\
|L_{2,1}w_1'(x)| &\leq C[\varepsilon_2^{-1}B_{2,2}(x) + \varepsilon_1^{-1}B_{1,1}(x-1) + \varepsilon_2^{-1}B_{1,2}(x-1)]\\
|L_{2,1}w_1'(x)| &\leq C[\varepsilon_1^{-1}B_{2,1}(x) + \varepsilon_2^{-1}B_{2,2}(x)], \text{since } B_{1,2}(x-1) = B_{2,2}(x).
\end{aligned}$$

Define the barrier functions

$$\psi^{\pm}(x) = C(\varepsilon_1^{-1}B_{2,1}(x) + \varepsilon_2^{-1}B_{2,2}(x)) \pm w'_{2,1}(x).$$

We have $\psi^{\pm}(1) \geq 0$

$$\begin{aligned} L_{2,1}\psi^{\pm}(x) &= C\{\varepsilon_1(\tfrac{-\alpha}{\varepsilon_1})\varepsilon_1^{-1}B_{2,1}(x)+\varepsilon_1(\tfrac{-\alpha}{\varepsilon_2})\varepsilon_2^{-1}B_{2,2}(x)\} \\ &\quad + a_{11}(x)\{C(\varepsilon_1^{-1}B_{2,1}(x)+\varepsilon_2^{-1}B_{2,2}(x))\} \pm |L_{2,1}w'_{2,1}(x)| \\ &\geq C(-\alpha + a_{11})[\varepsilon_1^{-1}B_{2,1}(x)+\varepsilon_2^{-1}B_{2,2}(x)] \pm C[\varepsilon_1^{-1}B_{2,1}(x)+\varepsilon_2^{-1}B_{2,2}(x)] \\ &\geq 0. \end{aligned}$$

Applying the maximum principle for the scalar operator $L_{2,1}$ (in [8]), the required bounds of $w'_1(x)$ follow.

Differentiating $(\vec{L}_2\vec{w})_1 = 0$ and $(\vec{L}_2\vec{w})_2 = 0$ once and using the estimates of $w'_1(x)$ and $w'_2(x)$, we have

$$\begin{aligned} |w''_1(x)| &\leq C\varepsilon_1^{-1}[\varepsilon_1^{-1}B_{2,1}(x) + \varepsilon_2^{-1}B_{2,2}(x)], \\ |w''_2(x)| &\leq C\varepsilon_2^{-1}[\varepsilon_1^{-1}B_{2,1}(x) + \varepsilon_2^{-1}B_{2,2}(x)]. \end{aligned}$$

□

Remark. The unique point x^* in $(0, 1]$, such that $\varepsilon_1^{-1}B_{1,1}(x^*) = \varepsilon_2^{-1}B_{1,2}(x^*)$, $1 + x^* \in (1, 2]$ and $\varepsilon_1^{-1}B_{2,1}(1 + x^*) = \varepsilon_2^{-1}B_{2,2}(1 + x^*)$ is introduced, which leads to the following novel estimates for the derivatives of the singular components. In [9], Linss et al. use a single point of this kind.

Lemma 5. *Suppose that* $\varepsilon_2 \in \left(2\epsilon_1, \frac{\alpha}{2}\right)$ *. Then, there are functions*

$$w_{1,1}(x), \quad w_{1,2}(x) \quad w_{2,1}(x), \quad w_{2,2}(x)$$

such that

$$w_1(x) = w_{1,1}(x) + w_{1,2}(x), \quad w_2(x) = w_{2,1}(x) + w_{2,2}(x)$$

and

$$|w'_{1,1}(x)| \leq C\varepsilon_1^{-1}B_{1,1}(x), \quad |w''_{1,2}(x)| \leq C\varepsilon_1^{-1}\varepsilon_2^{-1}B_{1,2}(x)$$

$$|w'_{2,1}(x)| \leq C\varepsilon_2^{-1}B_{1,1}(x), \quad |w''_{2,2}(x)| \leq C\varepsilon_2^{-2}B_{1,2}(x), \quad x \in [0, 1]$$

and

$$|w'_{1,1}(x)| \leq C\varepsilon_1^{-1}B_{2,1}(x), \quad |w''_{1,2}(x)| \leq C\varepsilon_1^{-1}\varepsilon_2^{-1}B_{2,2}(x)$$

$$|w'_{2,1}(x)| \leq C\varepsilon_2^{-1}B_{2,1}(x), \quad |w''_{2,2}(x)| \leq C\varepsilon_2^{-2}B_{2,2}(x), \quad x \in [1, 2].$$

Proof. For $x \in [0,1]$, the decompositions of $w_1(x)$ and $w_2(x)$, and hence their respective bounds, follow from [7].

For $x \in [1,2]$, define the function $w_{1,2}(x)$ as follows:

$$w_{1,2}(x) = \begin{cases} w_1(x), & \text{for } x \in [1+x^*, 2] \\ \displaystyle\sum_{k=0}^{2} \frac{(x-(1+x^*))^k}{k!} w_1^{(k)}(1+x^*), & \text{for } x \in [1, 1+x^*) \end{cases}$$

$$w_{1,1}(x) = w_1(x) - w_{1,2}(x).$$

w_2 is similarly decomposed.

Proceeding as in [7], the required bounds of $w_{1,1}(x)$ and $w_{1,2}(x)$ for $x \in [1,2]$ follow. □

3 Shishkin Mesh

Motivated by [8,10], a piecewise uniform Shishkin mesh $\overline{\Omega}^N = \overline{\Omega^{-}}^N \cup \Omega^{+N}$ where $\overline{\Omega^{-}}^N = \{x_j\}_0^{\frac{N}{2}}$ and $\Omega^{+N} = \{x_j\}_{\frac{N}{2}+1}^N$ with N mesh intervals is now constructed on $\overline{\Omega} = [0,2]$ as follows for the case $\varepsilon_1 < \varepsilon_2$. In the case $\varepsilon_1 = \varepsilon_2$ a simpler construction requiring just one parameter τ suffices. The interval $[0,1]$ is subdivided into three subintervals $[0,\tau_1] \cup (\tau_1,\tau_2] \cup (\tau_2,1]$. The parameters τ_r, $r = 1,2$, which determine the points separating the uniform meshes, are defined by $\tau_0 = 0$, $\tau_3 = \frac{1}{2}$,

$$\tau_2 = \min\left\{\frac{1}{2}, \frac{\varepsilon_2}{\alpha}\ln N\right\} \tag{12}$$

and

$$\tau_1 = \min\left\{\frac{\tau_2}{2}, \frac{\varepsilon_1}{\alpha}\ln N\right\}. \tag{13}$$

Clearly

$$0 < \tau_1 < \tau_2 \le \frac{1}{2}.$$

Then, on the subinterval $(\tau_2, 1]$, a uniform mesh with $\frac{N}{4}$ mesh points is placed and on each of the subintervals $(0,\tau_1]$ and $(\tau_1,\tau_2]$, a uniform mesh of $\frac{N}{8}$ mesh points is placed. Similarly, the interval $[1,2]$ is divided into 3 subintervals $[1, 1+\tau_1]$, $(1+\tau_1, 1+\tau_2]$, $(1+\tau_2, 2]$ having the same number of mesh intervals as in $[0,1]$.

Note that, when both of the parameters $\tau_r, r = 1, 2,$ take on their lefthand value, the Shishkin mesh becomes a classical uniform mesh on $[0, 2]$.

4 The Discrete Problem

The IVP (1), (2) is discretised using the backward Euler scheme applied on the piecewise uniform fitted mesh $\overline{\Omega}^N$. The discrete problem is

$$\vec{L}^N\vec{U}(x_j) = ED^-\vec{U}(x_j)+A(x_j)\vec{U}(x_j)+B(x_j)\vec{U}(x_j-1) = \vec{f}(x_j), \quad j = 1(1)N \tag{14}$$

$$\vec{U}(0) = \vec{u}(0), \tag{15}$$

where

$$D^-\vec{U}(x_j) = \frac{\vec{U}(x_j) - \vec{U}(x_{j-1})}{x_j - x_{j-1}}, \quad j = 1(1)N.$$

Lemma 6. *If $\vec{\Psi}(x_j)$ is any vector mesh function such that $\vec{\Psi}(0) \geq \vec{0}$ and $\vec{L}^N\vec{\Psi}(x_j) \geq \vec{0}$ for $1 \leq j \leq N$ then $\vec{\Psi}(x_j) \geq \vec{0}$ for $0 \leq j \leq N$.*

Proof. Let i^*, j^* be such that $\Psi_{i^*}(x_{j^*}) = \min\limits_{i,j} \Psi_i(x_j), \quad 1 \leq j \leq N.$

Hence $j^* \neq 0$. Then,

$$\begin{aligned}\left(\vec{L}^N\vec{\Psi}\right)_{i^*}(x_{j^*}) &= -\varepsilon_{i^*}D^-\Psi_{i^*}(x_{j^*}) + \sum_{k=1}^{2} a_{i^*k}(x_{j^*})\Psi_k(x_{j^*}) + b_{i^*}(x_{j^*})\Psi(x_{j^*}-1) \\ &< -\varepsilon_{i^*}D^-\Psi_{i^*}(x_{j^*}) + \sum_{k=1}^{2} a_{i^*k}(x_{j^*})\Psi_k(x_{j^*}) + b_{i^*}(x_{j^*})\Psi(x_{j^*}) \ < \ 0,\end{aligned}$$

which contradicts the hypothesis and proves the lemma.

An immediate consequence of the discrete maximum principle is the following discrete stability result. □

Lemma 7. *Let $\vec{\Psi}$ be any vector mesh function in the domain of $\vec{L}^N$. Then*

$$\| \vec{\Psi}(x_j) \| \leq \max \{\| \vec{\Psi}(0) \|, \tfrac{1}{\alpha} \| \vec{L}^N\vec{\Psi}(x_j) \|\}, \quad j = 0(1)N.$$

Proof. Let $M = \max \{\| \vec{\Psi}(0) \|, \frac{1}{\alpha} \| \vec{L}^N\vec{\Psi}(x_j) \|\}$. Define the barrier functions

$$\vec{\Theta}^{\pm}(x_j) = M(1, 1)^T \pm \vec{\Psi}(x_j).$$

Then,

$$\vec{\Theta}^{\pm}(0) = M(1,1)^T \pm \vec{\Psi}(0) \geq \vec{0}.$$

Also for $j = 1(1)N$,

$$\vec{L}^N \vec{\Psi}^{\pm}(x_j) = M(A(x_j) + B(x_j))(1,1)^T \pm \vec{L}^N \vec{\Psi}(x_j) > M\alpha(1,1)^T \pm \vec{f} \geq \vec{0}.$$

Hence the result follows from the discrete maximum principle. □

5 Error Estimate

Analogous to the continuous case, the discrete solution $\vec{U}$ can be decomposed into $\vec{V}$ and $\vec{W}$ which are defined to be the solutions of the following discrete problems:

$$(\vec{L}_1^N \vec{V})(x_j) = ED^-\vec{V}(x_j) + A(x_j)\vec{V}(x_j) = \vec{f}(x_j) - B(x_j)\vec{\phi}(x_j - 1) \text{ on } \Omega^{-N}$$
$$(\vec{L}_2^N \vec{V})(x_j) = ED^-\vec{V}(x_j) + A(x_j)\vec{V}(x_j) + B(x_j)\vec{V}(x_j - 1) = \vec{f}(x_j) \text{ on } \Omega^{+N}$$

and

$$(\vec{L}_1^N \vec{W})(x_j) = \vec{0}, \ x_j \in \Omega^{-N},$$

$$(\vec{L}_2^N \vec{W})(x_j) = \vec{0}, \ x_j \in \Omega^{+N},$$

$$\vec{W}(0) = \vec{w}(0).$$

The error at each point $x_j \in \overline{\Omega}^N$ is denoted by $\vec{e}(x_j) = \vec{U}(x_j) - \vec{u}(x_j)$. Then the local truncation error $\vec{L}^N \vec{e}(x_j)$ has the decomposition

$$\vec{L}^N \vec{e}(x_j) = \vec{L}^N(\vec{V} - \vec{v})(x_j) + \vec{L}^N(\vec{W} - \vec{w})(x_j).$$

The error in the smooth and singular components is bounded in the following theorems.

Theorem 1. *Let conditions (3) and (4) hold. If $\vec{v}$ denotes the smooth component of the solution of (1), (2) and $\vec{V}$ the smooth component of the solution of the problem (14), (15), then*

$$|(\vec{L}^N(\vec{V} - \vec{v}))_i(x_j)| \leq C\ N^{-1} \ln N. \tag{16}$$

Proof. Proceeding as in [7] in the domains Ω^{-N} and Ω^{+N} separately, the above estimate is derived. □

Theorem 2. *Let conditions (3) and (4) hold. If $\vec{w}$ denotes the singular component of the solution of (1), (2) and $\vec{W}$ the singular component of the solution of the problem (14), (15), then*

$$|(\vec{L}^N(\vec{W}-\vec{w}))_i(x_j)| \leq C\,N^{-1}\ln N. \tag{17}$$

Proof. Proceeding as in [7] in the domains Ω^{-N} and Ω^{+N} separately, the above estimate is derived. □

The main theoretical result of this paper is presented in

Theorem 3. *Let $\vec{u}$ be the solution of the continuous problem (1), (2) and $\vec{U}$ be the solution of the problem (14), (15). Then,*

$$\parallel \vec{U}(x_j)-\vec{u}(x_j) \parallel \leq C\,N^{-1}\ln N.$$

Proof. From Lemma 7, it is clear that, in order to prove the above theorem, it suffices to prove that $\parallel (\vec{L}^N(\vec{U}-\vec{u})) \parallel \leq C\,N^{-1}\ln N$. But, $\parallel (\vec{L}^N(\vec{U}-\vec{u})) \parallel \leq \parallel (\vec{L}^N(\vec{V}-\vec{v})) \parallel + \parallel (\vec{L}^N(\vec{W}-\vec{w})) \parallel$. Hence, using Theorems 1 and 2, the above result is derived. □

6 Numerical Results

The numerical method proposed in this paper is illustrated through an example presented in this section.

Example. Consider the initial value problem

$$E\vec{u}'(x)+A(x)\vec{u}(x)+B(x)u(x-1)=\vec{f}(x),\text{for}x\in(0,2],\ \vec{u}(x)=\vec{\phi}(x),\text{for}x\in[-1,0],$$

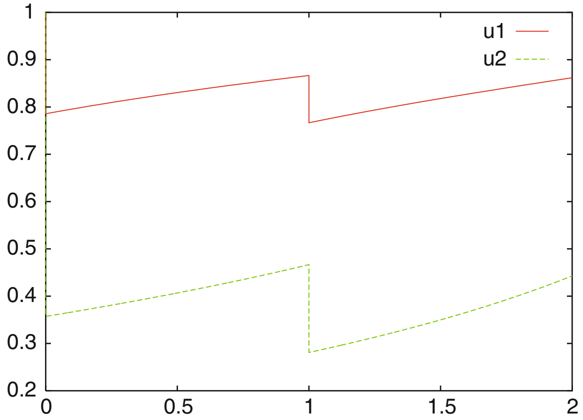
where $E=\text{diag}(\varepsilon_1,\ \varepsilon_2)$, $A=\begin{pmatrix}3+x & -1\\ -1 & 5-x\end{pmatrix}$, $B=\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}$ $\vec{f}=(1+x,\,0)^T$, $\vec{\phi}(x)=\vec{1}$ for $x\in[-1,0]$. The maximum pointwise errors and the rate of convergence for this IVP are presented in Table 1 and a graph of the numerical solution for $N=2048, \varepsilon_1=2^{-19}, \varepsilon_2=2^{-17}$ is given in Fig. 1.

Table 1 Values of D_ε^N, D^N, p^N, p^* and $C_{p^*}^N$ for $\varepsilon_1 = \frac{\eta}{16}$, $\varepsilon_2 = \frac{\eta}{4}$ and $\alpha = 0.9$

	Number of mesh points N							
η	128	256	512	$\cdots$	4096	8192	16384	32768
2^0	0.421E−02	0.215E−02	0.109E−02	$\cdots$	0.137E−03	0.686E−04	0.343E−04	0.172E−04
2^{-3}	0.148E−01	0.976E−02	0.589E−02	$\cdots$	0.104E−02	0.521E−03	0.261E−03	0.131E−03
2^{-6}	0.147E−01	0.970E−02	0.586E−02	$\cdots$	0.109E−02	0.598E−03	0.325E−03	0.175E−03
2^{-9}	0.147E−01	0.970E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.325E−03	0.175E−03
2^{-12}	0.147E−01	0.970E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
2^{-15}	0.147E−01	0.969E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
2^{-18}	0.147E−01	0.969E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
2^{-21}	0.147E−01	0.969E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
2^{-24}	0.147E−01	0.969E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
2^{-27}	0.147E−01	0.969E−02	0.585E−02	$\cdots$	0.109E−02	0.597E−03	0.324E−03	0.175E−03
D^N	0.148E−01	0.976E−02	0.589E−02	$\cdots$	0.109E−02	0.598E−03	0.325E−03	0.175E−03
p^N	0.599E+00	0.728E+00	0.665E+00	$\cdots$	0.865E+00	0.880E+00	0.890E+00	
C_p^N	0.796E+00	0.796E+00	0.728E+00	$\cdots$	0.467E+00	0.389E+00	0.320E+00	0.262E+00

Computed order of $\vec{\varepsilon}$ -uniform convergence, $p^* = 0.5991$
Computed $\vec{\varepsilon}$ -uniform error constant, $C_{p^*}^N = 0.7960$

Fig. 1 Numerical solution



Acknowledgement The first author wishes to acknowledge the financial support extended under the "INSPIRE" fellowship by the Department of Science and Technology, Government of India.

References

1. Ducrot, A., Magal, P., Seydi, O.: A singularly perturbed delay differential equation modeling nosocomial infections. Differ. Integr. Equat. (to appear)
2. Lange, C.G., Miura, R.M.: Singular perturbation analysis of boundary-value problems for differential-differnce equations. SIAM J. Appl. Math. **42**(3), 502–530 (1982)
3. Zhongdi, C.: A hybrid finite difference scheme for a class of singularly perturbed delay differential equations. Neural, Parallel Sci. Comput. **16**, 303–308 (2008)
4. Hongjiong, T.: Numerical methods for singularly perturbed delay differential equations. In: Proceedings of an International conference on Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2004, ONERA, Toulouse (2004)
5. Subburayan, V., Ramanujam, N.: An initial value technique for singularly perturbed convection diffusion problems with a negative shift. J. Optim. Theory Appl. **158**, 234–250 (2013)
6. Nicaise, S., Xenophontos, C.: Robust approximation of singularly perturbed delay differential equations by the hp finite element method. Comput. Methods Appl. Math. **13**(1), 21–37 (2013)
7. Valarmathi, S., Miller, J.J.H.: A parameter uniform finite difference method for a singularly perturbed linear dynamical systems. Int. J. Numer. Anal. Model. **7**(3), 535–548 (2010)
8. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems. World Scientific Publishing Co., Singapore (1996)
9. Linss, T., Madden, N.: Accurate solution of a system of coupled singularly perturbed reaction-diffusion equations. Computing **73**, 121–133 (2004)
10. Farrell, P.A., Hegarty, A., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust computational techniques for boundary layers. In: Knops, R.J., Mortonm, K.W. (eds.) Applied Mathematics and Mathematical Computation. Chapman & Hall/CRC Press, Boca Raton (2000)

A Technique to Construct Grid Methods of Higher Accuracy Order for a Singularly Perturbed Parabolic Reaction-Diffusion Equation

L. Shishkina and G. Shishkin

Abstract We consider a technique to construct ε-uniformly convergent in the maximum norm grid approximations of higher accuracy order on uniform grids for a singularly perturbed parabolic reaction-diffusion equation with a perturbation parameter ε ($\varepsilon \in (0, 1]$) multiplying the highest-order derivative, the solution of which has a parabolic boundary layer in a neighborhood of the lateral boundary.

Keywords Parabolic reaction-diffusion equation • Parabolic boundary layer • Uniform grids • Decomposition of grid solution • Asymptotic construction technique • Richardson extrapolation • Higher-order finite difference scheme • ε-uniform convergence • Maximum norm

1 Introduction

The use of *efficient* numerical methods developed for regular problems (see, e.g., [1]) and based on *standard* difference *schemes on uniform grids* does not provide ε-uniform convergence in the maximum norm for one-dimensional singularly perturbed parabolic reaction-diffusion equation with a perturbation parameter ε ($\varepsilon \in (0, 1]$) multiplying the highest-order derivative in the equation (the solution of such a problem has a parabolic boundary layer in a neighborhood of the lateral boundary; see, e.g., [2]).

In [3], a new approach is developed to construct ε-uniformly convergent difference schemes *on uniform grids* of high accuracy order for a one-dimensional singularly perturbed parabolic reaction-diffusion equation. Using an *asymptotic construction technique*, a basic scheme of *the solution decomposition method*

L. Shishkina (✉) • G. Shishkin
Institute of Mathematics and Mechanics, Russian Academy of Sciences, Ekaterinburg, Russia
e-mail: lida@convex.ru

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_13,

is constructed in which the regular and singular components of the discrete solution are solutions of *discrete subproblems* considered on *uniform meshes*. The basic difference scheme converges ε-uniformly in the maximum norm at the rate $\mathcal{O}\left(N^{-2}\ln^2 N + N_0^{-1}\right)$, where $N+1$ and N_0+1 are the number of nodes in the spatial and time meshes, respectively. The use of the Richardson extrapolation technique to the basic scheme leads to a scheme of higher accuracy order, i.e., the Richardson scheme of the solution decomposition method, which converges ε-uniformly in the maximum norm at the rate $\mathcal{O}\left(N^{-4}\ln^4 N + N_0^{-2}\right)$.

In this paper, we present a modified technique from [3] which simplifies the constructions under numerical solving of similar problems. This technique allows you to construct ε-uniformly convergent schemes of the fourth accuracy order in x up to a logarithmic factor and the second order in t, using the Richardson extrapolation on two embedded grids, and to construct ε-uniformly convergent schemes of the sixth accuracy order $\mathcal{O}\left(N^{-6}\ln^6 N + N_0^{-3}\right)$ in x up to a logarithmic factor and the third order in t, using the Richardson extrapolation on three embedded grids.

2 Problem Formulation and Aim of Research

On the set $\overline{G}$

$$\overline{G} = G \cup S, \quad G = D \times (0,T], \quad D = (0,d), \tag{1}$$

we consider a boundary value problem for the singularly perturbed parabolic reaction-diffusion equation[1]

$$L_{(2)}\, u(x,t) \equiv \left\{\varepsilon^2 a(x,t)\,\frac{\partial^2}{\partial x^2} - c(x,t) - p(x,t)\,\frac{\partial}{\partial t}\right\} u(x,t) = f(x,t), \ (x,t) \in G,$$

$$u(x,t) = \varphi(x,t), \quad (x,t) \in S. \tag{2}$$

The functions $a(x,t)$, $c(x,t)$, $p(x,t)$, $f(x,t)$, and $\varphi(x,t)$ are assumed to be sufficiently smooth on $\overline{G}$ and S, respectively; moreover[2] $a(x,t), c(x,t), p(x,t) > m$, $|f(x,t)| \le M$, $(x,t) \in \overline{G}$; $|\varphi(x,t)| \le M$, $(x,t) \in S$; the parameter ε takes arbitrary values in $(0,1]$. Here $S = S_0 \cup S^L$, S_0 and S^L are the lower and lateral sides of the boundary S, $S^L = S_1^L \cup S_2^L$, S_1^L and S_2^L are the left and right parts of the lateral boundary, and $S_0 = \overline{S}_0$. We assume that the data of problem (2), (1) on

[1] The notation $L_{(j.k)}$ $(M_{(j.k)},\ G_{h(j.k)})$ means that these operators (constants, grids) are introduced in formula $(j.k)$.

[2] By M (or m), we denote sufficiently large (small) positive constants independent of the parameter ε and of the discretization parameters.

the set of corner points $S^* = S_0 \cap \overline{S}^L$ satisfy the compatibility conditions ensuring the required smoothness of the solution on $\overline{G}$. For small values of ε, a parabolic boundary layer appears in a neighborhood of the set S^L.

Our aim is for initial boundary value problem (2), (1) on the basis of the solution decomposition method using only uniform grids and the Richardson extrapolation technique, to construct a difference scheme that converges ε-uniformly in the maximum norm with a higher accuracy order (two/three with respect to t and four/six with respect to x up to a logarithmic factor).

3 Difference Scheme of the Solution Decomposition Method

In this section we consider a decomposition of the solution to problem (2), (1) using an asymptotic construction technique. On the basis of this solution decomposition of the differential problem, we construct a difference scheme (scheme of the decomposition method to the grid solution), in which the grid regular and singular components of the discrete solution are computed on *uniform meshes*.

3.1 Solution Decomposition of the Differential Problem

We write the solution of problem (2), (1) as the sum of its regular $U(x,t)$ and singular $V(x,t)$ components:

$$u(x,t) = U(x,t) + V(x,t), \quad (x,t) \in \overline{G}. \tag{3a}$$

Because we will construct a scheme of improved accuracy order, we use the expansion of the regular component of the three members:

$$U(x,t) = U_0(x,t) + \varepsilon^2\, U_1(x,t) + v_U(x,t), \quad (x,t) \in \overline{G}. \tag{3b}$$

Here $U_0(x,t)$ is the main term in the expansion, and $v_U^0(x,t)$ is the remainder term. In (3b), the functions $U_0(x,t)$, $U_1(x,t)$ and $v_U^0(x,t)$ are solutions of the following problems:

$$L_{(4)}U_0(x,t) = f(x,t),\ (x,t) \in \overline{G} \setminus S_0,\ U_0(x,t) = \varphi(x,t),\ (x,t) \in S_0; \tag{4a}$$

$$L_{(4)}U_1(x,t) = -\varepsilon^2 a(x,t)\,\frac{\partial^2}{\partial x^2}\,U_0(x,t),\ (x,t) \in \overline{G} \setminus S_0,\ U_1(x,t) = 0,\ (x,t) \in S_0; \tag{4b}$$

$$L_{(2)}v_U(x,t) = -\varepsilon^4 a(x,t)\,\frac{\partial^2}{\partial x^2}\,U_1(x,t),\ \ (x,t) \in G, \quad v_U(x,t) = 0,\ \ (x,t) \in S. \tag{4c}$$

Here $L_{(4)}$ is the operator $L_{(2)}$ for $\varepsilon = 0$, i.e., $L_{(4)} \equiv -c(x,t) - p(x,t)\,\dfrac{\partial}{\partial t}$, $(x,t) \in \overline{G} \setminus S_0$.

The function $V(x,t)$, $(x,t) \in \overline{G}$, is the solution of the problem

$$L_{(2)}\, V(x,t) = 0, \;\; (x,t) \in G, \quad V(x,t) = \varphi_V(x,t), \;\; (x,t) \in S, \tag{5a}$$

where $\varphi_V(x,t) = \varphi(x,t) - U(x,t),\; (x,t) \in S,\; U(x,t) = U_{(3)}(x,t),\; (x,t) \in \overline{G}$. We write $V(x,t)$ as the sum of the functions

$$V(x,t) = V_1(x,t) + V_2(x,t), \quad (x,t) \in \overline{G}, \tag{3c}$$

where the functions $V_1(x,t)$ and $V_2(x,t)$ are solutions of the problems

$$L_{(2)}\, V_j(x,t) = 0, \quad (x,t) \in G, \tag{5b}$$

$$V_j(x,t) = \varphi_V(x,t), \;\; (x,t) \in S_j^L, \quad V_j(x,t) = 0, \;\; (x,t) \in S \setminus S_j^L, \quad j = 1, 2.$$

Thus, for the solution of problem (2), (1), we obtained representation (3) whose components are solutions of problems (4) and (5).

3.2 Construction of the Basic Scheme of the Solution Decomposition Method

Now, we construct a difference scheme for the boundary value problem (2), (1) by approximating problems (4), (5b). We consider two cases depending on the value of ε.

3.2.1 Difference Scheme for Not Too Small Values of ε

We construct a difference scheme for not too small values of the parameter ε, namely, provided

$$\varepsilon \geq \varepsilon_0(N), \quad \varepsilon_0(N) = m\,\ell^{-1}\, d\; \ln^{-1} N, \tag{6}$$

where m is an arbitrary number in $(0, m_0)$, $m_0 = \min_{\overline{G}}^{1/2}\,[a^{-1}(x,t)\,c(x,t)]$, and $\ell = 2$. In this case, problem (2), (1) is approximated by the standard difference scheme on a uniform grid

$$\Lambda z(x,t) \equiv \{\varepsilon^2\, a(x,t)\,\delta_{x\overline{x}} - c(x,t) - p(x,t)\,\delta_{\overline{t}}\} z(x,t) = f(x,t), \; (x,t) \in G_h,$$

$$z(x,t) = \varphi(x,t), \; (x,t) \in S_h. \tag{7}$$

Here

$$\overline{G}_h = \overline{\omega} \times \overline{\omega}_0 \tag{8}$$

is the uniform grid, $N + 1$ and $N_0 + 1$ are the numbers of nodes in the meshes $\overline{\omega}$ and $\overline{\omega}_0$, respectively, $\overline{G}_h = G_h \cup S_h$, and $\delta_{x\overline{x}}\, z(x,t)$ is the central second-order difference derivative. Using the solution of difference scheme (7), (8), we construct the interpolant

$$\bar{z}_u(x,t), \quad (x,t) \in \overline{G}, \ \text{under conditon} \ (6). \tag{9a}$$

We call the interpolant (9a) the solution of the difference scheme [(7), (8); (6)] which approximates the differential problem (2), (1) under the condition (6).

3.2.2 Difference Scheme for Sufficiently Small Values of *ε*

We construct a difference scheme for sufficiently small values of the parameter ε, namely, for

$$\varepsilon < \varepsilon_0(N)_{(6)}. \tag{10}$$

We approximate the components in the representation (3b) and the function $U(x,t)$ on the uniform grid (8) and the singular components in (5b) on uniform grids which are constructed on subdomains of $\overline{G}{}_j^{\sigma}$ of $\overline{G}$, adjacent to the boundaries S_j^L, $j = 1, 2$:

$$\begin{gathered}\overline{G}{}_j^{\sigma} = G_j^{\sigma} \bigcup S_j^{\sigma}, \quad G_j^{\sigma} = D_j^{\sigma} \times (0,T], \ j = 1,2, \\ D_1^{\sigma} = (0,\sigma), \ D_2^{\sigma} = (d-\sigma, d), \\ \sigma = \sigma(\varepsilon,\, N,\, l) = \min\left[d,\, m^{-1}\, l\, \varepsilon\, \ln N\right].\end{gathered} \tag{11}$$

Differential problems (4), (1) are approximated by the following problems on the grid (8):

$$\begin{aligned}\Lambda_{(12)} z_{U_0}(x,t) &= f(x,t), \quad (x,t) \in \overline{G}_h \setminus S_0, \\ z_{U_0}(x,t) &= \varphi(x,t), \quad (x,t) \in S_h \cap S_0;\end{aligned} \tag{12}$$

$$\begin{aligned}\Lambda_{(12)} z_{U_1}(x,t) &= -\varepsilon^2\, a(x,t)\, \delta_{x\,\overline{x}}\, z_{U_0}(x,t), \quad (x,t) \in \overline{G}_h \setminus S_0, \\ z_{U_1}(x,t) &= \varphi(x,t), \quad (x,t) \in S_h \cap S_0;\end{aligned} \tag{13}$$

$$\begin{aligned}\Lambda_{(7)} z_{v_U}(x,t) &= -\varepsilon^4\, a(x,t)\, \delta_{x\,\overline{x}}\, z_{U_1}(x,t), \quad (x,t) \in G_h, \\ z_{v_U}(x,t) &= 0, \quad (x,t) \in S_h.\end{aligned} \tag{14}$$

Here $G_h = G \cap \overline{G}_h$, $S_h = S \cap \overline{G}_h$, $\Lambda_{(12)} \equiv -c(x,t) - p(x,t)\,\delta_{\overline{t}}$, $(x,t) \in \overline{G}_h \setminus S_0$. We set

$$z_U(x,t) = z_{U_0}(x,t) + \varepsilon^2 z_{U_1}(x,t) + z_{v_U}(x,t), \quad (x,t) \in \overline{G}_h. \tag{15}$$

By $\bar{z}_U(x,t), (x,t) \in \overline{G}$, we denote the bilinear interpolant which is constructed using the values of $z_U(x,t)$ at the nodes of the grid $\overline{G}_h$ on the elementary partitions of the set $\overline{G}$, generated by the grid $\overline{G}_h$. The function $z_U(x,t)$, $(x,t) \in \overline{G}_h$ and also its interpolant $\bar{z}_U(x,t), (x,t) \in \overline{G}$ are called the grid and continual solutions, respectively, of the difference scheme [(12)–(14), (8); (10)], which approximates differential problems (4), (1) under condition (10).

Now, we construct an approximation of problem (5), (1).

On the set $\overline{G}^{\sigma}_{j\,(11)}$ we introduce the uniform grid

$$\overline{G}^{\sigma}_{j\,h} = \overline{G}^{\sigma\,u}_{j\,h} \equiv \overline{\omega}^{\sigma}_{j} \times \overline{\omega}_0, \quad j = 1, 2, \tag{16}$$

where $\overline{\omega}_0 = \overline{\omega}_{0\,(8)}$, $\overline{\omega}^{\sigma}_{j}$ is the mesh defined on $\overline{D}^{\sigma}_{j\,(11)}$ with the step size $h^{\sigma} = \sigma N^{-1}$, $N + 1$ is the number of nodes in the mesh $\overline{\omega}^{\sigma}_{j}$, and $\overline{G}^{\sigma}_{j\,h} = G^{\sigma}_{j\,h} \cup S^{\sigma}_{j\,h}$. On the grid $\overline{G}^{\sigma}_{j\,h}$, we solve the problem

$$\Lambda_{(7)} z_{V_j}(x,t) = 0, \quad (x,t) \in G^{\sigma}_{j\,h}, \tag{17}$$

$$z_{V_j}(x,t) = \left\{ \begin{array}{ll} \varphi(x,t) - z_U(x,t), & (x,t) \in S^{\sigma}_{j\,h} \cap \overline{S}^{L} \\ 0, & (x,t) \in S^{\sigma}_{j\,h} \setminus \overline{S}^{L} \end{array} \right\}, \quad (x,t) \in S^{\sigma}_{j\,h}, \quad j = 1, 2.$$

Using the values of the functions $z_{V_j}(x,t)$, $(x,t) \in \overline{G}^{\sigma}_{j\,h}$, we construct the interpolant $\bar{z}_{V_j}(x,t)$, $(x,t) \in \overline{G}^{\sigma}_{j}$. We assume that, outside of the set $\overline{G}^{\sigma}_{j}$, the functions $z_{V_j}(x,t)$ and $\bar{z}_{V_j}(x,t)$ vanish. We set

$$\bar{z}_V(x,t) = \bar{z}_{V_1}(x,t) + \bar{z}_{V_2}(x,t), \quad (x,t) \in \overline{G}.$$

The function $\bar{z}_V(x,t)$, $(x,t) \in \overline{G}^{\sigma}$, is called the solution of difference scheme [(17), (16); (10)], which approximates differential problem (5a), (1) under condition (10).

The function

$$\bar{z}_u(x,t) = \bar{z}_U(x,t) + \bar{z}_V(x,t), \quad (x,t) \in \overline{G}, \text{ under condition (10)} \tag{9b}$$

is called the solution of difference scheme [{(12)–(14), (8); (17), (16)}; (10)], which approximates differential problem (2), (1) under condition (10).

3.2.3 $\boldsymbol{\varepsilon}$-uniform Estimate for the Solution of the Basic Scheme

Thus, we have constructed the function $\bar{z}_{u\,(9\mathrm{a,b})}(x,t)$, $(x,t) \in \overline{G}$ approximating the solution of problem (2), (1). This function and the grid functions $z_{U_0}(x,t)$, $z_{U_1}(x,t)$, $z_{v_U}(x,t)$, $(x,t) \in \overline{G}_h$, and $z_{V_j}(x,t)$, $(x,t) \in \overline{G}^{\sigma}_{j\,h}$, $j = 1, 2$, are called the continual and grid solutions, respectively, of difference scheme [{(7), (8)}; {(12)–(14), (8); (17), (16)}], or, in short, *the basic scheme of the solution decomposition method.*

In [3] for the solution to the basic scheme of the solution decomposition method we have obtained the following ε-uniform estimate:

$$|u(x,t) - \bar{z}_{u\,(9\mathrm{a,b})}(x,t)| \le M\,[N^{-2}\,\ln^2 N + N_0^{-1}], \quad (x,t) \in \overline{G}. \tag{18}$$

4 Richardson Extrapolation on the Basis of Classical Scheme

We describe the Richardson extrapolation method, which is used for improving the accuracy of the solution to difference scheme (7) on uniform grid (8).

On the set $\overline{G}$, we introduce the grids

$$\overline{G}^{i}_{h} = \overline{\omega}^{i} \times \overline{\omega}^{i}_{0}, \quad i = 1, 2, 3, \tag{19a}$$

in which $\overline{\omega}^{i}$ and $\overline{\omega}^{i}_{0}$ are uniform meshes with respect to x and t, respectively. Here $\overline{G}^{1}_{h}$ is $\overline{G}_{h(8)}$, in which $h^1_x = dN^{-1}$ is the step size in $\overline{\omega}^1$ with the number of nodes $N+1$, and $h^1_t = TN_0^{-1}$ is the step size in $\overline{\omega}^1_0$ with the number of nodes N_0+1; $\overline{G}^{2}_{h}$ and $\overline{G}^{3}_{h}$ are "coarsened" grids. The step size h^2_x in $\overline{\omega}^2$ (on the interval $\overline{D}$) is k times larger than the step size h^1_x in $\overline{\omega}^1$, i.e., $h^2_x = k\,d\,N^{-1}$ and $k^{-1}N+1$ is the number of nodes in $\overline{\omega}^2$. The step size h^2_t in $\overline{\omega}^2_0$ (on the interval $[0,T]$) is k^2 times larger than the step size h^1_t in $\overline{\omega}^1_0$, i.e., $h^2_t = k^2\,T\,N_0^{-1}$ and $k^{-2}N_0+1$ is the number of nodes in $\overline{\omega}^2_0$. The step size h^3_x in $\overline{\omega}^3$ (on the interval $\overline{D}$) is k^2 times larger than the step size h^2_x in $\overline{\omega}^2$, i.e., $h_x^{=}k^2\,d\,N^{-1}$ è $k^{-2}N+1$ is the number of nodes in $\overline{\omega}^3$. The step size h^3_t in $\overline{\omega}^3_0$ (on the interval $[0,T]$) is k^2 times larger than the step size h^2_t in $\overline{\omega}^2_0$, i.e., $h^3_t = k^4\,T\,N_0^{-1}$ è $k^{-4}N_0+1$ is the number of nodes in $\overline{\omega}^3_0$.

We consider the case with two embedded grids $\overline{G}^{1}_{h}$ and $\overline{G}^{2}_{h}$, and let $\overline{G}^{0}_{h}$ be their intersection:

$$\overline{G}^{0}_{h} = \overline{G}^{1}_{h} \cap \overline{G}^{2}_{h} \tag{19b}$$

$\overline{G}^{0}_{h} = \overline{G}^{1}_{h}$ if k is an integer ($k \ge 2$), $\overline{G}^{0}_{h} \ne \overline{G}^{1}_{h}$ if k is a noninteger; $\overline{G}^{0}_{h} = \overline{\omega}^0 \times \overline{\omega}^0_0$.

Let $z^i(x,t)$, $(x,t) \in \overline{G}^{i}_{h}$, for $i = 1,\ 2$, be solutions of the difference schemes

$$\Lambda_{(7)} z^i(x,t) = f(x,t), \ \ (x,t) \in G_h^i, \quad z^i(x,t) = \varphi(x,t), \ \ (x,t) \in S_h^i, \quad i = 1,2. \tag{20a}$$

We set

$$z^0(x,t) = \gamma_1 \, z^1(x,t) + \gamma_2 \, z^2(x,t), \quad (x,t) \in \overline{G}_h^0, \tag{20b}$$

where $\gamma_i = \gamma_i(k), i = 1,2, \gamma_1 = -(k^2-1)^{-1}, \gamma_2 = 1 - \gamma_1 = k^2\,(k^2-1)^{-1}$.

Difference scheme (20), (19) constructed on the basis of scheme (7), (8) is called the Richardson scheme on two embedded grids. The function $z^0_{(20)}(x,t)$, $(x,t) \in \overline{G}_h^0$, is called the solution to Richardson scheme (20), (19), while the functions $z^1_{(20)}(x,t)$, $(x,t) \in \overline{G}_h^1$, and $z^2_{(20)}(x,t)$, $(x,t) \in \overline{G}_h^2$, are called the components generating the solution of scheme (20), (19).

In [3], justification of convergence of the solution $z^0(x,t)$ of Richardson scheme (20), (19) to the solution $u(x,t)$ of boundary value problem (2), (1) is performed, and we have obtained the following estimate:

$$|u(x,t) - z^0(x,t)| \le M \, \left[\varepsilon^{-4} \, N^{-4} + N_0^{-2}\right], \quad (x,t) \in \overline{G}_h^0. \tag{21}$$

Thus, the Richardson scheme (20), (19) converges with the fourth accuracy order in x but for fixed values of ε and under the sufficiently restrictive condition $N^{-1} = o(\varepsilon)$, $N_0^{-1} = o(1)$.

5 Richardson Extrapolation for Solution Decomposition Scheme

Here we will apply the Richardson extrapolation technique to improve accuracy order of discrete solutions obtained on the basis of the solution decomposition method.

To construct a scheme of higher accuracy, we approximate problem (2), (1) by the standard difference scheme (7), (8) under the condition

$$\varepsilon \ge \varepsilon_0(N), \quad \varepsilon_0(N) = m \, \ell^{-1} \, d \, \ln^{-1} N, \tag{22}$$

where $m = m_{(6)}$, $\ell = 4$, and under the condition

$$\varepsilon < \varepsilon_0(N), \quad \varepsilon_0(N) = \varepsilon_{0\,(22)}(N) \tag{23}$$

we use grid constructions similar to [(12)–(14), (8); (17), (16)]. Further, to improve accuracy of the scheme, we apply the Richardson extrapolation technique.

5.1 Solution of the Richardson Scheme Provided (22)

Let the condition (22) be fulfilled. Using the solution $z^0_{(20)}(x,t)$, $(x,t) \in \overline{G}^0_h$, of the Richardson scheme (20), (19) on two embedded grids, we construct the interpolant

$$\hat{z}_u(x,t), \quad (x,t) \in \overline{G} \text{ under condition } (22). \tag{24a}$$

We call this interpolant the solution of difference scheme [(20), (19), (22)] which approximates the differential problem (2), (1) under the condition (22).

5.2 Solution of the Richardson Scheme provided (23) for the Regular Component $U(x,t)$

Let the condition (23) be fulfilled. Using the Richardson extrapolation on two embedded grids, we construct a grid approximation of the component $U(x,t)$.

On the set $\overline{G}$, we introduce the embedded grids

$$\overline{G}^i_h = \overline{G}^i_{h\,(19)} = \overline{\omega}^i \times \overline{\omega}^i_0, \quad i = 1,2; \quad \overline{G}^0_h = \overline{G}^0_{h\,(19)}. \tag{25}$$

To approximate problems (4a) and (4b), we need an extension of problem (4a) to the set $\overline{G}^e$

$$\overline{G}^e = \overline{D}^e \times [0,T], \quad \overline{D}^e = [-h^2, d + h^2], \tag{26a}$$

where h^2 is the step size of the "coarsened" mesh $\overline{\omega}^1_{(25)}$. We extend the data specifying problem (2), (1) to the set $\overline{G}^e$ with preserving their properties; the extended functions $a(x,t), \ldots, f(x,t)$, $(x,t) \in \overline{G}$ and $\varphi(x,t)$, $(x,t) \in S_0$, are denoted by $a^e(x,t), \ldots, f^e(x,t)$, $(x,t) \in \overline{G}^e$ è $\varphi^e(x,t)$, $(x,t) \in S^e_0$. In decomposition (3b) and in (4b), we have

$$U_{0\,(3)}(x,t) = U^e_0(x,t), \quad (x,t) \in \overline{G}. \tag{26b}$$

Here $U^e_0(x,t)$, $(x,t) \in \overline{G}^e$ is the solution of the "extended" problem

$$L^e_{(4)}\, U^e_0(x,t) = f^e(x,t), \ (x,t) \in \overline{G}^e \setminus S^e_0, \quad U^e_0(x,t) = \varphi^e(x,t), \ (x,t) \in S^e_0. \tag{26c}$$

On the set $\overline{G}^e$, we construct the embedded grids

$$\overline{G}^{e\,i}_h = \overline{G}^{e\,i}_{h\,(27)} = \overline{\omega}^{e\,i} \times \overline{\omega}^i_{0\,(25)}, \quad i = 1,2; \quad \overline{G}^{e\,0}_h = \overline{G}^{e\,1}_h \cap \overline{G}^{e\,2}_h, \tag{27}$$

where $\overline{\omega}^{e\,i}$ are extended uniform meshes and $\overline{\omega}^{e\,i} \cap \overline{D} = \overline{\omega}^i_{(25)}$, $i = 1, 2$.

Problems (26c), (26a) and (4b), (4c), (1) are approximated by the difference schemes on the grids $\overline{G}_h^{e\,i}$ and $\overline{G}_h^{i}$, respectively:

$$\begin{aligned} &\Lambda^e_{(28)} z^{e\,i}_{U_0}(x,t) = f^e(x,t), \quad (x,t) \in \overline{G}_h^{e\,i} \setminus S_0^e, \\ &\qquad z^{e\,i}_{U_0}(x,t) = \varphi^e(x,t), \quad (x,t) \in \overline{G}_h^{e\,i} \cap S_0^e; \end{aligned} \tag{28}$$

$$\begin{aligned} &\Lambda_{(12)} z^{i}_{U_1}(x,t) = -\varepsilon^2\, a(x,t)\, \delta_{x\,\overline{x}}\, z^{e\,i}_{U_0}(x,t), \quad (x,t) \in \overline{G}_h^{i} \setminus S_0, \\ &\qquad z^{i}_{U_1}(x,t) = 0, \quad (x,t) \in \overline{G}_h^{i} \cap S_0; \end{aligned} \tag{29}$$

$$\begin{aligned} &\Lambda_{(7)} z^{i}_{v_U}(x,t) = -\varepsilon^4\, a(x,t)\, \delta_{x\,\overline{x}}\, z^{i}_{U_1}(x,t), \quad (x,t) \in G_h^{i}, \\ &\qquad z^{i}_{v_U}(x,t) = 0, \ \ (x,t) \in S_h^{i}, \ i = 1,2. \end{aligned} \tag{30}$$

For the problem on the grid $\overline{G}_h^{e}$, the operator Λ^e is defined by the relation

$$\Lambda^e_{(28)} \equiv -c^e(x,t) - p^e(x,t)\delta_{\overline{t}}, \quad (x,t) \in \overline{G}_h^{e} \setminus S_0^e.$$

We set

$$z^i_U(x,t) = z^{e\,i}_{U_0}(x,t) + \varepsilon^2 z^i_{U_1}(x,t) + z^i_{v_U}(x,t), \quad (x,t) \in \overline{G}_h^{i}, \ i = 1,2. \tag{31}$$

On the set $\overline{G}_h^{0}$, we define the function $z^0_U(x,t)$ in the following way:

$$z^0_U(x,t) = \gamma_1\, z^1_U(x,t) + \gamma_2\, z^2_U(x,t), \quad (x,t) \in \overline{G}_h^{0}, \tag{32}$$

where $\gamma_i = \gamma_{i\,(20)}(k)$.

The function $z^0_U(x,t)$, $(x,t) \in \overline{G}_h^{0}$ is the grid approximation of $U(x,t)$ constructed with the use of the Richardson technique.

Using the function $z^0_U(x,t)$, $(x,t) \in \overline{G}_h^{0}$, we construct its interpolant

$$\hat{z}^0_U(x,t), \quad (x,t) \in \overline{G} \tag{33}$$

as follows. First, for the function $z^0_U(x,t)$ for $t \in \overline{\omega}_0^0$, we construct the interpolant $\tilde{z}^0_U(x,t)$, $x \in \overline{D}$, $t \in \overline{\omega}_0^0$. This is the cubic interpolant on each elementary interval of the partition $\overline{D}$ generated by the mesh $\overline{\omega}^0$. It is constructed from the values of $z^0_U(x,t)$ at three adjacent nodes of the mesh $\overline{\omega}^0$ (see, e.g., in [4] and in [5]). The function $\hat{z}^0_U(x,t)$, $(x,t) \in \overline{G}$ is obtained by applying linear interpolation with respect to t for the function $\tilde{z}^0_U(x,t)$.

The function $\hat{z}^0_U(x,t)$, $(x,t) \in \overline{G}$, is the continual approximation of the function $U(x,t)$ constructed with the use of the Richardson technique.

5.3 Solution of the Richardson Scheme provided (23) for the Singular Component $V(x, t)$

Under the condition (23), using the Richardson technique, we construct a grid approximation of the singular component $V(x, t)$. On the set $\overline{G}$, we define the subsets $\overline{G}^{\sigma}_{j}$

$$\overline{G}^{\sigma}_{j} = \overline{G}^{\sigma}_{j\,(11)} = G^{\sigma}_{j} \bigcup S^{\sigma}_{j}, \quad \sigma = \sigma_{(11)}(\varepsilon,\, N,\, l) \;\; \mathit{for} \;\; l = 4, \quad j = 1, 2. \tag{34}$$

On the sets $\overline{G}^{\sigma}_{j}$, we construct the embedded sets (similar to the grids $\overline{G}^{\,i}_{h\,(25)}$, $\overline{G}^{\,0}_{h\,(25)}$)

$$\overline{G}^{\sigma\,i}_{j\,h} = \overline{G}^{\sigma\,i}_{j\,h\,(35)} = \overline{\omega}^{\sigma\,i}_{j} \times \overline{\omega}^{i}_{0}, \;\; i = 1, 2; \;\; \overline{G}^{\sigma\,0}_{j\,h} = \overline{G}^{\sigma\,0}_{j\,h\,(35)} = \overline{G}^{\sigma\,1}_{j\,h} \cap \overline{G}^{\sigma\,2}_{j\,h}, \;\; j = 1, 2. \tag{35}$$

Problem (5b), (1) is approximated by the difference scheme

$$\Lambda_{(7)} z^{i}_{V_j}(x, t) = 0, \quad (x, t) \in G^{\sigma\,i}_{j\,h}, \tag{36}$$

$$z^{i}_{V_j}(x, t) = \left\{ \begin{array}{ll} \varphi(x, t) - z_U(x, t), & (x, t) \in S^{\sigma\,i}_{j\,h} \cap \overline{S}^{L} \\ 0, & (x, t) \in S^{\sigma\,i}_{j\,h} \setminus \overline{S}^{L} \end{array} \right\}, \;\; (x, t) \in S^{\sigma\,i}_{j\,h}, \;\; i, j = 1, 2.$$

On the set $G^{\sigma\,0}_{j\,h}$ we define the function $z^{0}_{V_j}(x, t)$ as follows:

$$z^{0}_{V_j}(x, t) = \gamma_1\, z^{1}_{V_j}(x, t) + \gamma_2\, z^{2}_{V_j}(x, t), \quad (x, t) \in \overline{G}^{\sigma\,0}_{j\,h}, \quad j = 1, 2, \tag{37}$$

where $\gamma_i = \gamma_{i\,(20)}$, $i = 1, 2$.

The function $z^{0}_{V_j}(x, t)$, $(x, t) \in \overline{G}^{\sigma\,0}_{j\,h}$ is the grid approximation of the function $V_j(x, t)$ constructed with the use of the Richardson technique. We construct its interpolant:

$$\hat{z}^{0}_{V_j}(x, t), \quad (x, t) \in \overline{G}^{\sigma\,0}_{j}, \quad j = 1, 2; \tag{38}$$

outside of $\overline{G}^{\sigma 0}_{j}$, the function $\hat{z}^{0}_{V_j}(x, t)$ is assumed to be zero. Set

$$\hat{z}^{0}_{V}(x, t) = \hat{z}^{0}_{V_1}(x, t) + \hat{z}^{0}_{V_2}(x, t), \quad (x, t) \in \overline{G}.$$

We call the function

$$\hat{z}_u(x,t) = \hat{z}^0_U(x,t) + \hat{z}^0_V(x,t), \quad (x,t) \in \overline{G} \text{ under condition (23)} \tag{24b}$$

the solution of the Richardson difference scheme [{(28)–(30), (25), (27)}; {(36), (35); (23)}], which approximates differential problem (2), (1) under condition (23).

5.4 ε-uniform Estimate for the Solution of the Richardson Solution Decomposition Scheme

Thus, we have constructed the function $\hat{z}_{u\,(24\text{a,b})}(x,t)$, $(x,t) \in \overline{G}$ approximating the solution of problem (2), (1). This function and the grid functions $z^{e\,0}_{U_0}(x,t)$, $z^0_{U_1}(x,t)$, $z^0_{v_U}(x,t)$, $(x,t) \in \overline{G}^0_h$ and $z^0_{V_j}(x,t)$, $(x,t) \in \overline{G}^{\sigma\,0}_{j\,h}$, $j = 1, 2$, are called the continual and grid solutions, respectively, of the Richardson difference scheme [(20), (19); (28)–(30), (25), (27)], or, in short, the solutions of *the Richardson solution decomposition scheme*.

For the solution to the Richardson scheme of the solution decomposition method, in [3] we have obtained the following ε-uniform estimate:

$$|u(x,t) - \hat{z}_u(x,t)| \leq M\,[N^{-4} \ln^4 N + N_0^{-2}], \quad (x,t) \in \overline{G}. \tag{39}$$

5.5 Construction of a Scheme with Higher Accuracy Order

The technique described above allows us to construct a Richardson scheme of type (20) on the three embedded grids $\overline{G}^1_h$, $\overline{G}^2_h$ and $\overline{G}^3_h$ with the solution $z^0(x,t)$ on the set $\overline{G}^0_h$, which is the intersection of the sets, $\overline{G}^0_h = \overline{G}^1_h \cap \overline{G}^2_h \cap \overline{G}^3_h$. Application of this Richardson scheme to the basic scheme of the solution decomposition (similar to constructions in Sect. 5) leads to the scheme of higher accuracy order whose solution $\hat{z}_u(x,t)$ converges ε-uniformly in the maximum norm at the rate $\mathcal{O}\left(N^{-6} \ln^6 N + N_0^{-3}\right)$ on the set $\overline{G}$.

Acknowledgements This research was supported by the Russian Foundation for Basic Research under grant no.13-01-00618.

References

1. Marchuk, G.I., Shaidurov, V.V.: Difference Methods and Their Interpolations. Springer, New York (1983)
2. Shishkin, G.I., Shishkina, L.P.: Difference Methods for Singular Perturbation Problems. Monographs and Surveys in Pure and Applied Mathematics. Chapman and Hall/CRC, Boca Raton (2009)
3. Shishkin, G.I., Shishkina, L.P.: A Richardson scheme of the decomposition method for solving singularly perturbed parabolic reaction-diffusion equation. Comp. Math. Math. Phys. **50**(12), 2003–2022 (2010)
4. Bakhvalov, N.S.: Numerical Methods. Nauka, Moscow (1973) (in Russian)
5. Marchuk, G.I.: Methods of Numerical Mathematics. Nauka, Moscow (1989) (in Russian)

Spectral Analysis of Large Sparse Matrices for Scalable Direct Solvers

Ahmet Duran, M. Serdar Celebi, Mehmet Tuncel, and Figen Oztoprak

Abstract It is significant to perform structural analysis of large sparse matrices in order to obtain scalable direct solvers. In this paper, we focus on spectral analysis of large sparse matrices. We believe that the approach for exception handling of challenging matrices via Gerschgorin circles and using tuned parameters is beneficial and practical to stabilize the performance of sparse direct solvers. Nearly defective matrices are among challenging matrices for the performance of solver. Such matrices should be handled separately in order to get rid of potential performance bottleneck. Clustered eigenvalues observed via Gerschgorin circles may be used to detect nearly defective matrix. We observe that the usage of super-nodal storage parameters affects the number of fill-ins and memory usage accordingly.

Keywords Spectral analysis • Sparse solver • Defective matrices

1 Introduction

We design and implement a new hybrid algorithm and solver for large sparse linear systems. We consider scalable direct solvers because of their robustness and examine the SuperLU_DIST 3.3 for distributed memory parallel machines

A. Duran (✉)
Department of Mathematics, National Center for High Performance Computing of Turkey (UHeM), Istanbul Technical University, Istanbul 34469, Turkey
e-mail: aduran@itu.edu.tr, http://web.itu.edu.tr/aduran

M.S. Celebi • F. Oztoprak
Informatics Institute, Istanbul Technical University, Istanbul 34469, Turkey

M. Tuncel
Department of Mathematics and Informatics Institute, Istanbul Technical University, Istanbul 34469, Turkey

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_14,

among several sparse direct solvers (see Li et al. [1], Li and Demmel [2], Amestoy et al. [3], Schenk and Gartner [4, 5], Duran and Saunders [6], Duran et al. [7], and references contained therein). Duran et al. [8] and Celebi et al. [9] discussed the advantages and limitations of the SuperLU solvers and tested the code of SuperLU_DIST 3.0 in order to measure the performance scalability for various patterned sparse matrices and randomly populated sparse matrices (see [10] for the theoretical foundation regarding the distribution of eigenvalues for some sets of random matrices). Although the existing versions of SuperLU work well for many matrices, they need to be improved for certain types of sparse matrices.

It is important to estimate the elapsed time to solve large sparse linear systems for time-restricted real-life decision-making applications such as oil and gas reservoir simulators and financial applications (see [11–13] and references therein). Challenging matrices should be distinguished and handled separately because they may lead to performance bottleneck. Therefore, structural analysis of large sparse matrices for scalable direct solvers is needed. In this work, we focus on spectral analysis of large sparse matrices and check whether there is relationship between the eigenvalue distribution of matrix and the performance of the solver. We try to examine the eigenvalue distribution of various sparse matrices. We may find all eigenvalues in order to obtain the distribution graph of eigenvalues, if possible. However, it is very expensive to find all eigenvalues. Therefore, Gerschgorin's theorem may be used to bound the spectrum of square matrices. Several behaviors such as being disjoint, overlapped, or clustered of, Gerschgorin circles may give clue regarding the distribution of the eigenvalues and the performance of the solver for that matrix.

The presence of repeated eigenvalues can be one of the sources of challenges. The repeated eigenvalue may have fewer eigenvectors than the multiplicity of eigenvalue. While such eigenvalue is called defective eigenvalue, the corresponding matrix is referred as a defective matrix (see [14]). If the matrix of eigenvectors is singular, then the matrix cannot be diagonalizable and the matrix is defective. We observe that it takes longer time to solve sparse linear system having defective or nearly defective matrix than regular matrix. Moreover, defective matrix may lead to memory restriction due to the appearance of more fill-ins than that of diagonalizable matrix.

The existing versions of SuperLU are sensitive to challenging matrices and need exception handling. Apart from the solver, spectral analysis can be done and tuned parameters may be used accordingly. The exception handling is one of the new properties of SuperLU_MCDT (Multi Core Distributed) solver (see Duran et al. [8] and Celebi et al. [9]). The remainder of this work is organized as follows. In Sect. 2, the test matrices including randomly populated matrices and patterned matrices are described. Later, the computation for spectral properties is presented and several illustrative examples are given. Section 3 concludes this work.

Table 1 Description of patterned matrices

Name	Order	NNZ	NNZ/N	Pattern symmetry	Numeric symmetry	Origin
EMILIA_923	923136	40373538	43,74	100 %	100 %	UFSMC
HELM2D03LOWER_20K	392257	1939353	4,94	0 %	0 %	UHeM

Table 2 Description of randomly populated matrices

Name	Order	NNZ	NNZ/N	Condition number	Origin
RAND_30K_3	30000	90000	3	1,20E+006	UHeM
RAND_30K_5	30000	150000	5	4,22E+006	UHeM
RAND_30K_7	30000	210000	7	1,76E+006	UHeM
RAND_30K_9	30000	270000	9	2,51E+006	UHeM
RAND_30K_11	30000	330000	11	8,82E+005	UHeM
RAND_30K_30	30000	900000	30	1,13E+006	UHeM
RAND_30K_50	30000	1500000	50	7,03E+005	UHeM
RAND_30K_75	30000	2250000	75	1,16E+006	UHeM
RAND_30K_100	30000	3000000	100	3,39E+006	UHeM
RAND_10K_3	10000	30000	3	7,10E+005	UHeM
RAND_20K_3	20000	60000	3	3,19E+005	UHeM
RAND_30K_3	30000	90000	3	1,20E+006	UHeM
RAND_40K_3	40000	120000	3	3,90E+006	UHeM
RAND_50K_3	50000	150000	3	1,20E+006	UHeM
RAND_60K_3	60000	180000	3	2,14E+006	UHeM

2 Methods and Results

We consider a portfolio of test matrices containing randomly populated sparse matrices in addition to patterned matrices. We generate 30 different randomly populated matrices RAND_30K_3, …, RAND_30K_100 for each. We describe the matrices in Tables 1 and 2, respectively.

2.1 Description of Matrices

2.2 Computation for Spectral Properties

The selected eigenvalues of large matrices are computed using the Scalable Library for Eigenvalue Problem Computations (SLEPc) software (see [15]), which is developed based on the Portable, Extensible Toolkit for Scientific Computation (PETSc) (see [16]). The code has been tested up for all sparse matrices in the list on HP Integrity Superdome SD32B (see [17]), a computing server with shared memory architecture at UHeM. The software package includes implementations of a set of

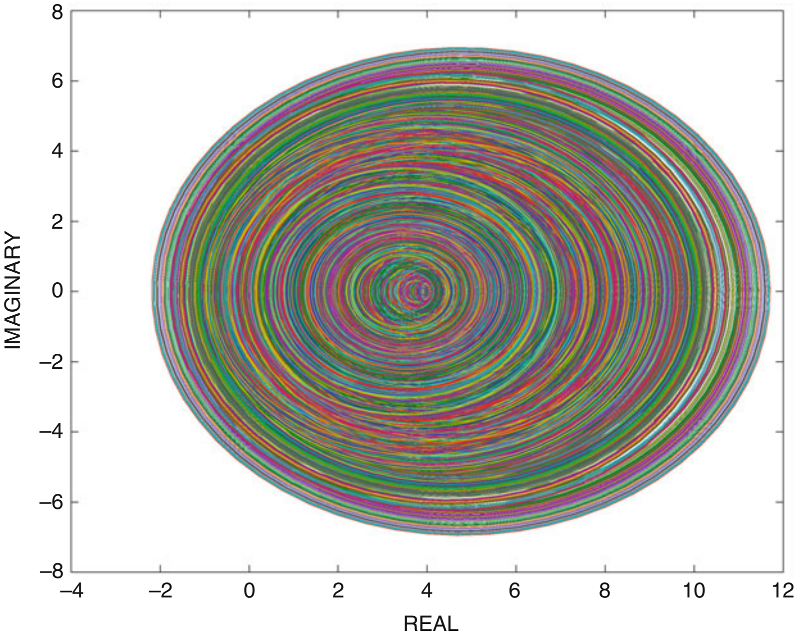


Fig. 1 Gerschgorin circles of matrix HELM2D03LOWER_20K

methods for the solution of large sparse eigenproblems on parallel computers. It is applicable to both symmetric and nonsymmetric matrices. In our computations, we used the Krylov-Schur method available in the package.

The computation of all eigenvalues may not be feasible for large sparse matrices, mainly due to memory constraints. Therefore, we followed two strategies to get an idea about the eigenvalue distribution of the test matrices: For the large sparse matrices, we compute the extreme eigenvalues. We try to see a rough picture of the distribution for the rest of the eigenvalues by using Gerschgorin's theorem. For example, we show the Gerschgorin circles of matrix HELM2D03LOWER_20K and matrix EMILIA_923 in Figs. 1 and 2, respectively.

We can compute all eigenvalues of the small randomly populated matrices and show the distribution of eigenvalues for RAND_30K_100 in Fig. 3. We observe that nearly all eigenvalues can be found within the circle except for the largest eigenvalue that is indicated by cross in figure.

Although the existing versions of SuperLU work well for many reasonable matrices, they need to be improved for certain types of sparse matrices. For example, we generated a new unsymmetric matrix HELM2D03LOWER_20K (see Duran et al. [8]), shown in Fig. 4, which consists of the lower triangular part of a symmetric matrix HELM2D03 from the University of Florida sparse matrix collection [18] and an upper subdiagonal with 20,000 distance from the main diagonal. We reported in our PRACE WP43 paper (see Duran et al. [8]) that SuperLU_DIST 3.0 failed for

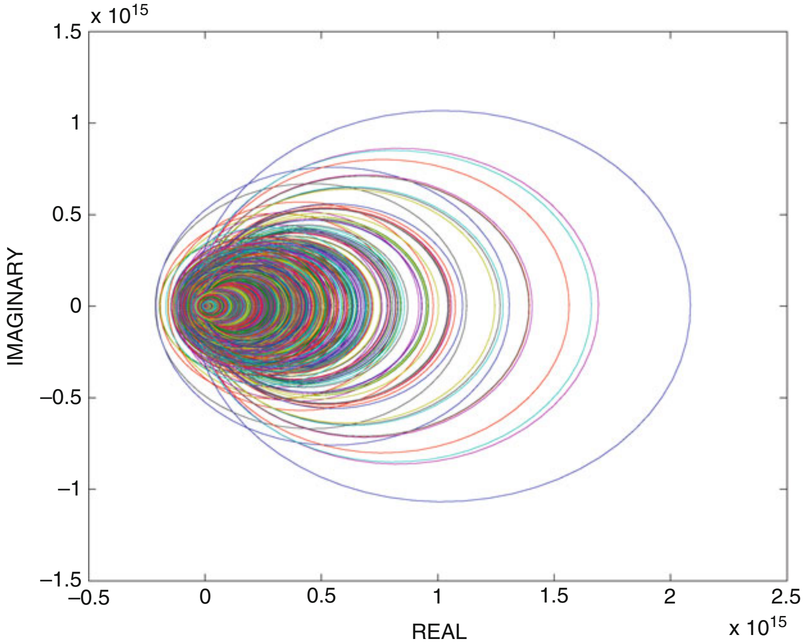
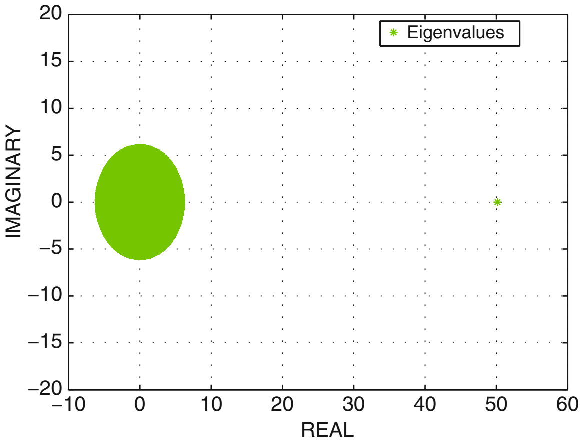


Fig. 2 Gerschgorin circles of matrix EMILIA_923

Fig. 3 Distribution of eigenvalues for matrix RAND_30K_100



HELM2D03LOWER_20K due to symbolic factorization error, although it works well for the matrix HELM2D03 on the Linux Nehalem Cluster (see [19]) available at UHeM. Later, the bug in the factorization routine was fixed in April 2013.

We used the SuperLU_DIST 3.3 with tunings of super-nodal storage parameters. However, it runs slowly for the matrix HELM2D03LOWER_20K compared to EMILIA_923 (see Table 3), because HELM2D03LOWER_20K is a challenging

Fig. 4 Matrix picture of HELM2D03LOWER_20K

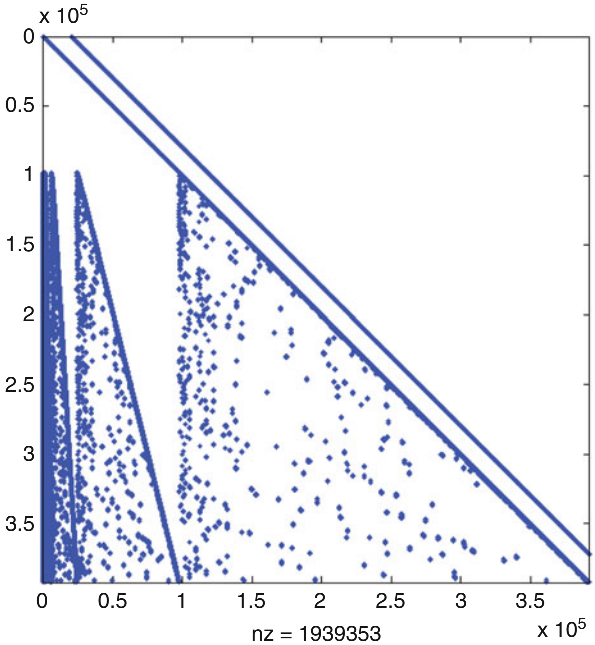


Table 3 The performance of the SuperLU_DIST 3.3 for HELM2D03LOWER_20K and EMILIA_923

Wall clock time(s)	BLAS		MKL	
Patterned matrices	Default parameters	Tuned parameters	Default parameters	Tuned parameters
HELM2D03LOWER_20K	5,594.72	3,047.56	5,310.04	2,324.00
EMILIA_923	743,29			

matrix. It takes approximately 7.5 times longer than EMILIA_923, although HELM2D03LOWER_20K's order, total number of nonzeros, and the number of nonzeros per row are less than that of EMILIA_923. Table 3 shows the performance of the SuperLU_DIST 3.3 for HELM2D03LOWER_20K and EMILIA_923 by using standard BLAS [20] and Intel's Math Kernel Library (MKL) [21] which is a kind of optimized BLAS on the Linux Nehalem Cluster with 64 (8 × 8 mesh) cores. The tunings of super-nodal storage parameters are important. For example, the usage of tuned parameters (relax:100 and maxsuper:110) outperforms at least 1.8 times faster than that of default parameters (relax:12 and maxsuper:60) for HELM2D03LOWER_20K using the SuperLU_DIST 3.3. Moreover, the usage of super-nodal storage parameters affects the number of fill-ins. For instance, there are 3,208,629,380 nonzeros in L+U with the default parameters compared to 3,477,287,771 nonzeros of L+U in the presence of the tuned parameters.

When we examine the spectral properties of HELM2D03LOWER_20K in Fig. 1, the real parts of the eigenvalues range between 2.294563 and 4.944602 with many repeated eigenvalues. Those clustered eigenvalues can be observed via Gerschgorin circles as in Fig. 1. Therefore, HELM2D03LOWER_20K is a nearly defective matrix.

3 Conclusions

We believe that the approach of exception handling of challenging matrices via Gerschgorin circles and using tuned parameters is beneficial and practical to stabilize the performance of sparse direct solvers. Nearly defective matrices are among challenging matrices. Such matrices should be handled separately in order to get rid of potential performance bottleneck. Clustered eigenvalues observed via Gerschgorin circles may be used to detect nearly defective matrix.

We reported in our PRACE WP43 paper (see Duran et al. [8]) that SuperLU_DIST 3.0 failed for HELM2D03LOWER_20K due to symbolic factorization error. Later, the bug in the factorization routine was fixed in April 2013. We noticed that the SuperLU_DIST 3.3 with tunings of super-nodal storage parameters works for HELM2D03LOWER_20K but slowly.

The tunings of super-nodal storage parameters are important. For example, the usage of tuned parameters outperforms at least 1.8 times faster than that of default parameters for HELM2D03LOWER_20K using the SuperLU_DIST 3.3. Moreover, we observe that the usage of super-nodal storage parameters affects the number of fill-ins and memory usage.

Acknowledgements This work was financially supported by the PRACE project funded in part by the EUs 7th Framework Programme (FP7/2011–2013) under grant agreement no. RI-283493. Computing resources used in this work were provided by the National Center for High Performance Computing of Turkey (UHeM) (see [22]) under grant number 1001682012.

References

1. Li, X.S., Demmel, J.W., Gilbert, J.R., Grigori, L., Shao, M., Yamazaki, I.: SuperLU users' guide. Technical Report UCB. Computer Science Division, University of California, Berkeley, update: 2011 (1999)
2. Li, X.S., Demmel, J.W.: Superlu-dist: a scalable distributed-memory sparse direct solver for unsymmetric linear systems. ACM Trans. Math. Softw. **29**, 110–140 (2003)
3. Amestoy, P.R., Duff, I.S., J.-Y. L'Excellent, Koster, J.: A fully asynchronous multifrontal solver using distributed dynamic scheduling. SIAM J. Matrix Anal. Appl. **23**, 15–41 (2001)
4. Schenk, O., Gartner, K.: Solving unsymmetric sparse systems of linear equations with PARDISO. Future Generat. Comput. Syst. **20**, 475–487 (2004)
5. Schenk, O., Gartner, K.: On fast factorization pivoting methods for sparse symmetric indefinite systems. Electron. Trans. Numer. Anal. **23**, 158–179 (2006)

6. Duran, A., Saunders, B.D.: Gen_SuperLU package (version 1.0, August 2002), referenced as GSLU also, a part of LinBox package. GSLU contains a set of subroutines to solve a sparse linear system A*X=B over any field. http://web.itu.edu.tr/~aduran/Gen_SuperLU.pdf
7. Duran, A., Saunders, B.D., Wan, Z.: Hybrid algorithms for rank of sparse matrices. In: Proceedings of the SIAM International Conference on Applied Linear Algebra (SIAM-LA), Williamsburg, VA, 15–19 July 2003
8. Duran, A., Celebi, M.S., Tuncel, M., Akaydin, B.: Design and implementation of new hybrid algorithm and solver on CPU for large sparse linear systems. PRACE-2IP White Paper, Libraries, WP 43. http://www.prace-ri.eu/IMG/pdf/wp43-newhybridalgorithmfo_lsls.pdf. Accessed 13 July 2012
9. Celebi, M.S., Duran, A., Tuncel, M., Akaydin, B.: Scalable and improved SuperLU on GPU for heterogeneous systems. PRACE-2IP White Paper, Libraries, WP 44. http://www.prace-ri.eu/IMG/pdf/scalablesuperluongpu.pdf. Accessed 13 July 2012
10. Marchenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sb **83**(7), 457–486 (1967)
11. Duran, A., Bommarito, M.J.: A profitable trading and risk management strategy despite transaction cost. Quant. Finance **11**(6), 829–848 (2011)
12. Dogru, A.H., Fung, L.S.K., Middya, U., Al-Shaalan, T.M., Pita, J.A., Kumar, K.H., Su, H.J., Hoy, H., Al-Harbi, R., Tan, J.C.T., Dreiman, W.T., Hahn, W.A., Mezghani, M., Al-Zamel, N.M., Al-Youbi, A., Al-Mani, T.: A next-generation parallel reservoir simulator for giant reservoirs. SPE 119272 (2009)
13. Duran, A., Celebi, M.S., Tuncel, M., Akaydin, B.: Scalability of SuperLU solvers for large scale complex reservoir simulations. In: SPE and SIAM Conference on Mathematical Methods in Fluid Dynamics and Simulation of Giant Oil and Gas Reservoirs, Istanbul, Turkey, 3–5 Sept 2012
14. Strang, G.: Linear Algebra and Its Applications, 3rd edn. Harcourt, San Diego (1988)
15. http://www.grycap.upv.es/slepc
16. http://www.mcs.anl.gov/petsc
17. nPartition Administrator's Guide.: HP Part Number: 5991-1247B, 1st edn. Hewlett-Packard Development Company, Palo Alto (2007)
18. Davis, T.A.: University of Florida sparse matrix collection. http://www.cise.ufl.edu/research/sparse/matrices/
19. http://www.uybhm.itu.edu.tr/eng/inner/duyurular.html#karadeniz
20. http://www.netlib.org/blas
21. http://software.intel.com/en-us/intel-mkl
22. http://www.uybhm.itu.edu.tr

Numerical Study of Two-Dimensional Jet Flow Issuing from a Funnel

Abdelkader Gasmi

Abstract In this paper, the problem of steady two-dimensional flow emerging from a slot of a funnel is considered. The fluid is assumed to be incompressible and inviscid and the flow is irrotational. The problem is reformulated using conformal mappings and the resulting problem is then solved by using the series truncation method. We computed solutions for various values of the Weber numbers. The contraction coefficient for different forms of the funnel has been found. The shape of the free surface of the jet has been determined and presented.

Keywords Free surface • Inviscid flow • Weber number • Jet • Series truncation

1 Introduction

In this work, we consider two-dimensional potential flow of an incompressible and inviscid fluid emerging from an opening located at the end of a semi-infinite tube. The width of the tube is $2H_0$ and the inclination angle of the end faces of the tube to the horizontal is β (see Fig. 1). The surface tension effects are incorporated into the nonlinear free-surface boundary conditions and gravity is neglected. Far upstream the flow is uniform with a constant velocity U_0, the thickness of the fluid is $2H$ and the velocity approaches a constant U far downstream. The mathematical problem is characterized by four parameters, the width of the tube $2H_0$, the inclination angle β, the angle at the separation point between the plate and the free streamlines γ, and the Weber number α. If one neglects the effect of surface tension, the solution can be computed exactly by using the free streamline theory proposed by Kirchhoff (1869) and the method of conformal mapping [1, 2].

A. Gasmi (✉)
Department of Computer Science, Faculty of Mathematics and Computer Science,
University of M'sila, M'sila, Algeria
e-mail: gasmi_a@yahoo.fr

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_15,

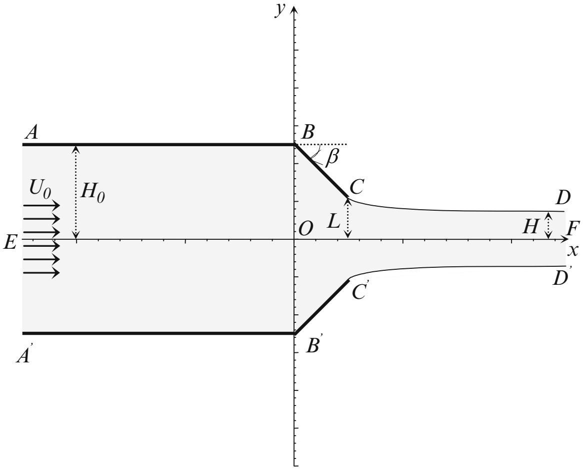


Fig. 1 Sketch of the flow and the coordinates

When the gravity or the effects of surface tension are included, the free-surface equation is generally a nonlinear partial differential equation. These equations are generally very difficult to solve analytically. In cases where the effect of surface tension is neglected and the effect of gravity is considered there are several variants of this problem, for example, the flow emerging from vessels, the flow under sluice gate, the flow over an obstacle, etc. Many authors have investigated these problems: Yoon and Semenov [3] and Vanden-Breock [4].

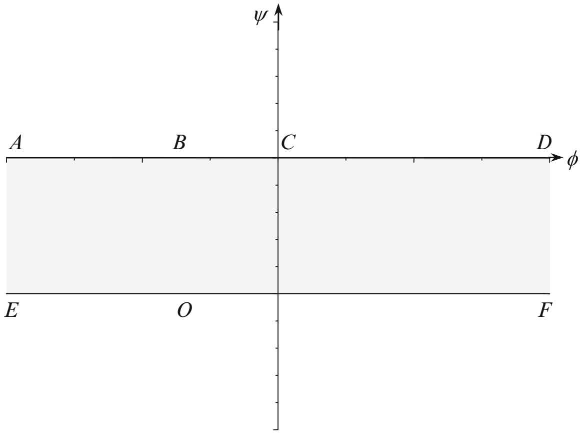
In this paper, we solved the fully nonlinear problem numerically and the mesh points were only on the free surface. For each value of the inclination angle β, we found that there exists a unique solution for all $\alpha \geq 0$. Gasmi and Mekias [5] considered the problem of a free-surface flow past an infinite flat plate in a channel; this configuration can be obtained when the inclination angle $\beta = \frac{\pi}{2}$. In this work we extend the calculations of Gasmi and Mekias [5, 6, 8].

The problem is formulated in Sect. 2. The numerical procedure is described in Sect. 3 and the results are presented and discussed in Sect. 4.

2 Formulation

Let us consider a two-dimensional steady irrotational flow issuing from an orifice of length $2L$ located at the end of a semi-infinite tube of width $2H_0$ in the presence of surface tension forces (see Fig. 1). The fluid is inviscid and incompressible. We introduce the cartesian coordinates with the streamline EOF on the x-axis and the y-axis is perpendicular and passing through the points B and B'. Far upstream the flow is uniform with a constant velocity U_0. Far downstream, we assume that the velocity approaches a constant U and the thickness of the fluid tends to a constant $2H$. Because of the symmetry of the flow field, we need only consider the half of the flow region which is contained between the x-axis and the streamline $ABCD$.

Fig. 2 The complex potential f-plane



The governing equations for the flow are

$$\Delta\phi = 0 \qquad \text{in the flow field,} \tag{1}$$

where ϕ is the velocity potential,

$$\frac{\partial\phi}{\partial\eta} = 0 \qquad \text{on the walls,} \tag{2}$$

where η is a normal vector of the boundaries,

$$\frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 + \frac{1}{2}\left(\frac{\partial\phi}{\partial y}\right)^2 - \frac{T}{\rho}K = cts \qquad \text{on the two free surfaces.} \tag{3}$$

Here ρ is the density, T is the surface tension, and K is the curvature of the free surface:

$$\phi \to cts \qquad x \to -\infty, \tag{4}$$

$$\phi \to Ux \qquad x \to +\infty. \tag{5}$$

We introduce the complex potential function f and the complex velocity ζ: $f = \phi + i\psi$, $\zeta = \frac{df}{dz} = u - iv$ in which ψ, u, and v represent, respectively, the stream function and the horizontal and vertical components of the fluid velocity.

The physical quantities are made dimensionless by using U as the velocity unit and L as the length unit. Without loss of generality, we choose $\phi = 0$ at the separation point C and $\psi = 0$ on the streamline $ABCD$. It follows from the choice of the dimensionless variables that $\psi = -1$ on streamline EOF, and the flow configuration in the complex potential plane is sketched in Fig. 2.

We define the function $\tau - i\theta$ as

$$w = u - iv = e^{\tau - i\theta}. \tag{6}$$

In these new variables (3) becomes

$$e^{2\tau} - \frac{2}{\alpha} e^{\tau} \frac{\partial\theta}{\partial\phi} = 1 \qquad 0 < \phi < \infty. \tag{7}$$

α is the Weber number defined by

$$\alpha = \frac{\rho U^2 L}{T}. \tag{8}$$

The kinematic condition on AB, BC, and EOF can be expressed as

$$\mathrm{Im}\zeta = 0 \qquad \text{on} \quad \psi = 0 \quad \text{and} \ -\infty < \phi < \phi_B, \tag{9}$$

$$\frac{\mathrm{Re}\zeta}{\mathrm{Im}\zeta} = \cot\beta \qquad \text{on} \quad \psi = 0 \quad \text{and} \ \phi_B < \phi < 0, \tag{10}$$

$$\mathrm{Im}\zeta = 0 \qquad \text{on} \quad \psi = -1 \quad \text{and} \ -\infty < \phi < \infty. \tag{11}$$

This completes the formulation of the problem of determining $\tau - i\theta$. This function must be analytic in the strip $-1 < \psi < 0$ and satisfies the conditions (7), (9), (10), and (11).

3 Numerical Procedure

The nonlinear flow problem is numerically solved via a series truncation technique, similar to that used by Vanden-Broeck. We define a new variable t by the relation

$$f = \frac{2}{\pi} \log\left(\frac{2it}{1-t^2}\right). \tag{12}$$

The wall ABC goes onto the imaginary interval $(0, i)$, the wall EOF onto the real interval $(0, 1)$, and the free surface CD onto the circumference (see Fig. 3). The domain of the fluid in the f-plane is then transformed into the first quadrant of the unit disk in a t-plane.

Since there is stagnation point, an angle at B and discontinuity in the derivative $\frac{\partial\theta}{\partial\phi}$ at the separation point $\phi = 0$, ζ must have singularities at these points. Local analysis shows that appropriate singularities are

$$\zeta \sim (b^2 + t^2)^{\frac{\beta}{\pi}} \qquad \text{as} \qquad t \longrightarrow ib. \tag{13}$$

$$\zeta \sim (t^2 + 1)^{1-\frac{\gamma}{\pi}} \qquad \text{as} \qquad t \longrightarrow i. \tag{14}$$

Note that the free surfaces in the t-plane are described by the points $t = e^{i\sigma}$.

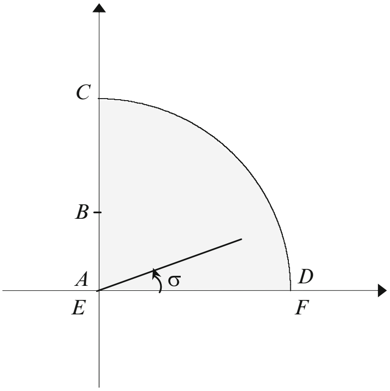


Fig. 3 The complex potential t-plane

Next we define the function $\Omega(t)$ by the relation

$$e^{\tau - i\theta} = \zeta(t) = (b^2 - t^2)^{\frac{\beta}{\pi}}(1 - t^2)^{1-\frac{\gamma}{\pi}} e^{\Omega(t)}. \tag{15}$$

At the points $t = i$ and $t = ib$, ζ has singularities associated with a flow around a corner and stagnant flow (see Fig. 1). These singularities are removed in (12) by the factor $(b^2 - t^2)^{\frac{\beta}{\pi}}(1 - t^2)^{1-\frac{\gamma}{\pi}}$ (see Birkhof and Zarantonello [1] and Vanden-Broeck [4] for details). It follows that $\Omega(t)$ can be represented by a Taylor expansion in powers of t. Furthermore, the kinematic conditions (9) imply that the expansion for a_n has real coefficients and involves only even powers of t. Thus we write

$$\Omega(t) = \sum_{n=1}^{\infty} a_n t^{2n}. \tag{16}$$

Using (12) we rewrite (7) in the form

$$e^{2\bar{\tau}} + \frac{\pi}{\alpha} e^{\bar{\tau}} \tan(\sigma) \frac{\partial \bar{\theta}}{\partial \sigma} = 1. \tag{17}$$

Here $\bar{\tau}(\sigma)$ and $\bar{\theta}(\sigma)$ denote the values of τ and θ on the free surface CD.
We solve the problem approximately by truncating the infinite series in (16) after N terms. We find the N coefficients a_n and the separation angle γ by collocation. Thus we introduce the $N + 1$ mesh points

$$\sigma_I = \frac{\pi}{2(N+1)}\left(I - \frac{1}{2}\right), \qquad I = 1, \ldots, N + 1. \tag{18}$$

Using (18) we obtain $[\bar{\tau}(\sigma)]_{\sigma=\sigma_I}$, $[\bar{\theta}(\sigma)]_{\sigma=\sigma_I}$ and $[\frac{\partial \bar{\theta}}{\partial \sigma}]_{\sigma=\sigma_I}$ in terms of coefficients a_n and the separation angle γ. This leads to a system of $N + 1$ nonlinear algebraic equations of $N + 1$ unknown $(a_{n,\ n=1,\ldots,N}, \gamma)$. This system is also solved by Newton's method.

4 Results

Numerical schemes of Sect. 3 were used to compute solutions for different values of the inclination angle β and several values of the Weber number α. We found that the coefficients a_n decrease rapidly as n increases. Figure 4 presents the variation of $\log|a_n|$ against n for $\alpha = 100$ and $\beta = \frac{\pi}{3}$, which shows how the coefficients a_n decrease rapidly. Some of the coefficients of the series (16) and the corresponding Weber number for different values of β are shown in Table 1.

Most of the calculations were done and presented with $N = 60$.

We note that as the Weber number α decreases, the contraction coefficient C_c and the angle in the separation point increase. Numerical values of C_c vs. $\frac{1}{\alpha}$ are shown in Fig. 5.

In Fig. 6 we present values of the angle at the separation point between the wall and the free streamlines γ vs. $\frac{1}{\alpha}$. It is seen that numerical solutions exist for all $\alpha \geq 0$.

As $b \to 0$ and for all values of the inclination angle $0 \leq \beta \leq \frac{\pi}{2}$ we obtain the same results as Gasmi and Mekias [6] and our result also agrees with the results

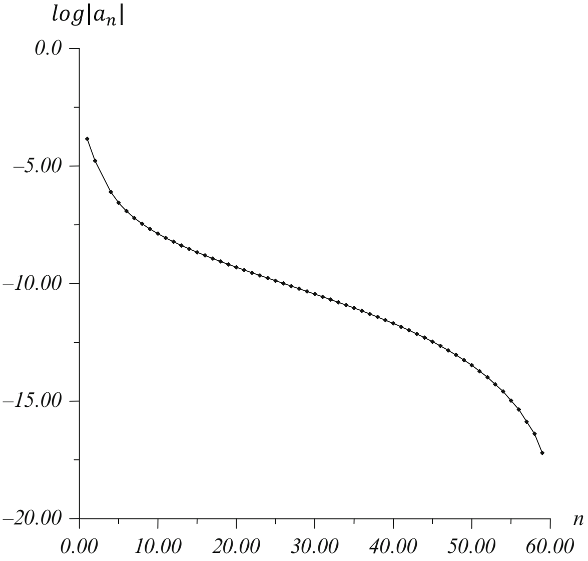


Fig. 4 The variation of $\log|a_n|$ against n for $\alpha = 100$ and $\beta = \frac{\pi}{3}$

Table 1 Some values of the coefficients a_n of the series (12) for several values of the angle β, $b = 0.5$ and different values of Weber number α

β	α	a_1	a_{20}	a_{40}	a_{60}
$\frac{\pi}{3}$	1.5	2.7332×10^{-1}	4.8052×10^{-5}	1.0300×10^{-5}	5.7705×10^{-7}
	10	2.2080×10^{-1}	1.3708×10^{-4}	1.4792×10^{-5}	1.7920×10^{-6}
	$\alpha \to \infty$	1.1387×10^{-8}	3.1298×10^{-9}	2.8399×10^{-10}	1.4792×10^{-13}
$\frac{\pi}{6}$	1.5	3.4142×10^{-1}	5.6654×10^{-5}	1.1014×10^{-6}	6.0315×10^{-8}
	10	2.4580×10^{-1}	1.3470×10^{-4}	1.4681×10^{-5}	1.8478×10^{-6}
	$\alpha \to \infty$	1.4871×10^{-7}	4.1034×10^{-8}	3.1948×10^{-9}	1.4145×10^{-12}

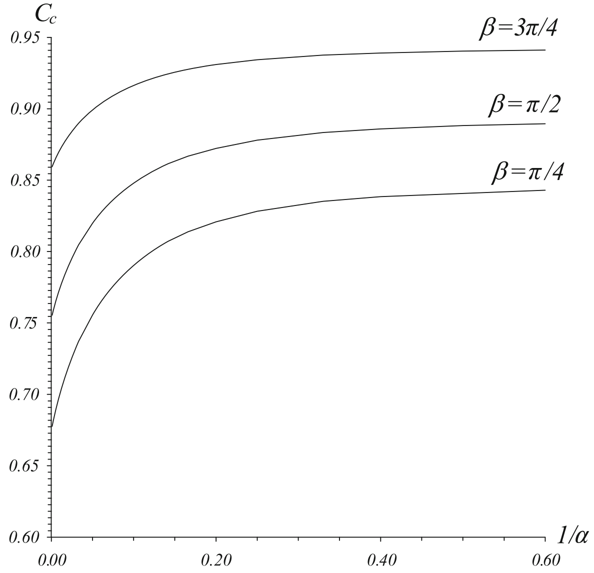


Fig. 5 Coefficient of contraction C_c vs. $1/\alpha$

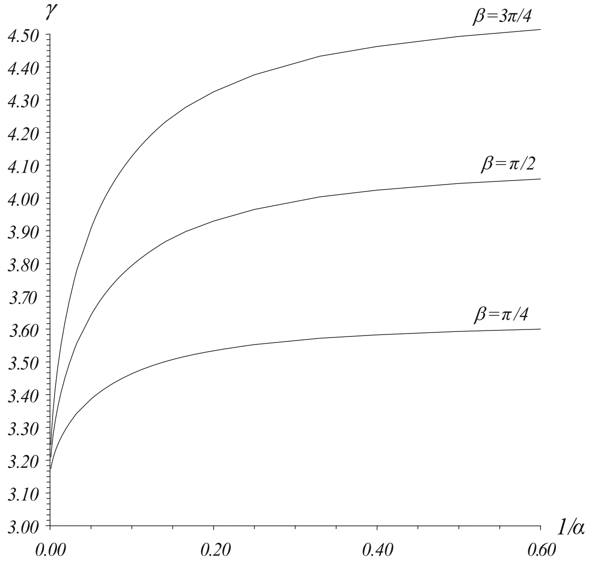


Fig. 6 The angle of separation point γ vs. $1/\alpha$

of Ackerberg and Liu [7] for different Weber number $\alpha \geq \tilde{\alpha} = 6.801483$. These authors solved the problem via the finite difference method and the mesh points were throughout the fluid domain; they could find solution for all $\alpha \geq \tilde{\alpha} = 6.801483$. In our procedure mesh points are only needed on the free surface and we computed solutions for $\alpha \geq 0$.

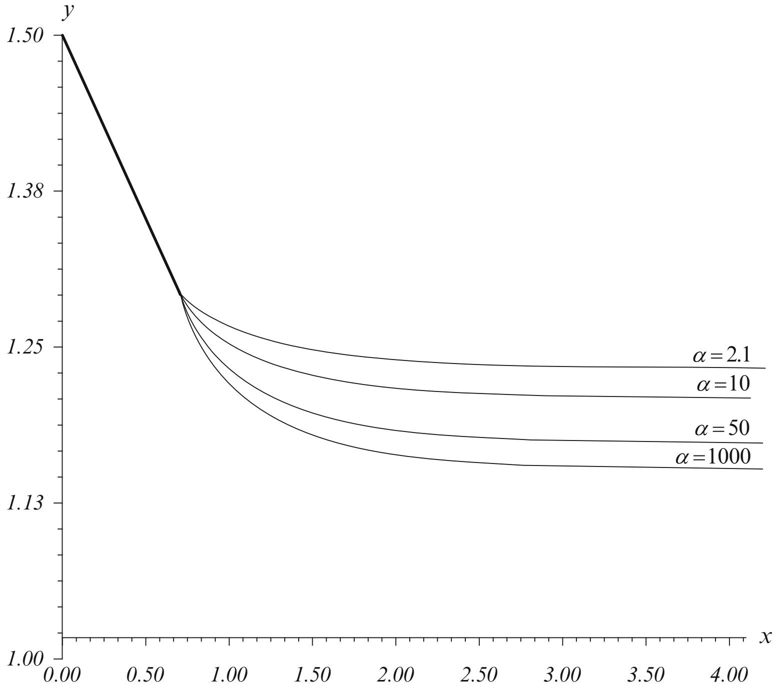


Fig. 7 Free streamline shape for $\beta = \frac{\pi}{6}$, $H_0 - L = 1$ and various Weber numbers

It is also observed that where $\alpha \to 0$, the free surface tends to a horizontal line, the contraction coefficient $C_c \to 1$, and the angle of separation $\gamma \to 3\pi/2$. For this limiting case, all boundaries are rectilinear; hence, an exact solution can be found via Schwarz-Christoffel transformation [1].

Typical profiles for various Weber numbers of the free surface are presented in Fig. 7 for $H_0 - L = 1$ and $\beta = \frac{\pi}{4}$. This work is a generalization of our previous paper [6].

5 Conclusion

We have employed the series truncation method to find the velocity and the shape of the jet flow. We also have precisely determined the type of singularity at the separation points. The obtained solution showed that the surface tension forces act in such a way as to reduce the free-surface curvature of the jet. The main advantage of this method is to reduce the two-dimensional problems into one-dimensional and find the solution only on the boundary of the flow field.

Acknowledgements This work was supported in part by the Algerian Ministry of Higher Education and Research under the CNEPRU-Project: B05620110014.

References

1. Birkhof, G., Zarantonello, E.H.: Jets, Wakes and Cavities. Academic, New York (1957)
2. Batchelor, G.K.: An Introduction to Fluid Dynamics. Cambridge University Press, Cambridge (1967)
3. Yoon, B.S., Semenov, Y.A.: Capillary cavity flow past a circular cylinder. Eur. J. Mech. B/Fluids **28**, 670 (2009)
4. Vanden-Broeck, J.-M.: Gravity-Capillary Free-Surface Flows. Cambridge University Press, Cambridge (2010)
5. Gasmi, A., Mekias, H.: A jet from container and flow past a vertical flat plate in a channel with the surface tension effects. Appl. Math. Sci. **1**(54), 2687 (2007)
6. Gasmi, A., Mekias, H.: The effect of surface tension on the contraction coefficient of a jet. J. Phys. A Math. Gen. **36**, 851 (2003)
7. Ackerberg, R.C., Liu, T.-J.: The effects of capillarity on the contraction coefficient of a jet emanating from a slot. Phys. Fluids **30**, 289–90 (1987)
8. Gasmi, A.: Two-dimensional cavitating flow past an oblique plate in a channel. J. Comput. Appl. Math. **259**, 828–834 (2014)

Higher-Order Immersed Finite Element Spaces for Second-Order Elliptic Interface Problems with Quadratic Interface

Mohamed Ben-Romdhane, Slimane Adjerid, and Tao Lin

Abstract In this manuscript, we present quadratic immersed finite element (IFE) spaces to be used with the interior penalty IFE method proposed in Adjerid (Int. J. Numer. Anal. Model., 2013, accepted) to solve interface problems with a quadratic interface. Quadratic IFE spaces for interface problems with quadratic interfaces are developed using an affine mapping between the reference and the physical elements. Two different approaches for imposing the interface jump conditions are proposed: (i) a weak form of jump conditions using Legendre polynomials and (ii) a pointwise form by imposing the conditions at some particular points. We give a procedure to construct IFE shape functions, investigate the optimal approximation capability of the proposed IFE spaces, and present numerical results showing optimal convergence.

Keywords Quadratic finite element spaces • Interface problem • Quadratic interface • Interior penalty immersed finite element method

1 Introduction and Model Problem

Interface problems with discontinuous coefficients across the interface are encountered in many engineering and scientific problems such as problems in material sciences, electromagnetism, fluid dynamics, and biological processes [1]. Solving such interface problems efficiently and accurately remains a challenge because

M. Ben-Romdhane (✉)
Department of Mathematics and Natural Sciences, Gulf University for Science
and Technology, P.O. Box: 7207, Hawally 32093, Kuwait
e-mail: romdhane.m@gust.edu.kw

S. Adjerid • T. Lin
Department of Mathematics, Virginia Polytechnic Institute
and State University, Blacksburg, VA 24061, USA

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_16,

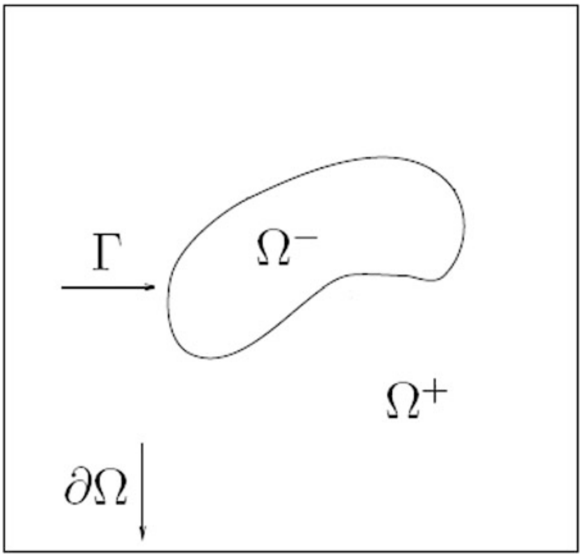


Fig. 1 A two-material domain Ω

of the non-smoothness or the discontinuity of the input data and/or the solutions as well as the non-smoothness of the interface geometry in some applications. To handle the mentioned type of interface problems, the immersed finite element (IFE) methods are proposed. Unlike the standard finite element method where body-fitted meshes are used in order to guarantee optimal convergence of the solutions [2–4], the IFE method does not require any restrictions on the mesh used to solve the interface problem. It allows using finite elements that are cut by the interface which eliminates the need for using body-fitted meshes and uses interface-independent meshes to solve interface problems. As a result, meshes used by IFE methods consist of the following two types of elements: (i) *non-interface elements* which do not intersect the interface and are equipped with standard local FE basis functions and (ii) *interface elements* which are cut by the interface and are equipped with IFE basis functions satisfying interface jump conditions.

We consider the following model interface problem:

$$\begin{cases} -\nabla(\beta\nabla u) = f, & \text{in} \quad \Omega, \\ u_{|\partial\Omega} = g. \end{cases} \tag{1a}$$

Without loss of generality, we assume that $\Omega \subset \mathbb{R}^2$ is a rectangular domain consisting of two sub-domains Ω^+ and Ω^- separated by an interface Γ as illustrated in Fig. 1, while β is given by

$$\beta(x, y) = \begin{cases} \beta^+, & \text{in } \Omega^+, \\ \beta^-, & \text{in } \Omega^-, \end{cases} \tag{1b}$$

where β^+ and β^- are two positive constants. Following the same notations as [5], we let $\mathcal{P}_2$ denote the two-dimensional quadratic polynomial space in $\mathbb{R}^2$, and let

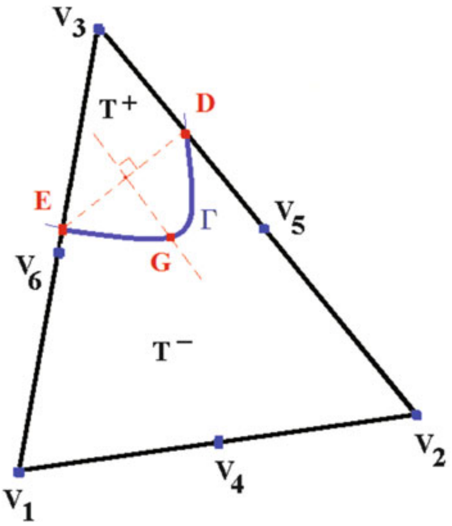


Fig. 2 A physical interface element

$\mathscr{T}_h$ be a regular triangulation of the domain Ω, where h is the maximum diameter. The set of interface elements that are cut by the interface is denoted by $\mathscr{T}_h^i$, and we call the set of non-interface elements as $\mathscr{T}_h^c = \mathscr{T}_h \setminus \mathscr{T}_h^i$. Similarly, edges that are cut by the interface are called interface edges; otherwise, they are referred to as non-interface edges. We let $\mathscr{E}_h$, $\mathscr{E}_{h,0}$, $\mathscr{E}_h^i$, respectively, denote the set of all edges, interior edges, and interface edges. As illustrated in Fig. 2, every interface element T can be split as $T = T^+ \cup T^-$, where $T^\pm = T \cap \Omega^\pm$.

In the discussion from now on, for a triangular element $T = \Delta V_1 V_2 V_3$, we will use V_4, V_5, and V_6 to denote the midpoints of the edges V_1V_2, V_2V_3, and V_1V_3, respectively.

We will also use D and E to denote the intersection points of Γ with the edges of T. Guided by the standard isoparametric finite element ideas [6], we will denote by G the intersection point of Γ with the line orthogonal to the line segment DE and passing through its midpoint, as shown in Fig. 2.

2 Piecewise Quadratic IFE Spaces for Quadratic Interfaces

We introduce the following piecewise quadratic IFE spaces on an arbitrary interface element T:

$$\mathcal{R}_1(T) = \{U, \mid U|_{T^\pm} \in \mathcal{P}_2, \int_{T\cap\Gamma} [U]_{T\cap\Gamma}\, v_i \, ds = 0, \; i = 0, 1, 2,$$

$$\int_{T\cap\Gamma} [\, \beta \, \boldsymbol{n} \cdot \nabla U \,]_{T\cap\Gamma}\, v_i \, ds = 0, \; i = 0, 1,$$

$$\int_{T\cap\Gamma} [\,\beta\; \Delta U\,]_{T\cap\Gamma}\; v_0\; ds = 0\,\}, \tag{2}$$

$$\mathcal{R}_2(T) = \{U,\; |\; U|_{T^\pm} \in \mathcal{P}_2,\; [U]_{D,E,G} = [\,\beta\, \boldsymbol{n}\cdot\nabla U\,]_{D,E} = [\,\beta\; \Delta U\,]_G = 0\,\}, \tag{3}$$

with $v_i,\ i = 0, 1, 2$, being the mappings from the reference interface element $\hat{T}$ to the physical element T, of the three one-dimensional Legendre polynomials $\hat{v}_i,\ i = 0, 1, 2$, of degree i, shifted to the interface $\hat{\Gamma}$ on the reference element.

To define the global quadratic IFE spaces over the whole simulation domain Ω, we define the set of nodes $\mathcal{N}_h$ for the usual Lagrange quadratic finite element space defined on the mesh $\mathcal{T}_h$, and for each node $\mathbf{v}_i \in \mathcal{N}_h$, we define a piecewise quadratic IFE basis function ψ_i^k, for $k = 1, 2$, over Ω as follows:

$$\psi_i^k|_T \in \begin{cases} \mathcal{R}_k(T)\ \forall\ T \in \mathcal{T}_h^i \\ \\ \mathcal{P}_2\ \forall\ T \in \mathcal{T}_h \setminus \mathcal{T}_h^i \end{cases}, \qquad \psi_i^k(\mathbf{v}_j) = \delta_{ij}\ \forall\ \mathbf{v}_j \in \mathcal{N}_h.$$

We also recall that $\mathcal{N}_h$ contains N nodes among which the first N_I nodes are inside Ω while the rest of them are on the boundary of Ω, and we define the global quadratic IFE spaces over the domain Ω as

$$\mathcal{W}_h = \text{span}\{\psi_j^1,\ j = 1, \ldots, N\}, \tag{4}$$

and

$$\mathcal{J}_h = \text{span}\{\psi_j^2,\ j = 1, \ldots, N\}, \tag{5}$$

and the subsets of the spaces $\mathcal{W}_h$ and $\mathcal{J}_h$, consisting of functions interpolating the essential boundary condition g

$$\mathcal{W}_{h,E} = \left\{ U \in \mathcal{W}_h \,|\, U = \sum_{i=1}^{N_I} c_i \psi_i^1 + \sum_{i=N_I+1}^{N} g(\mathbf{v}_i)\psi_i^1 \right\}. \tag{6}$$

and

$$\mathcal{J}_{h,E} = \left\{ U \in \mathcal{J}_h \,|\, U = \sum_{i=1}^{N_I} c_i \psi_i^2 + \sum_{i=N_I+1}^{N} g(\mathbf{v}_i)\psi_i^2 \right\}. \tag{7}$$

In a forthcoming paper, we will detail the mapping of the piecewise quadratic IFE spaces $\mathcal{R}_k(T),\ k = 1, 2$, defined on an arbitrary element T, into the reference interface triangle and discuss the new interface jump conditions on the reference element.

3 Approximation Capability

For all our numerical experiments, we consider the rectangular domain $\Omega = [0,1]^2$ and the uniform triangular mesh $\mathscr{T}_h = \mathscr{T}_h^i \cup \mathscr{T}_h^c$ of size h. $\mathscr{T}_h$ is formed by partitioning Ω into $(1/h)^2$ squares, with $h = \frac{1}{2^m}$, $m = 2, \ldots, 7$, then forming the triangular elements by joining the lower right and upper left vertices of the squares. We define a piecewise Lagrange-type IFE interpolant $I_h u(x,y)$ of $u(x,y)$ such that $\forall\, T \in \mathscr{T}_h$

$$I_h u(x,y)|_T = \sum_{i=1}^{6} u(V_i)\phi_i(x,y), \tag{8}$$

where V_i, $i = 1, \ldots, 6$, are the nodes on T and $\phi_i(x,y)$, $i = 1, \ldots, 6$, are the six Lagrange FE or IFE shape functions depending on whether T is a non-interface or an interface element.

To compute the interpolation errors, we use the usual L^2 norm and the following broken H^1 norm

$$||u_t - (I_h u)_t||_{0,h}^2 = \sum_{T \in \mathscr{T}_h} \iint_T (u_t - (I_h u)_t)^2 dxdy,\ t = x, y, \tag{9}$$

using the constructed Lagrange piecewise quadratic IFE shape functions and discuss the approximation capability of the piecewise quadratic IFE space $\mathcal{J}_h$.

Example 3.1. We consider the domain $\Omega = [0,1]^2$ cut by the quadratic interface $\Gamma : x^2 + \frac{\pi}{7} = y$, as illustrated in Fig. 3. Let us denote $\Omega^+ = \{x^2 + \frac{\pi}{7} < y\}$ and $\Omega^- = \{x^2 + \frac{\pi}{7} > y\}$. We test our IFE space on the piecewise function

$$u(x,y) = \begin{cases} \frac{1}{\beta^+}\left(y^3 - \left(x^2 + \frac{\pi}{7}\right)^3\right) e^{x+y} \ ; & \text{in } \Omega^+ \\ \frac{1}{\beta^-}\left(y^3 - \left(x^2 + \frac{\pi}{7}\right)^3\right) e^{x+y} \ ; & \text{in } \Omega^-, \end{cases} \tag{10}$$

with $r = \frac{\beta^+}{\beta^-} = 5$ and $r = \frac{\beta^+}{\beta^-} = 10^3$ representing a moderate and a large jump in the coefficient β.

We present interpolation errors in $I_h u$ in the L^2 norm $||u - I_h u||_0$ and the broken weighted H^1 norms $||u_x - (I_h u)_x||_{0,h}$ and $||u_y - (I_h u)_y||_{0,h}$ and compute their orders of convergence in Tables 1 and 2 for the quadratic IFE space $\mathcal{J}_h$. We conclude that $\mathcal{J}_h$ has optimal approximation capability, $(p + 1 = 3)$ in L^2 norm, and $(p = 2)$ in broken weighted H^1 norms, for both moderate and large jumps in β.

Fig. 3 Geometry of Ω and the quadratic interface Γ in Example 3.1

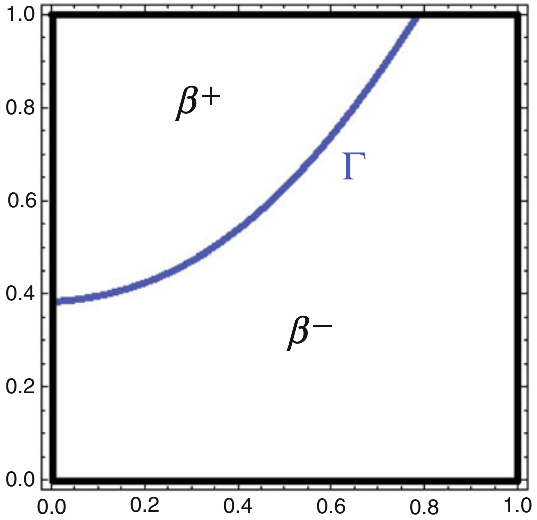


Table 1 Interpolation errors and orders for u, u_x, and u_y for the function (10) in Example 3.1 with $r = 5$, with the IFE space $\mathcal{J}_h$

h	$\|\|u - I_h u\|\|_0$	Order	$\|\|u_x - (I_h u)_x\|\|_{0,h}$	Order	$\|\|u_y - (I_h u)_y\|\|_{0,h}$	Order
1/ 10	2.233135e−03	N/A	1.490673e−01	N/A	4.638299e−02	N/A
1/ 20	2.816030e−04	2.987336	3.752784e−02	1.989931	1.168816e−02	1.988547
1/ 30	8.356524e−05	2.996240	1.670054e−02	1.996824	5.201587e−03	1.996752
1/ 40	3.527386e−05	2.998050	9.398318e−03	1.998422	2.927384e−03	1.998229
1/ 50	1.806504e−05	2.998803	6.016243e−03	1.999017	1.873946e−03	1.998996
1/ 60	1.045593e−05	2.999149	4.178425e−03	1.999372	1.301490e−03	1.999414
1/ 70	6.585241e−06	2.999262	3.070096e−03	1.999508	9.562708e−04	1.999498
1/ 80	4.411826e−06	2.999612	2.350624e−03	1.999738	7.321765e−04	1.999676

Table 2 Interpolation errors and orders for u, u_x, and u_y for the function (10) in Example 3.1 with $r = 1{,}000$, with the IFE space $\mathcal{J}_h$

h	$\|\|u - I_h u\|\|_0$	Order	$\|\|u_x - (I_h u)_x\|\|_{0,h}$	Order	$\|\|u_y - (I_h u)_y\|\|_{0,h}$	Order
1/ 10	2.231791e−03	N/A	1.489825e−01	N/A	4.631764e−02	N/A
1/ 20	2.814414e−04	2.987295	3.750711e−02	1.989907	1.167267e−02	1.988428
1/ 30	8.351753e−05	2.996232	1.669135e−02	1.996817	5.194732e−03	1.996731
1/ 40	3.525368e−05	2.998055	9.393162e−03	1.998418	2.923542e−03	1.998211
1/ 50	1.805472e−05	2.998799	6.012948e−03	1.999013	1.871490e−03	1.998987
1/ 60	1.044995e−05	2.999154	4.176138e−03	1.999370	1.299785e−03	1.999412
1/ 70	6.581474e−06	2.999260	3.068417e−03	1.999505	9.550187e−04	1.999493
1/ 80	4.409302e−06	2.999612	2.349338e−03	1.999741	7.312180e−04	1.999674

4 Application to Interface Problems

To solve the interface problem, we use the interior penalty immersed finite element (IP-IFE) formulation detailed in [5]. We present numerical results for the IP-IFE method with the quadratic IFE space $\mathcal{J}_h$.

For all our numerical experiments, we consider the rectangular domain $\Omega = [0,1]^2$ and the uniform triangular mesh $\mathscr{T}_h = \mathscr{T}_h^i \cup \mathscr{T}_h^c$ of size h. Here, $\mathscr{T}_h$ is formed by partitioning Ω into $(1/h)^2$ squares, with $h = \frac{1}{2^m}$, $m = 2, \ldots, 7$, then forming the triangular elements by joining the lower right and upper left vertices of the squares.

Example 4.1. We consider the quadratic interface $\Gamma : y = x^2 + \frac{\pi}{7}$, defined in Example 3.1. We solve the interface problem (1a), where the true solution is given by (10) from the same example, with $r = \frac{\beta+}{\beta-} = 5$, then with $r = \frac{\beta+}{\beta-} = 10^3$. The L^2 error $||u - U^h||_0$, the weighted errors $||u_x - U_x^h||_{0,h}$ and $||u_y - U_y^h||_{0,h}$, and their orders of convergence are presented in Tables 3 and 4.

Moreover, a computation of the global rates of convergence using least-squares fit reveals the following:

Table 3 L^2 errors and orders for u, u_x, and u_y in Example 4.1 with $r = 5$, using the interior penalty IFE method with the IFE space $\mathcal{J}_h$

h	$\|\|u - U^h\|\|_0$	Order	$\|\|u_x - U_x^h\|\|_{0,h}$	Order	$\|\|u_y - U_y^h\|\|_{0,h}$	Order
1/ 10	2.239261e−03	N/A	1.494187e−01	N/A	4.761907e−02	N/A
1/ 20	2.820069e−04	2.989221	3.755771e−02	1.992180	1.178206e−02	2.014948
1/ 30	8.376481e−05	2.993891	1.672086e−02	1.995786	5.240740e−03	1.997990
1/ 40	3.532076e−05	3.001724	9.403894e−03	2.000588	2.939983e−03	2.009367
1/ 50	1.808355e−05	3.000167	6.019455e−03	1.999283	1.880580e−03	2.002403
1/ 60	1.046425e−05	3.000404	4.179944e−03	2.000306	1.305182e−03	2.003263
1/ 70	6.591189e−06	2.998564	3.071384e−03	1.999145	9.585988e−04	2.002100
1/ 80	4.414898e−06	3.001159	2.351411e−03	2.000375	7.338535e−04	2.000752

Table 4 L^2 errors and orders for u, u_x, and u_y in Example 4.1 with $r = 1{,}000$, using the interior penalty IFE method with the IFE space $\mathcal{J}_h$

h	$\|\|u - U^h\|\|_0$	Order	$\|\|u_x - U_x^h\|\|_{0,h}$	Order	$\|\|u_y - U_y^h\|\|_{0,h}$	Order
1/ 10	3.337899e−03	N/A	1.715437e−01	N/A	9.378451e−02	N/A
1/ 20	2.890901e−04	3.529349	3.782933e−02	2.180999	1.305892e−02	2.844314
1/ 30	8.802201e−05	2.932808	1.688684e−02	1.989197	6.020309e−03	1.909740
1/ 40	3.568347e−05	3.138532	9.416284e−03	2.030347	3.053520e−03	2.359701
1/ 50	1.827893e−05	2.997795	6.040323e−03	1.989675	1.963467e−03	1.978920
1/ 60	1.053416e−05	3.022823	4.184462e−03	2.013363	1.343519e−03	2.081046
1/ 70	6.629628e−06	3.004037	3.073598e−03	2.001478	9.843718e−04	2.017791
1/ 80	4.506741e−06	2.890515	2.364600e−03	1.963882	7.797157e−04	1.745464

$$||u - I_h u||_0 \approx Ch^{2.9968}, \quad ||ux - (I_h u)_x||_{0,h} \approx Ch^{1.9975},$$
$$||uy - (I_h u)_y||_{0,h} \approx Ch^{2.0053}, \quad \text{for } r = 5.$$
$$||u - I_h u||_0 \approx Ch^{3.1251}, \quad ||ux - (I_h u)_x||_{0,h} \approx Ch^{2.0415},$$
$$||uy - (I_h u)_y||_{0,h} \approx Ch^{2.2276}, \quad \text{for } r = 10^3.$$

From these results, we can easily observe that the interior penalty IFE method with the IFE space $\mathcal{J}_h$ performs optimally for interface problems with quadratic interfaces.

5 Conclusion

In summary, we developed a quadratic IFE space that has an optimal approximation capability of $p + 1 = 3$ in L^2 norm and $p = 2$ in the weighted H^1 norms. The quadratic IFE space performs optimally in producing solutions to interface problems with quadratic interfaces via the interior penalty IFE method in [5]. In a forthcoming paper, an extension of the method to handle arbitrary smooth interfaces will be proposed.

Acknowledgements This research was partially supported by the National Science Foundation (Grant Number DMS 1016313).

References

1. Gong, Y., Li B., Li, Z.: Immersed-interface finite-element methods for elliptic interface problems with non-homogeneous jump conditions. SIAM J. Numer. Anal. **46**, 472–495 (2008)
2. Babuska, I., Osborn, J.E.: Can a finite element method perform arbitrarily badly? Math. Comput. **69**(230), 443–462 (2000)
3. Bramble, J.H., King J.T.: A finite element method for interface problems in domains with smooth boundary and interfaces. Adv. Comput. Math. **6**, 109–138 (1996)
4. Chen, Z., Zou, J.: Finite element methods and their convergence for elliptic and parabolic interface problems. Numer. Math. **79**, 175–202 (1998)
5. Adjerid, S., Ben-Romdhane, M., Lin, T.: Higher degree immersed finite element methods for second-order elliptic interface problems. Int. J. Numer. Anal. Model. **11**(3), 541–566 (2014)
6. Barrett, J.W., Elliott, C.M.: Fitted and unfitted finite-element methods for elliptic equations with smooth interfaces. IMA J. Numer. Anal. **7**, 283–300 (1987)

Calendering Analysis of a Non-Newtonian Material

A.M. Siddiqui, M. Zahid, M.A. Rana, and T. Haroon

Abstract In this investigation, the study of a non-Newtonian material when it is dragged through the narrow region between two corotating rolls is carried out. Theoretical analysis based on the lubrication approximation theory (LAT) shows that LAT is a good predictive tool for calendering, where the sheet thickness is very small compared with the roll size. By considering the influence of the material parameter, the dimensionless leave-off distance in the calendering process is determined. The leave-off distance is expressed in terms of eigenvalue problem. Quantities of engineering interest like the maximum pressure, the roll-separating force, the power transmitted to the fluid by rolls, and the normal stress effect are calculated. It is observed that the material parameter has great influence on detachment point, velocity, and pressure distribution, which are useful for the calendering process.

Keywords Third-order fluid • Calendering • Sheet thickness • Lubrication theory

A.M. Siddiqui
Department of Mathematics, Pennsylvania State University, York Campus 1031, Edgecomb Avenue, York, PA 17403, USA

M. Zahid (✉)
Department of Basic Sciences, Riphah International University, Sector I-14, Islamabad, Pakistan

Department of Mathematics, COMSATS Institute of Information Technology, University Road 22060, Abbottabad, Pakistan
e-mail: zahid315@hotmail.com

M.A. Rana
Department of Basic Sciences, Riphah International University, Sector I-14, Islamabad, Pakistan

T. Haroon
Department of Mathematics, COMSATS Institute of Information Technology,
Islamabad, Pakistan

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_17,

1 Introduction

Calendering is one of the final processes in production line. It takes place by conveying the material through the contact line, the nip, between two rotating parallel cylindrical rollers. Due to the line load in the nip, the material undergoes a deformation that smoothens the surface and reduces its thickness. The technological challenge of calendering is to achieve the required surface smoothness and other surface properties, together with the proper shape of the material. The calendering process of forming a flowable material into film or sheet is used in a variety of industries, such as paper, plastics and rubber, leather cloth, shrink films for packaging, resilient flooring tiles, and so on. This process for shaping of materials into sheets and films was introduced in the 1830s by Edwin Chaffee and Charles Goodyear [1] in the United States; however, Gaskell [2] was the first to analyze the process by developing a one-dimensional mathematical procedure for Newtonian and Bingham plastics fluids. Following Gaskell's work, a great deal of effort was made by numerous researchers to improve the model [2]. All works prior to 1990 have been summarized in the textbook by Agassant et al. [3]. Sofou and Mitsoulis [4] used the lubrication theory to provide numerical results for isothermal viscoplastic calendering sheets with a desired final thickness. Arcos et al. [5] reported the influence of the temperature-dependent consistency index on the exiting sheet thickness in the calendering process of a power-law fluid. Hernandez et al. [6] studied theoretically the analysis of calendering for an incompressible Newtonian fluid flow, with pressure-dependent viscosity. Their analysis is based on the regular perturbation technique and the resulting governing equations are based on the well-known lubrication theory. Recently Siddiqui et al. [7] presented the effects of magnetohydrodynamics on calendering process of an incompressible Newtonian fluid. They successfully found that the magnetic field provides the controlling parameter to increase or decrease power transmission, separation force, and distance between attachment and detachment points. Most recently Siddiqui et al. [8] consider thermodynamically compatible model for third-grade fluid [9] to provide numerical results for calendering process.

Various constitutive models currently exist to describe the properties of non-Newtonian fluids. The major problem however is that none of these models can adequately describe all non-Newtonian fluids. Among the several constitutive equations that have been suggested in the literature is a Rivlin-Ericksen model of third-order fluid that is capable of describing the normal stress effects for steady unidirectional flow and predicting shear thinning/thickening effects [10, 11].

From the aforementioned works, it appears that the no work has been reported on third-order calendering. The objective of this paper is to derive the flow mechanism of such a material and to investigate the effects of fluid physical properties on calendering operation. The paper is organized as follows. In the following sections the governing equations and formulation of the problem are given. The later section deals with the analytical solution of the flow variables. Finally results and discussions and conclusions are provided.

2 Governing Equations

The fundamental equations governing the flow of an incompressible, isothermal fluid are the field equations

$$\text{div}\mathbf{V} = 0, \tag{1}$$

$$\rho\frac{D\mathbf{V}}{Dt} = -\nabla p + \text{divS}, \tag{2}$$

where V is the fluid velocity, ρ is the density of the fluid, $\frac{D}{Dt}$ is the material time derivative, and S is the extra stress tensor which for a third-order fluid satisfies the constitutive equation

$$\mathbf{S} = \mu\mathbf{A}_1 + \alpha_1\mathbf{A}_2 + \alpha_2\mathbf{A}_1^2 + \beta_1\mathbf{A}_3 + \beta_2(\mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_1) + \beta_3(tr\mathbf{A}_1^2)\mathbf{A}_1, \tag{3}$$

$$\frac{D(.)}{Dt} = \frac{\partial(.)}{\partial t} + (\mathbf{V}\cdot\nabla)(.), \tag{4}$$

$$\mathbf{A}_1 = (\nabla\mathbf{V}) + (\nabla\mathbf{V})^T, \tag{5}$$

$$\mathbf{A}_n = \left(\frac{\partial}{\partial t} + (\nabla\cdot\mathbf{V})\right)\mathbf{A}_{n-1} + \mathbf{A}_{n-1}(\nabla\mathbf{V}) + (\nabla\mathbf{V})^T\mathbf{A}_{n-1}, \tag{6}$$

μ denotes the dynamic viscosity, $\alpha_i \quad (i = 1, 2)$, $\beta_i \quad (i = 1, 3)$ are the material constants, T indicates the matrix transpose, and A_i $(i = 1, 3)$ are the first three Rivlin-Ericksen tensors.

Equation (3) reduces to second-order fluid when β_1, β_2, and β_3 are zero and reduces to classical Newtonian fluid model when all material moduli except μ are zero.

3 Problem Formulation

Consider an incompressible, laminar, steady third-order fluid, which is dragged through the narrow region between the two corotating cylinders of the same radii R in such a way as to produce a sheet. The x-axis is taken parallel to the sheet and the y-axis normal to it. The upper roller is rotating anticlockwise while the lower roller is rotating clockwise with the same angular velocity ω, resulting in a linear velocity at its surface given by $U = \omega R$, their separation at the nip is H_0, and $-x_f$ is the location where material first bites the rolls, which is known, as shown in Fig. 1. The length of the curved channel formed by the rolls is very large compared with the separation at the nip, i.e., $H_0 \ll R$; hence, the flow can be taken two dimensional.

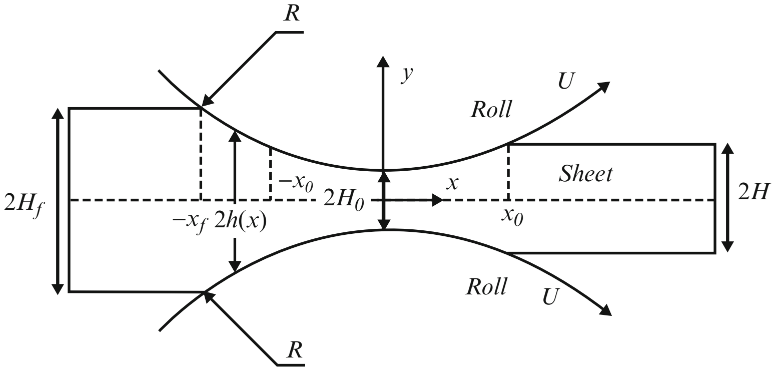


Fig. 1 Geometry of the studied physical model

$$V = [u(x, y), v(x, y)]. \tag{7}$$

Due to the symmetry of the physical model, we consider the upper half of this configuration.

We begin with the lubrication approximation theory (LAT) that the most important dynamic events occur in the nip region. In that region and extending to either side by a distance of the order of x_0, the roll surfaces are nearly parallel. Then it is reasonable to assume that $v \ll u$ and $\frac{\partial}{\partial x} \ll \frac{\partial}{\partial y}$. The material moves in the x-direction and there is no velocity in the y-direction. Thus, Eq. (7) implies $\partial u/\partial x = 0$, which means $u = u(y)$. Then, the continuity Eq. (1) is satisfied identically, the material derivative DV/Dt vanishes and the momentum Eq. (2) reduces to $-\nabla p + \text{div}\mathbf{S} = 0$. This leads Eq. (2) in component form as

$$\frac{dS_{xy}}{dy} - \frac{\partial p}{\partial x} = 0, \tag{8}$$

$$\frac{dS_{yy}}{dy} - \frac{\partial p}{\partial y} = 0, \tag{9}$$

where

$$S_{xy} = \frac{du}{dy} + 2(\beta_2 + \beta_3)\left(\frac{du}{dy}\right)^3 \quad \text{and} \quad S_{yy} = (2\alpha_1 + \alpha_2)\left(\frac{du}{dy}\right)^2. \tag{10}$$

On introducing generalized pressure P

$$P(x, y) = p(x, y) - (2\alpha_1 + \alpha_2)\left(\frac{du}{dy}\right)^2. \tag{11}$$

Using Eqs. (10) and (11), Eqs. (8) and (9) take the form

$$\mu \frac{d^2u}{dy^2} + 2(\beta_2 + \beta_3) \frac{d}{dy} \left(\frac{du}{dy} \right)^3 = \frac{\partial P}{\partial x}, \tag{12}$$

$$\frac{\partial P}{\partial y} = 0. \tag{13}$$

Eq. (13) that P can be a function of x alone. Therefore Eq. (12) can be written as

$$\mu \frac{d^2u}{dy^2} + 2\beta_4 \frac{d}{dy} \left(\frac{du}{dy} \right)^3 = \frac{dP}{dx}, \tag{14}$$

where for simplicity we have introduced $\beta_4 = \beta_2 + \beta_3$.

If both the rolls are identical and rotating with the speed U, then appropriate boundary conditions are

$$\begin{cases} u = U \text{ on } y = h(x), \\ \frac{\partial u}{\partial y} = 0 \text{ on } y = 0, \end{cases} \tag{15}$$

where $h(x)$ is the y-distance from the center line to the roll surface, that is,

$$h(x) = H_0 + R - \left(R^2 - x^2\right)^{\frac{1}{2}}. \tag{16}$$

Confining the analysis to values of x such that $x \ll R$, a good approximation to $h(x)$ is

$$h(x) = H_0 \left(1 + \frac{x^2}{2H_0R} \right). \tag{17}$$

4 Dimensionless Equations

In this section, dimensionless governing equations are presented to solve the third-order calendering process. Based on LAT analysis carried out previously, consider the following dimensionless variables:

$$x^* = \frac{x}{\sqrt{2RH_0}}, \quad u^* = \frac{u}{U}, \quad y^* = \frac{y}{H_0}, \quad P^* = \sqrt{\frac{H_0}{2R}} \frac{PH_0}{\mu U}, \quad \beta = \frac{2\beta_4 U^2}{\mu H_0},$$
$$h^*(x^*) = \frac{h(x)}{H_0}. \tag{18}$$

Introducing the above dimensionless variables, Eqs. (14) and (15) after removing the “*” sign yield

$$\frac{d^2u}{dy^2}+2\beta\frac{d}{dy}\left(\frac{du}{dy}\right)^3=\frac{dP}{dx}, \tag{19}$$

$$\begin{cases} u=1 \quad \text{on} \quad y=h(x)=1+x^2, \\ \frac{\partial u}{\partial y}=0 \text{ on } \quad y=0. \end{cases} \tag{20}$$

In case of finite sheet, there will be zero pressure and pressure gradient at exit as well as zero pressure at entry where the sheet first bites the rolls; the dimensionless forms of these boundary conditions, required for the solution of Eq. (19), are

$$\begin{cases} \frac{dP}{dx}_{x=\lambda}=P(x=\lambda)=0, \\ P(x=-x_f)=0. \end{cases} \tag{21}$$

Along with Eq. (19), the dimensionless volumetric flow rate is required, which can be written in the form

$$Q=1+\lambda^2=\int_0^{1+x^2}udy, \tag{22}$$

here in Eq. (22), λ represents an unknown eigenvalue of mathematical problem, which is related to the existing sheet thickness in the calendering process by the relationship defined in the next subsection.

4.1 Sheet Thickness

Once λ is found, then all other engineering quantities of interest are immediately available. The exiting sheet thickness H is given by

$$\frac{H}{H_0}=1+\lambda^2. \tag{23}$$

The thickness of the entering sheet H_f entering the analysis according to the definition is

$$x_f=\sqrt{\frac{H_f}{H_0}-1}. \tag{24}$$

5 Solution for $\beta \ll 1$

In order to find the dimensionless velocity and pressure profiles, and leave-off distance of the sheet, we try to find out an asymptotic solution to Eqs. (19)–(22). Applying the regular perturbation technique and using β as the perturbation parameter,

$$u(x,\ y) = u_0(x,\ y) + \beta u_1(x,\ y) + \cdots, \tag{25}$$

$$P(x) = P_0(x) + \beta P_1(x) + \cdots, \tag{26}$$

$$Q = Q_0 + \beta Q_1 + \cdots, \tag{27}$$

$$\lambda = \lambda_0 + \beta\lambda_1 + \cdots, \tag{28}$$

where u_0, P_0, Q_0, and λ_0 are the leading-order solutions, which represent the Newtonian case [2, 8], and u_1, P_1, Q_1, and λ_1 are the corrections up to first-order terms and contain the contribution of the non-Newtonian effect.

5.1 Zeroth-Order Problem and Its Solution

By introducing the relationships Eqs. (25)–(28) into Eqs. (19)–(22) and collecting terms of the same power of β, the zeroth-order boundary value problem becomes

$$\frac{dP_0}{dx} = \frac{d^2u_0}{dy^2}, \quad \text{for} \quad -x_f \leq x \leq \lambda_0, \tag{29}$$

$$Q_0 = 1 + \lambda_0^2 = \int_0^{1+x^2} u_0 dy \tag{30}$$

$$\begin{cases} \frac{du_0}{dy} = 0 \quad \text{at} \quad y = 0, \\ u_0 = 1 \quad \text{at} \quad y = 1 + x^2, \\ \frac{dP_0}{dx} = P_0 = 0 \text{ at } x = \lambda_0. \end{cases} \tag{31}$$

The solution for Eqs. (29) and (30) subject to boundary conditions (31) is given by

$$u_0 = 1 - \frac{1}{2}\left(\frac{dP_0}{dx}\right)\left[(1 + x^2)^2 - y^2\right], \tag{32}$$

with

$$\frac{dP_0}{dx} = -3\frac{(\lambda_0^2 - x^2)}{(1 + x^2)^3} \quad \text{for} \quad -x_f \leq x \leq \lambda_0, \tag{33}$$

and

$$P_0(x) = \frac{3}{8}\left[\begin{array}{l} \frac{x^2(1-3\lambda_0^2)-1-5\lambda_0^2}{(1+x^2)^2}x + (1-3\lambda_0^2)\left(\tan^{-1}x - \tan^{-1}x_f\right) \\ + \frac{\left(1+x_f^2\right)(1-3\lambda_0^2)-2(1+\lambda_0^2)}{\left(1+x_f^2\right)^2}x_f \end{array}\right]. \tag{34}$$

From Eq. (34), if we assume that $P_0 \to 0$ as $x \to -\infty$, we get $\lambda_0 = 0.4751$ (keeping $x_f = -3.0$). For this value of λ_0 we have, the maximum sheet thickness to minimum gap width ratio equal to 1.226.

The zeroth-order velocity profile is obtained by substituting Eq. (33) into Eq. (32):

$$u_0(x,y) = 1 - \frac{3}{2}\frac{\left(x^2-\lambda_0^2\right)}{(1+x^2)^3}\left[\left(1+x^2\right)^2 - y^2\right]. \tag{35}$$

Here we must emphasize that the solutions given by Eqs. (32)–(35) were obtained in previous works [2, 5, 8].

5.2 First-Order Problem and Its Solution

The first-order boundary value problem takes the form

$$\frac{dP_1}{dx} = \frac{d^2u_1}{dy^2} + \frac{d}{dy}\left(\frac{du_0}{dy}\right)^3, \tag{36}$$

$$Q_1 = 2\lambda_0\lambda_1 = \int_0^{1+x^2} u_1 dy, \tag{37}$$

$$\begin{cases} \frac{du_1}{dy} = 0 \quad \text{at} \quad y = 0, \\ u_1 = 0 \quad \text{at} \quad y = 1 + x^2, \\ \frac{dP_1}{dx} = P_1 = 0 \text{ at } x = \lambda_1. \end{cases} \tag{38}$$

Using Eq. (35) in Eq. (36) and integrating the resulting equation twice and using the boundary conditions (38), we get the following first-order solution for velocity:

$$u_1(x,y) = \frac{27}{4}\frac{\left(\lambda_0^2 - x^2\right)^3}{\left(1+x^2\right)^9}\left[y^4 - \left(1+x^2\right)^4\right] + \frac{1}{2}\frac{dP_1}{dx}\left[y^2 - \left(1+x^2\right)^2\right]. \tag{39}$$

To find the first-order dimensionless flow rate, Q_1, we substitute Eq. (39) into Eq. (37) and get

$$Q_1 = \frac{27}{5}\frac{\left(x^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^4} - \frac{1}{3}\frac{dP_1}{dx}\left(1+x^2\right)^3, \tag{40}$$

here Q_1 involves unknown pressure gradient dP_1/dx. Flow rate Q_1 and dP_1/dx can be obtained by applying the boundary condition given in Eq. (38). Thus, an explicit expression for dP_1/dx is

$$\frac{dP_1}{dx} = -\frac{81}{5\left(1+x^2\right)^3}\left[\frac{\left(\lambda_1^2-\lambda_0^2\right)^3}{\left(1+\lambda_1^2\right)^4} - \frac{\left(x^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^4}\right]. \tag{41}$$

Equation (41) is valid for $-x_f \leq x \leq \lambda_1$, where λ_1 must be determined as a part of the problem. The pressure distribution is obtained by integrating Eq. (41). Hence,

$$P_1(x) = \frac{-81}{5}\int_{-x_f}^{x}\frac{1}{\left(1+x^2\right)^3}\left[\frac{\left(\lambda_1^2-\lambda_0^2\right)^3}{\left(1+\lambda_1^2\right)^4} - \frac{\left(x^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^7}\right]dx. \tag{42}$$

The dimensionless leave-off distance λ_1 may be found from the above equation, since it has been assumed that $P_1 = 0$ at $x = -x_f$. Therefore Eq. (42) can be written as

$$\int_{-x_f}^{\lambda_1}\left[\frac{\left(x^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^7} - \frac{\left(\lambda_1^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^3\left(1+\lambda_1^2\right)^4}\right]dx = 0. \tag{43}$$

By integrating Eq. (42) and applying the boundary conditions (38), we have the following form of pressure distribution:

$$\begin{aligned}
P_1(x) = {} & 0.0129\tan^{-1}(x) - \left(6.0750\lambda_1^6 - 4.1137\lambda_1^4 + 0.9285\lambda_1^2 - 0.0699\right)\tfrac{\tan^{-1}(x)}{\left(1+\lambda_1^2\right)^4} \\
& + \left[\tfrac{-2.4860}{\left(1+x^2\right)^6} + \tfrac{4.5669}{\left(1+x^2\right)^5} - \tfrac{2.3084}{\left(1+x^2\right)^4} + \tfrac{0.0069}{\left(1+x^2\right)^3} + \tfrac{0.0086}{\left(1+x^2\right)^2} + \tfrac{0.0129}{\left(1+x^2\right)}\right]x \\
& - \left[\begin{array}{l}\left(\tfrac{4.0500}{\left(1+x^2\right)^2} + \tfrac{6.0750}{\left(1+x^2\right)}\right)\lambda_1^6 - \left(\tfrac{2.7424}{\left(1+x^2\right)^2} + \tfrac{4.1137}{\left(1+x^2\right)}\right)\lambda_1^4 \\ + \left(\tfrac{0.6190}{\left(1+x^2\right)^2} + \tfrac{0.9285}{\left(1+x^2\right)}\right)\lambda_1^2 - \left(\tfrac{0.0466}{\left(1+x^2\right)^2} + \tfrac{0.0699}{\left(1+x^2\right)}\right)\end{array}\right]\tfrac{x}{(1+\lambda_1)^4} \\
& - 0.0129\tan^{-1}\left(x_f\right) \\
& + \left(6.0750\lambda_1^6 - 4.1137\lambda_1^4 + 0.9285\lambda_1^2 - 0.0699\right)\tfrac{\tan^{-1}\left(x_f\right)}{\left(1+\lambda_1^2\right)^4} \\
& - \left[\tfrac{-2.4860}{\left(1+x_f^2\right)^6} + \tfrac{4.5669}{\left(1+x_f^2\right)^5} - \tfrac{2.3084}{\left(1+x_f^2\right)^4} + \tfrac{0.0069}{\left(1+x_f^2\right)^3} + \tfrac{0.0086}{\left(1+x_f^2\right)^2} + \tfrac{0.0129}{\left(1+x_f^2\right)}\right]x_f \\
& + \left[\begin{array}{l}\left(\tfrac{4.0500}{\left(1+x_f^2\right)^2} + \tfrac{6.0750}{\left(1+x_f^2\right)}\right)\lambda_1^6 - \left(\tfrac{2.7424}{\left(1+x_f^2\right)^2} + \tfrac{4.1137}{\left(1+x_f^2\right)}\right)\lambda_1^4 \\ + \left(\tfrac{0.6190}{\left(1+x_f^2\right)^2} + \tfrac{0.9285}{\left(1+x_f^2\right)}\right)\lambda_1^2 - \left(\tfrac{0.0466}{\left(1+x_f^2\right)^2} + \tfrac{0.0699}{\left(1+x_f^2\right)}\right)\end{array}\right]\tfrac{\left(x_f\right)}{(1+\lambda_1)^4}
\end{aligned} \tag{44}$$

In Eq. (44) assuming $P_1 \to 0$ as $x \to -\infty$ and $x_f = -3.0$, we get $\lambda_1 = 0.3336$.

The first-order velocity profile is obtained by substituting Eq. (41) into Eq. (39):

$$u_1(x, y) = \frac{27}{4}\frac{\left(\lambda_0^2 - x^2\right)^3}{\left(1 + x^2\right)^9}\left[y^4 - \left(1 + x^2\right)^4\right] + \frac{81}{10}\left[\frac{\left(x^2 - \lambda_0^2\right)^3}{\left(1 + x^2\right)^7} - \frac{\left(\lambda_1^2 - \lambda_0^2\right)^3}{\left(1 + x^2\right)^3\left(1 + \lambda_1^2\right)^4}\right]$$

$$\left[y^2 - \left(1 + x^2\right)^2\right]. \tag{45}$$

Combining the solutions at each order of approximation yields the solutions up to first order for velocity, pressure gradient, and pressure distribution.

5.3 *Operating Variables*

Once the velocity, pressure gradient, and pressure distribution are found, then all other quantities of interest are readily available. The operating variables of engineering interest are computed in the following manners.

5.3.1 Roll-Separating Force

The roll-separating force F is defined as

$$F = \int_{-\infty}^{\lambda} P(x)dx, \tag{46}$$

where $F = \frac{\bar{F} H_0}{\mu U R W}$, $\bar{F}$ is the dimensional roll-separating force per unit width W.

5.3.2 Power Input

The power transmitted to the fluid by roll is calculated by integrating the product of shear stress and the roll surface speed over the surface of roll which is obtained by setting $\bar{y} = 1$ in Eq. (25):

$$P_w = \int_{-\infty}^{\lambda} S_{xy}(x, 1)dx, \tag{47}$$

here $P_w = \frac{\bar{P}_w}{\mu W U^2}$ is the dimensionless power and $S_{xy} = \frac{\bar{S}_{xy} H_0}{\mu U}$ is the dimensionless stress tensor defined by

$$S_{xy} = \frac{\partial u}{\partial y} + \beta\left(\frac{\partial u}{\partial y}\right)^3, \tag{48}$$

where

$$\frac{\partial u}{\partial y}=3\frac{\left(x^2-\lambda_0^2\right)}{\left(1+x^2\right)^3}y+\beta\left(\frac{81}{5}\left(\frac{\left(x^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^7}-\frac{\left(\lambda_1^2-\lambda_0^2\right)^3}{\left(1+x^2\right)^3\left(1+\lambda_1^2\right)^4}\right)y+27\frac{\left(\lambda_0^2-x^2\right)^3}{\left(1+x^2\right)^9}y^3\right). \tag{49}$$

The expression for power can be calculated from Eq. (47) with the help of Eq. (48) when $y = 1$.

5.3.3 Normal Stress Effect

Equation (11) in a dimensionless form can be written as

$$p\left(x,y\right)=P\left(x,y\right)+\alpha\left(\frac{du}{dy}\right)^2, \tag{50}$$

where $\alpha = \frac{(2\alpha_1+\alpha_2)U}{\mu H_0}\sqrt{\frac{H_0}{2R}}$.

By using Eq. (25) and Eq. (26) in the above equation, one can easily find normal stress.

5.3.4 Adiabatic Temperature

The power input causes to raise the temperature of the fluid by an amount which at most is given by an adiabatic temperature rise $(\Delta T)_{\text{ave}}$:

$$(\Delta T)_{\text{ave}}=\frac{P_w}{Q\rho C_p}, \tag{51}$$

where ρ is the melt density and C_p is the melt heat capacity at constant pressure.

6 Results and Discussions

In this paper, we analyze the calendering process for incompressible third-order fluid. The lubrication theory is used to simplify the equations of motion. The numerical results for the volumetric flow rate, the separation point λ, the exit sheet thickness H/H_0, the power input, and the roll-separating force are presented in Tables 1 and 2. Table 1 is generated for positive values of β, whereas Table 2 is reserved for negative values of β. For a physical point of view, an increase in third-order parameter β corresponds to shear thickening effect i.e., an increase in viscosity of the fluid. More viscous fluid diffuses more momentum. Consequently, the magnitude of velocity decreases. This fact is obvious from Figs. (3–10). It is noted from Table 1 that sheet thickness, power input, and roll-separating force increase with an increase in β. This was physically expected because of shear thickening effect.

Table 1 The effect of material parameter β on leave-off distance, final sheet thickness, power input, and roll-separating force

β	λ	H/H_0	P_w	F
0.01	0.4784	1.2289	0.0062	0.5527
0.03	0.4851	1.2353	0.0114	0.5542
0.05	0.4917	1.2418	0.0174	0.5557
0.07	0.4984	1.2484	0.0216	0.5572
0.09	0.5051	1.2551	0.0243	0.5587
0.1	0.5084	1.2585	0.0253	0.5594
0.3	0.5751	1.3308	0.0747	0.5744
0.5	0.6419	1.4120	0.1267	0.5901
0.7	0.7086	1.5021	0.1761	0.6068
0.9	0.7753	1.6011	0.2210	0.6248
1.0	0.8087	1.6539	0.2422	0.6343

Table 2 The effect of material parameter β on leave-off distance, final sheet thickness, power input, and roll-separating force

β	λ	H/H_0	P_w	F
−0.01	0.4712	1.2225	0.0054	0.5507
−0.03	0.4650	1.2163	0.0017	0.5492
−0.05	0.4584	1.2101	−0.0017	0.5477
−0.07	0.4517	1.2040	−0.0048	0.5462
−0.09	0.4450	1.1980	−0.0047	0.5446
−0.1	0.4417	1.1951	−0.0094	0.5439
−0.3	0.3750	1.1406	−0.0229	0.5286
−0.5	0.3083	1.0950	0.0046	0.5133
−0.7	0.2415	1.0583	0.0940	0.4980
−0.9	0.1748	1.0305	0.2598	0.4827
1.0	0.1415	1.0200	0.3655	0.4751

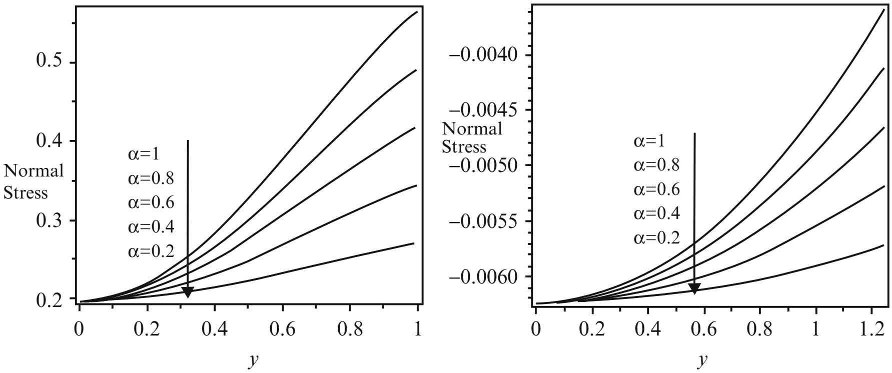


Fig. 2 Normal stress effect at position $x = 0$ and at $x = 0.5$, respectively, fixing $\beta = 0.5$

Figure 2 shows the normal stress effects at different positions of calendering process keeping $\beta = 0.5$ for different values of α. It is observed that the stress in increasing through out the region with the increase in α.

Fig. 3 Effect of β on velocity at $x = -0.5$

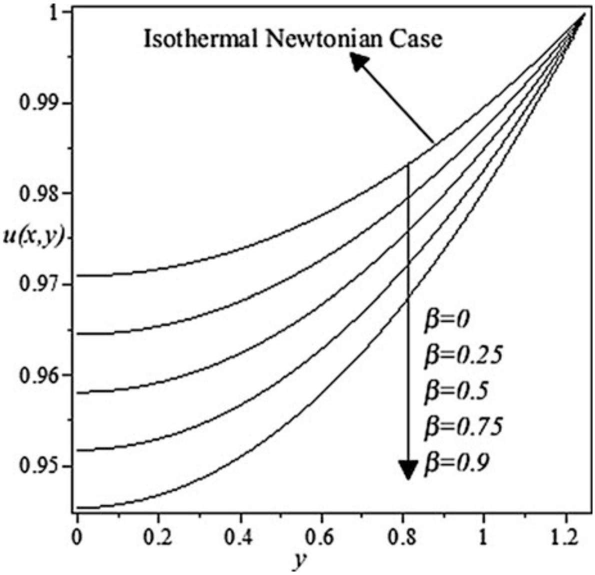
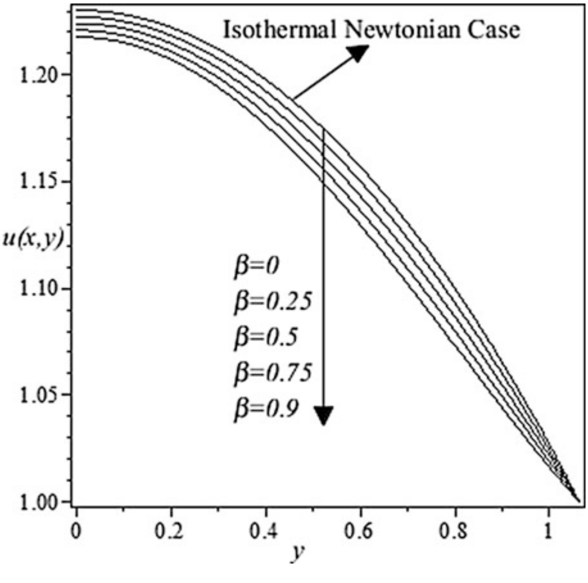


Fig. 4 Effect of β on velocity at $x = -0.25$



In Figs. (3–10) the dimensionless velocity $u(x, y)$ is presented as a function of the transversal coordinate y for different values of β. The dimensionless velocity profiles are evaluated at eight different positions of x. In Figs. 3 and 4 the velocity profile is plotted in the domain $-x_f \le x \le -\lambda$, corresponding to the region near the entrance where the pressure gradient is positive. We can see that velocity decreases with the material parameter for a fixed value of λ, and this increment is

Fig. 5 Effect of β on velocity at $x = 0$

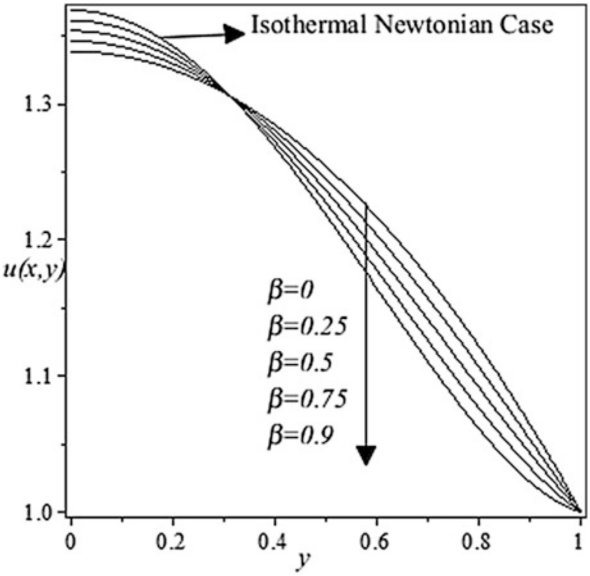
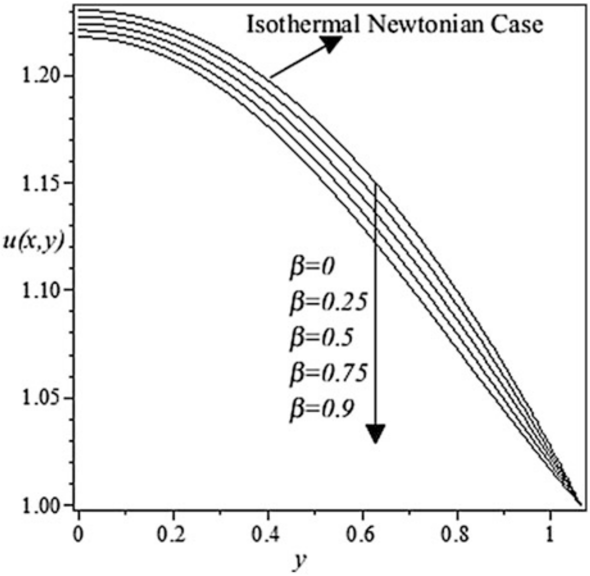


Fig. 6 Effect of β on velocity at $x = 0.25$



more pronounced at the center plane. In contrast, at the vicinity of the rolls, velocity diminishes weakly compared with the Newtonian case. The velocity profile was also plotted in the domain $-\lambda \leq x \leq \lambda$, velocity increases weakly at the vicinity of the rolls, and it decreases at the center plane when β is increased as shown in Figs. (5–9).

Fig. 7 Effect of β on velocity at $x = 0.4$

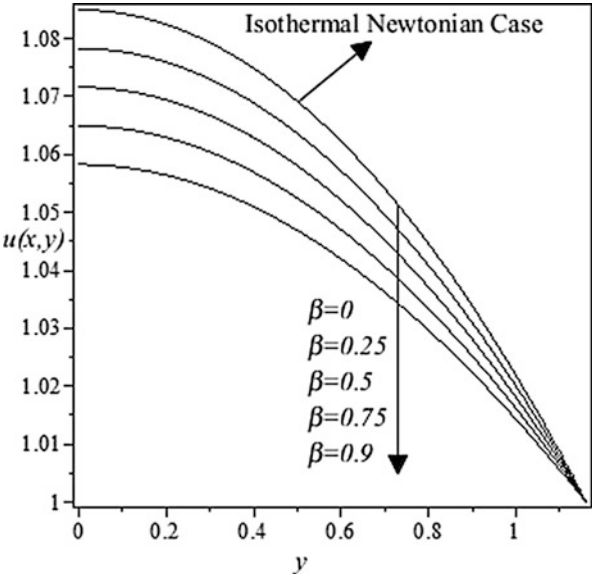


Fig. 8 Effect of β on velocity at $x = 0.5$

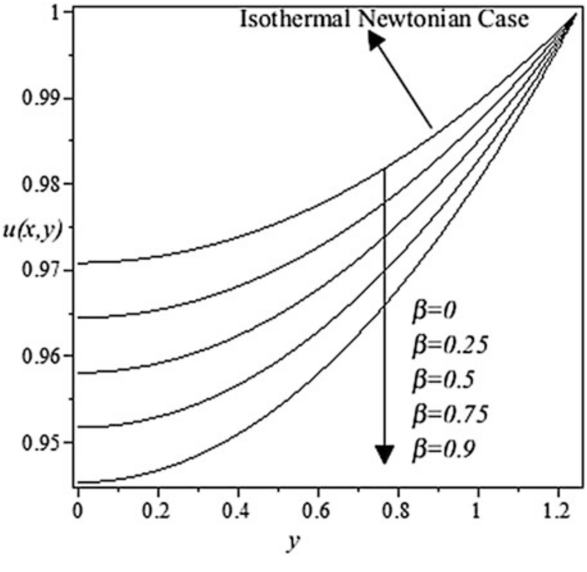


Figure 11 shows the numerical solution for the dimensionless pressure gradient as a function of the dimensionless axial coordinate x, for different values of the material parameter $\beta(=0.0, 0.25, 0.5, 0.75, 0.9)$, and two fixed values of the dimensionless leave-off distance $\lambda(=0.4751, 0.3336)$. Symmetric profiles about the nip point are obtained. Pressure gradient is negative at $x = 0$ and increases symmetrically about this point, attains maximum value, and then decreases exponentially and reaches

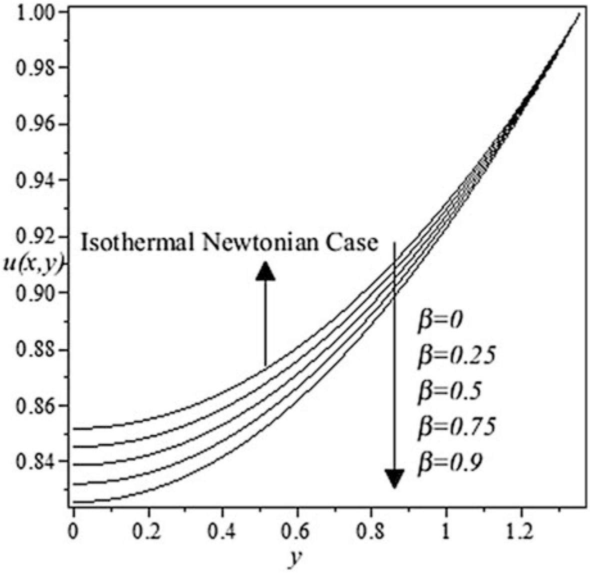


Fig. 9 Effect of β on velocity at $x = 0.6$

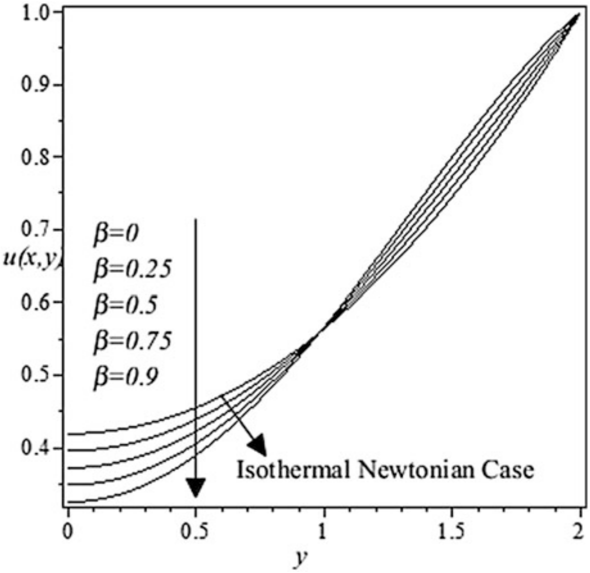


Fig. 10 Effect of β on velocity at $x = 1.0$

to zero value at $x = \pm 4$. Moreover, an increase in β causes to increase pressure gradient. Also, material parameter has major effect on pressure gradient at $x = 0$ because the pressure has a maximum absolute value at this point. In contrast, it has zero effect near $x = \pm\lambda$ because at this point flat velocity profiles are reached corresponding to a rigid motion of the sheet.

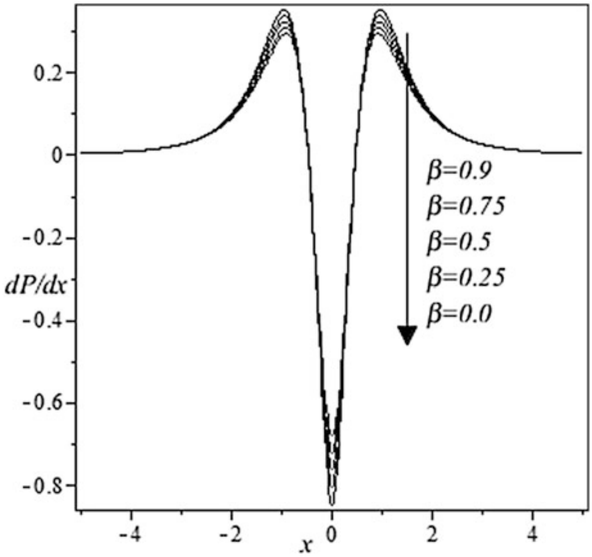


Fig. 11 Effect of β on pressure gradient

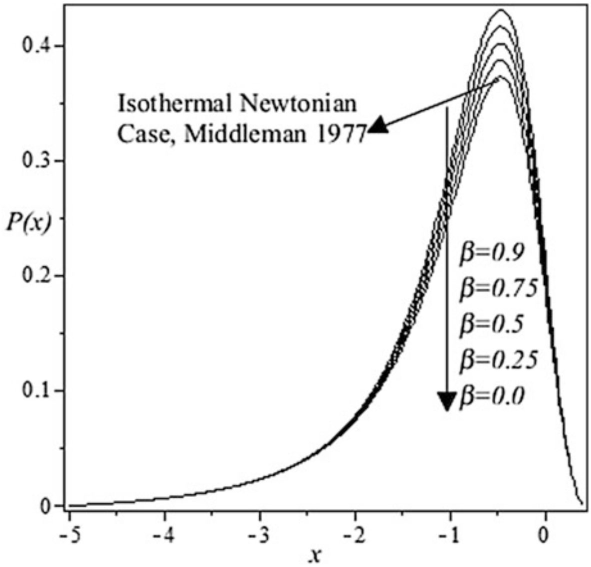


Fig. 12 Effect of β on pressure distribution

The results for the dimensionless pressure distribution are shown in Fig. 12. Starting from zero value at $x = \lambda$, the pressure increases monotonically up to a maximum value at $x = -\lambda$ and then decreases until the entry point for a finite sheet thickness is reached. By varying the material parameter β, the pressure is drastically affected at $x = -\lambda$. Also, an increase in the value of β tends to extend the length of contact between the rolls and the material as can be noted from Tables 1 and 2.

7 Conclusion

The findings of the present work can be summarized as follows:

1. Shear thickening/thinning phenomena are observed.
2. Shear thickening causes to enhance sheet thickness, power input, and roll-separating force.
3. Third-order parameter β provides mechanism to control sheet thickness, power input, roll-separating force, and leave-off distance.
4. Sheet thickness, power input, roll-separating force, and leave-off distance for non-Newtonian material are larger than those of Newtonian fluid.

References

1. Osswald, T.A., Menges, G.: Historical background. In: Material Science of Polymers for Engineers, 2nd edn, p.18. Hanser Publishers, Munch (2003)
2. Gaskell, R.E.: The calendering of plastic materials. J. Appl. Mech. **17**, 334–337 (1950)
3. Agassant, J.-F., Avenas, P., Sergent, J.-Ph., Carreau, P.J.: Polymer Processing: Principles and Modeling. Hanser Publishers, Munich (1991)
4. Sofou, S., Mitsoulis, E.: Calendering of pseudoplastic and viscoplastic sheets of finite thickness. J. Plast. Film Sheeting **20**, 185–222 (2004)
5. Arcos, J.C., Méndez, F., Bautista, O.: Effect of temperature-dependent consistency index on the exiting sheet thickness in the calendering of power-law fluids. Int. J. Heat Mass Tran. **54**, 3979–3986 (2011)
6. Hernandez, A., Arcos, J.C., Méndez, F., Bautista, O.L.: Effect of pressure-dependent consistency index on the exiting sheet thickness in the calendering of Newtonian fluids. Appl. Math. Model. (2013) http://dx.doi.org/10.1016/j.apm.
7. Siddiqui, A.M., Zahid, M., Rana, M.A., Haroon, T.: Effect of magnetohydrodynamics on Newtonian calendering. J. Plast. Film Sheeting **29**(4), 347–364 (2013)
8. Siddiqui, A.M., Zahid, M., Rana, M.A., Haroon, T.: Calendering analysis of a third-order fluid. J. Plast. Film Sheeting **0**(0), 1–24 (2013)
9. Fosdick, R.L., Rajagopal, K.R.: Thermodynamics and stability of fluids of third grade. Proc. R. Soc. **A339**, 351–377 (1980)
10. Coleman, B.D.: On the stability of equilibrium states of general fluids. Archs Ration. Mech. Anal. **36**, 1–32 (I970)
11. Coleman, B.D., Noll, W.: An approximation theorem for functionals, with application in continuum mechanics. Archs Ration. Mech. Anal. **6**, 355–370 (1960)

Wavelet Solution of Convection-Diffusion Equation with Neumann Boundary Conditions

A.H. Choudhury

Abstract In this paper, we derive a highly accurate numerical method for the solution of one-dimensional convection-diffusion equation with Neumann boundary conditions. This parabolic problem is solved by using semidiscrete approximations. The space direction is discretized by wavelet-Galerkin method and the time variable is discretized by using various classical finite difference schemes. The numerical results show that this method gives high favourable accuracy compared with the exact solution.

Keywords Parabolic equation • Semidiscrete approximations • Stability • Wavelet-Galerkin method

1 Introduction

In this paper, we consider numerical solution of one-dimensional convection-diffusion equation

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} + \beta \frac{\partial u}{\partial x} = f, \quad a < x < b, \quad t > 0 \tag{1}$$

with initial condition

$$u(x,0) = g(x), \quad a < x < b \tag{2}$$

A.H. Choudhury (✉)
Department of Mathematics, Srikishan Sarda College, Hailakandi 788151, India
e-mail: ahchoudhury27@yahoo.com

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_18,

and boundary conditions

$$\frac{\partial u}{\partial x}(a,t) = c(t), \quad \frac{\partial u}{\partial x}(b,t) = d(t), \quad t > 0, \tag{3}$$

where α is a positive constant and β and f are constants or functions of any or of both the independent variables x and t. Parabolic partial differential equation (PDE) like (1) appears in connection with fluid mechanics, financial mathematics and many other fields. Several methods exist for the solution of convection-diffusion equation (1), mostly with Dirichlet and periodic boundary conditions, for example, [1–4]. But its solution with Neumann boundary conditions is hardly available in the literature. A particle method was proposed by Mas-Gallic [5] which is only theoretical and lacks computational aspects.

Usually, parabolic problems are solved by using semidiscrete approximations. For the solution of problem (1)–(3) in the present paper, the space direction is discretized by using wavelet-Galerkin method and the time variable is discretized by using various classical finite difference schemes. Wavelets in consideration here are Daubechies compactly supported wavelets [6] which are differentiable.

Wavelet applications to the solution of PDE problems are relatively new. Some recent applications are [2, 4, 7, 8], among many more. To discretize a PDE problem by wavelet-Galerkin method, the Galerkin bases are constructed from orthonormal bases of compactly supported wavelets. This can be done in a number of ways. In this paper, we construct these basis functions as in Choudhury and Deka [7]. This approach has been used to solve wave equation in Choudhury and Deka [8].

In Sect. 2, we explain the approximation of the Sobolev space $H^m(a,b)$ using Daubechies scaling functions. Section 3 elaborates the method for the solution of problem (1)–(3). In Sect. 4, we demonstrate the method with the help of a numerical example. Section 5 concludes the paper.

2 Approximation of Sobolev Spaces in Daubechies Bases

For a positive integer N, consider two functions $\phi, \psi \in L^2(\mathbb{R})$ defined by

$$\phi(x) = \sum_k a_k \phi(2x-k), \quad \psi(x) = \sum_k b_k \phi(2x-k), \tag{4}$$

where $\{a_k\}_{k\in\mathbb{Z}}$ and $\{b_k\}_{k\in\mathbb{Z}}$ are two specific sequences [6] such that $a_k = b_k = 0$ for $k \notin \{0, 1, \ldots, S\}$, $S = 2N - 1$. The functions ϕ and ψ are called dbN scaling function and dbN wavelet function, respectively, where N is called their order. These functions are compactly supported with $\text{supp}(\phi) = \text{supp}(\psi) = [0, S]$. They are available in wavelet toolbox of MATLAB 6 for $1 \leqslant N \leqslant 45$. They satisfy the properties (3.9)–(3.12) in [7].

The integer translates and dilates of ϕ and ψ are defined as

$$\phi_{n,k}(x) = 2^{\frac{n}{2}}\phi(2^n x - k), \quad \psi_{n,k}(x) = 2^{\frac{n}{2}}\psi(2^n x - k), \quad n, k \in \mathbb{Z}. \tag{5}$$

Now, for all $n \in \mathbb{Z}$, we define

$$V_n = L^2\text{-closure } (\text{span}\{\phi_{n,k} : k \in \mathbb{Z}\}). \tag{6}$$

We recall here that for an open interval (a, b) and for an integer $m \geqslant 1$, the space

$$H^m(a, b) = \{u \in H^{m-1}(a, b) : u' \in H^{m-1}(a, b)\} \tag{7}$$

is called the Sobolev space of order m, which is a Hilbert space with inner product $\langle u, v\rangle_m = \sum_{i=0}^{m} \int_a^b u^{(i)} v^{(i)} dx$ and associated norm $\|.\|_m$. It may be noted here that $H^0(a, b) = L^2(a, b)$.

Let N be any positive integer and let ϕ and ψ be the dbN scaling function and wavelet function, respectively. Then, by Theorem 1.1 in [9], there exists an integer m, $0 \leqslant m < N$, such that the Sobolev space $H^m(a, b)$ can be approximated by the restrictions of translates and dilates of ϕ to (a, b).

We shift the support of ϕ from $[0, S]$ to $[a, b]$ by using the transformation $y = \frac{b-a}{S}x + a$ and let

$$I_n = \{k \in \mathbb{Z} : \text{supp}(\phi_{n,k}) \cap (a, b) \neq \emptyset\}. \tag{8}$$

Considering V_n as defined in (6), we define the space $V_n(a, b)$ to be the set of restrictions of all functions in V_n to (a, b). In fact, we take

$$V_n(a, b) = \text{span}\{\phi_{n,k}|_{(a,b)} : k \in I_n\}. \tag{9}$$

Since (a, b) is bounded, the space $V_n(a, b)$ is finite dimensional and is a closed subspace of $H^m(a, b)$. By Proposition 4.1 in [7], $\dim(V_n(a, b)) = 2^n S + S - 1$ and a basis for $V_n(a, b)$ is given by

$$\{\phi_{n,k} \in V_n(a, b) : 1 - S \leqslant k \leqslant 2^n S - 1\}. \tag{10}$$

3 Solution Methodology

Since PDE (1) is of second order in space with Neumann boundary conditions (3), the solution space for spatial direction for problem (1)–(3) is $H^1(a, b)$. Multiplying equation (1) by a function $v \in H^1(a, b)$ and integrating by parts with respect to x in (a, b), we get

$$\int_a^b \left(\frac{\partial u}{\partial t}v + \alpha \frac{\partial u}{\partial x}\frac{dv}{dx} + \beta \frac{\partial u}{\partial x}v\right) dx = \int_a^b f v dx + \alpha[d(t)v(b) - c(t)v(a)], \quad (11)$$

which is the variational (weak) form of problem (1)–(3).

In Glowinski et al. [9], it is established that $N \geqslant 3$ is sufficient for the solution of problems of second order (in space). So, we let $N \geqslant 3$ be any integer and let ϕ be the dbN scaling function. Considering the basis $\{\phi_{n,j}\}$ of $V_n(a,b)$ in Sect. 2, the approximate solution of the variational problem (11) can be taken as

$$u_n(x,t) = \sum_j z_{n,j}(t)\phi_{n,j}(x). \quad (12)$$

Applying the Galerkin method to problem (11) with the approximate solution (12), we get a system of first-order ordinary differential equations in $z = [z_{n,j}]$:

$$M\dot{z} + Az = F, \quad (13)$$

where A, M and F are the stiffness matrix, the mass matrix and the force vector, respectively, whose elements are given by

$$\begin{cases} A_{ij} = \int_a^b \left(\alpha\phi'_{n,j}\phi'_{n,i} + \beta\phi'_{n,j}\phi_{n,i}\right) dx, \; M_{ij} = \int_a^b \phi_{n,j}\phi_{n,i}\, dx, \\ F_i = \int_a^b f\phi_{n,i}\, dx + \alpha[d(t)\phi_{n,i}(b) - c(t)\phi_{n,i}(a)]. \end{cases} \quad (14)$$

There are several methods for the solution of equation (13). The most commonly used method is a θ*-family of approximation*:

$$\theta\dot{z}_{s+1} + (1-\theta)\dot{z}_s = \frac{z_{s+1} - z_s}{\Delta t}, \quad 0 \leqslant \theta \leqslant 1, \quad (15)$$

where z_s refers to the value of z at time $t = t_s = s\Delta t$.

Using approximation (15) for times t_s and t_{s+1} in (13), we get

$$\hat{M}_{s+1} z_{s+1} = \hat{A}_s z_s + \hat{F}_{s,s+1}, \quad (16)$$

where

$$\begin{cases} \hat{M}_{s+1} = M + \theta\Delta t A_{s+1}, \\ \hat{A}_s = M - (1-\theta)\Delta t A_s, \\ \hat{F}_{s,s+1} = \Delta t[\theta F_{s+1} + (1-\theta)F_s]. \end{cases} \quad (17)$$

The solution z_{s+1} at time t_{s+1} is obtained in terms of the solution z_s at time t_s by inverting the matrix $\hat{M}_{s+1}$, where the initial solution z_0 at time $t = t_0 = 0$ is $z(0)$, which can be obtained by multiplying the initial condition (2) by v and integrating and approximating with (12). For details about this scheme, we refer to Reddy [10].

The above θ-family of time approximation schemes, for which $\theta < \frac{1}{2}$, is stable only if

$$\Delta t < \frac{2}{(1-2\theta)\lambda}, \tag{18}$$

where λ is the largest eigenvalue of the problem associated to the original PDE problem.

For different values of θ, the time discretization scheme (15) is classified as follows:

1. *The forward difference (or Euler) scheme*: $\theta = 0$; conditionally stable; order of accuracy=O(Δt).
2. *The backward difference scheme*: $\theta = 1$; unconditionally stable; order of accuracy=O(Δt).
3. *The Galerkin scheme*: $\theta = \frac{2}{3}$; unconditionally stable; order of accuracy= O($(\Delta t)^2$).
4. *The Crank-Nicholson scheme*: $\theta=\frac{1}{2}$; unconditionally stable; order of accuracy=O($(\Delta t)^2$).

4 Numerical Results

Here the methodology for the solution of the parabolic IBVP (1)–(3) described above is demonstrated with an example. The solution is performed using all the four time discretization schemes. The computations are performed in MATLAB 6.

For the test problem, we take $\alpha = \frac{1}{16}$ and $\beta = \frac{1}{4}$ in the convection-diffusion equation (1) with space domain [0, 1]. The right-hand function $f(x)$, the initial value function $g(x)$ and the boundary value functions $c(t)$ and $d(t)$ in (2) and (3) are obtained as per the exact solution

$$u(x,t) = 2e^{-\frac{\pi^2}{4}t}\sin(2\pi x) + 8x^2 t. \tag{19}$$

For $\theta = 0$, $db3$ scaling function is used for spatial discretization. As this scheme is conditionally stable, we have to find an upper limit of the time step Δt using the stability condition (18). The largest eigenvalue of the associated problem for $n = 0$ is 320.93. Therefore, the maximum time step is $\frac{2}{320.93} \approx 0.0062$. Figure 1 shows the exact, unstable and stable solutions due to $db3(n = 0)$ scaling function at $x = 1$.

For $\theta = 1$ and $\theta = \frac{2}{3}$, we also use $db3$ scaling function for spatial discretization. As this scheme is unconditionally stable, there is no restriction on Δt. Table 1 shows the decay of maximum absolute error for both the schemes with decreasing time step due to $db3(n = 1, 2, 3)$ scaling functions at $t = 1$.

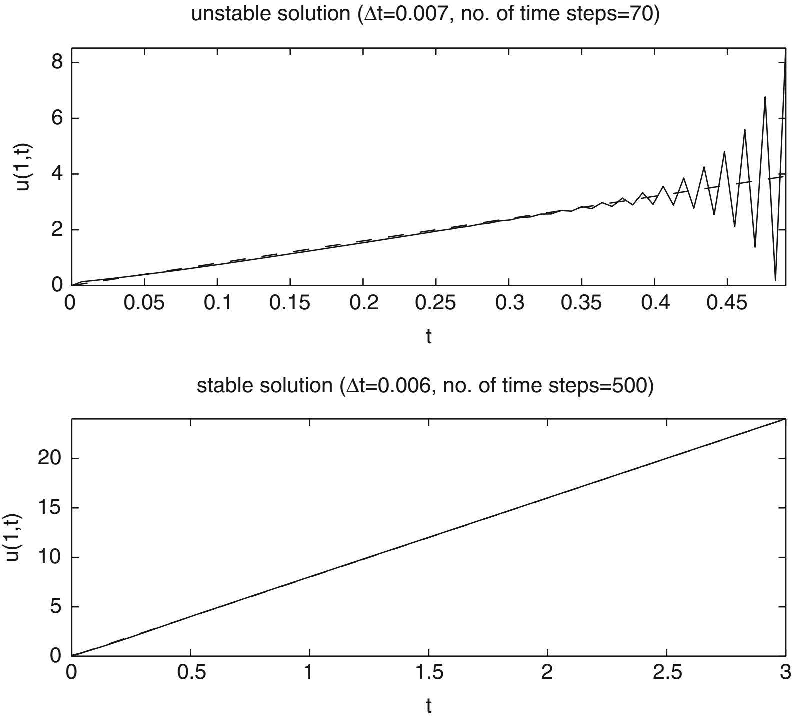


Fig. 1 ($\theta = 0$): $db3(n = 0)$ solution (*solid line*) and exact solution (*dashed line*) at $x = 1$

Table 1 $\left(\theta = 1, \frac{2}{3}\right)$ Maximum absolute errors at $t = 1$ due to $db3$ scaling functions

Scheme	Time step	Maximum absolute error		
(θ)	(Δt)	$n = 1$	$n = 2$	$n = 3$
1	$\frac{1}{5}$	0.2276	0.2262	0.2261
	$\frac{1}{10}$	0.1212	0.1181	0.1180
	$\frac{1}{20}$	0.0645	0.0605	0.0602
	$\frac{1}{40}$	0.0351	0.0307	0.0305
	$\frac{1}{80}$	0.0201	0.0157	0.0153
$\frac{2}{3}$	$\frac{1}{5}$	0.0661	0.0621	0.0619
	$\frac{1}{10}$	0.0406	0.0362	0.0360
	$\frac{1}{20}$	0.0240	0.0196	0.0193
	$\frac{1}{40}$	0.0148	0.0104	0.0100
	$\frac{1}{80}$	0.0099	0.0055	0.0051

For $\theta = \frac{1}{2}$, Table 2 shows the maximum absolute errors between the exact and the computed solutions at $t = 1$ due to $db3(n = 0, 1, 2, 3)$ and $db4(n = 0, 1, 2, 3)$ scaling functions for $\Delta t = 0.01$ and $\Delta t = 0.001$, respectively.

Table 2 $(\theta = \frac{1}{2})$ Maximum absolute errors at $t = 1$ due to $db3$ and $db4$ scaling functions

Scaling functions	n	Δt	Maximum absolute error
$db3$	0	10^{-2}	5.2205×10^{-2}
	1	10^{-2}	7.3238×10^{-3}
	2	10^{-2}	1.0011×10^{-3}
	3	10^{-2}	1.3354×10^{-4}
$db4$	0	10^{-3}	7.2287×10^{-3}
	1	10^{-3}	3.4199×10^{-4}
	2	10^{-3}	1.7572×10^{-5}
	3	10^{-3}	1.4965×10^{-6}

5 Conclusion

In this paper, we have analysed a method for numerical solution of one-dimensional convection-diffusion equation with Neumann boundary conditions. The space direction is discretized by using wavelet-Galerkin method and the time variable is discretized by using classical finite difference schemes. The main advantages of this method are that the schemes are unconditionally stable (except one) and are useful for problems with time-dependent boundary conditions and with time-dependent source term. The method gives high favourable accuracy. The efficiency of the developed algorithm has been illustrated by a test problem.

References

1. Bazán, F.S.V.: Chebyshev pseudospectral method for computing numerical solution of convection-diffusion equation. Appl. Math. Comput. **200**, 537–546 (2008)
2. Choudhury, A.H., Deka, R.K.: Wavelet method for numerical solution of convection-diffusion equation. In: Nadarajan, R., Vijayalakshmi Pai, G.A., Sai Sundara Krishnan, G. (eds.) Mathematical and Computational Models: Recent Trends, pp. 101–107. Narosa Publishing House, New Delhi (2010)
3. Dehghan, M.: On the numerical solution of the one-dimensional convection-diffusion equation. Math. Prob. Eng. **1**, 61–74 (2005)
4. Mehra, M., Kumar, B.V. Rathish: Time-accurate solution of advection-diffusion problems by wavelet-Taylor-Galerkin method. Commun. Numer. Meth. Eng. **21**, 313–326 (2005)
5. Mas-Gallic, S.: Particle approximation of a linear convection-diffusion problem with Neumann boundary conditions. SIAM J. Numer. Anal. **32**, 1098–1125 (1995)
6. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Commun. Pure Appl. Math. **41**, 909–996 (1988)
7. Choudhury, A.H., Deka, R.K.: Wavelet-Galerkin solutions of one dimensional elliptic problems. Appl. Math. Model. **34**, 1939–1951 (2010)
8. Choudhury, A.H., Deka, R.K.: Wavelet method for numerical solution of wave equation with neumann boundary conditions. In: Proceedings of International MultiConference of Engineers and Computer Scientists, vol. II, pp. 1513–1516. Hong Kong (2011)

9. Glowinski, R., et al.: Wavelet solutions of linear and non-linear elliptic, parabolic and hyperbolic problems in one space dimension. In: Glowinski, R., Lichnewsky, A. (eds.) Computing Methods in Applied Sciences and Engineering, pp. 55–120. SIAM, Philadelphia (1990)
10. Reddy, J.N.: An Introduction to the Finite Element Method. Tata McGraw Hill, New Delhi (2003)

Test of Causality Between Oil Price and GDP Growth in Algeria

Zouaoui Chikr Elmezouar, A. Mazri, M. Benzaire, and AEK. Boudi

Abstract This paper seeks to investigate the causal relationship between oil prices and economic growth in Algeria. The empirical analysis starts by analyzing the time series properties of the data which is followed by examining the nature of causality among the variables. Algeria is an oil-producing rather oil-exporting country. An increase in oil price increases economic growth. This study analyzes how change in real crude oil price affects the real GDP of Algeria positively. The empirical analysis involves testing the time series characteristics of the data series (stationary) using ADF test and running the pairwise Granger causality test based on EViews software.

Keywords Oil price • Economic growth • Cointegration • Granger causality • Algeria

1 Introduction

Algeria is the second largest oil producer, after Nigeria, in Africa. It became a member of the Organization of the Petroleum Exporting Countries (OPEC) in 1969, shortly after it began oil production in 1958. Currently, the country is heavily reliant on its hydrocarbon sector, which accounted for almost 70 percent of government budget revenue and grants and about 98 percent of export earnings in 2011, according to the International Monetary Fund.

In recent years, crude oil production has been stagnant, because new production and infrastructure projects have repeatedly been delayed. Additionally, in the last

Z. Chikr Elmezouar (✉) • A. Mazri • M. Benzaire • AEK. Boudi
Laboratory of Economic Studies and Local Development in South West
of Algeria, University of Bechar, Bechar, Algeria
e-mail: chikrtime@yahoo.fr; mazeriabdelhafid@yahoo.fr; benzairmebarek@yahoo.fr; dr.boudi@yahoo.fr

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_19,

three licensing rounds, there has been limited interest from investors to undertake new oil projects under the government's current terms. As a result, the Algerian Parliament recently approved amendments to the current.

Any major disruption to Algeria's hydrocarbon production would not only be detrimental to the local economy but, depending on the scale of lost production, could affect world oil prices.

Moreover, several studies draw additional conclusions. Thus, the effect of oil shocks is asymmetric. Indeed, rising oil prices have a larger impact that cuts on economic growth (and to a lesser extent inflation). This finding may be explained by downward rigidities of wages and prices. Moreover, they allocate effects on the labor market and uncertainty in financial markets as a result of fluctuations in oil prices. An obvious conclusion is that the impact of dearer oil is generally more pronounced in developing countries than in advanced countries. The oil has indeed a more important place in those countries mainly because of the weight of manufacturing and generally less modern machinery. This leads us to ask :

What is the impact of a continuous increase of oil prices on economic growth in Algeria? This raises a number of important questions: Are oil prices a stimulus for economic growth? (Or, alternatively, do oil prices "cause" GDP?) Is economic growth a stimulus for oil prices? (Or, alternatively, does GDP "cause" oil prices?).

The remaining part of the paper is organized in the following way. Section 2 dwells on literature review. Section 3 presents the econometric methodology, Sect. 4 contains empirical results and discussion, and finally, conclusions are drawn in Sect. 5.

2 Literature Review

A number of empirical studies have explored the relationship between economic growth and oil price. Hamilton (1983) showed a negative relationship between oil prices and macroeconomic activity in the United States. Hooker (1994) confirmed Hamilton's results and demonstrated that from 1948 to 1972, oil price variability exerts influence on GDP growth. Later, Mork (1989), Lee et al. (1995), and Hamilton (1996) introduced nonlinear transformations into the models and Granger causality tests. Results confirmed the incidence of negative relationship between oil price fluctuations and economic downturns as well as Granger causality from oil prices to growth before 1973 but no Granger causality from 1973 to 1994. Other studies include Mork (1989), Federer (1996), Hamilton (1997), Lee and Ni (2002), and Balke et al. (2002).

Recently, Gounder and Barleet (2007) using both linear and nonlinear oil price transformations discovered a direct link between net oil price shock and economic growth in New Zealand. In addition, oil price shock was discovered to have substantial effect on inflation and exchange rate. In a comparative study of the impact of oil price shock and exchange rate volatility on economic growth, Jin (2008) discovered that oil price increases exert a negative impact on economic growth in Japan and China and a positive impact on economic growth of Russia.

3 Methodology

3.1 Granger Causality Tests

Several studies have been devoted to the study of causality between variables (Granger 1969; Sims 1972). Furthermore, we carried out the Granger causality test where Granger (1969) proposed a time series data-based approach in order to determine causality. For example, if we want to explore the causal relationship between oil prices (pt) and economic growth (yt),

$$p_t = \underset{i=1\overset{n}{\sum}_i}{\alpha}\ y_{t-i} + \sum_{i=1}^{n}\beta_i p_{t-i} + \epsilon_{1t} \tag{1}$$

$$y_t = \sum_{i=1}^{n}\lambda_i p_{t-i} + \sum_{i=1}^{n}\delta_i y_{t-i} + \epsilon_{2t} \tag{2}$$

where n is the number of lags.

If α_i coefficients are jointly significantly different from zero, the Granger test suggests that economic growth (y_t) causes the oil prices (p_t) and if (λ_i) is jointly significantly different from zero, the Granger test suggests that the oil prices (p_t) cause the economic growth (y_t).

If the two causalities are verified, we can conclude the return causality "feedback causality" between the two variables.

3.2 Causality Test and Cointegration Variables

The relationship causality between different time series is based on the following.

3.2.1 Unit Root Tests

A stochastic process is stationary if its first and second moments are constant.

Analytically, y_t is stationary if

$$E\,(y_t) = \mu, \forall t \tag{3}$$

$$cov\,(y_t, y_{t+k}) = \delta\,(h)\,, \forall t \tag{4}$$

Dickey and Fuller (DF) proposed a basic model of a unit root test

$$\begin{aligned}\Delta y_t &= (\phi - 1)y_{t-1} + \varepsilon_t\\ \Delta y_t &= (\phi - 1)y_{t-1} + \beta + \varepsilon_t\\ \Delta y_t &= (\phi - 1)y_{t-1} + \beta + \partial t + \varepsilon_t\end{aligned}$$

The hypothesis tests are

$$\begin{cases} H_0 : (\phi - 1) = 0\\ H_1 : (\phi - 1) < 0\end{cases}$$

To get a broader view, Dickey and Fuller took an autoregressive process of higher order known as the augmented Dickey-Fuller (ADF) test. This test is represented as follows:

$$\begin{aligned}\Delta y_t &= \phi y_{t-1} + \sum_i \theta_i \Delta y_{t-i} + \varepsilon_t\\ \Delta y_t &= \phi y_{t-1} + \sum_i \theta_i \Delta y_{t-i} + \beta + \varepsilon_t\\ \Delta y_t &= \phi y_{t-1} + \sum_i \theta_i \Delta y_{t-i} + \beta + \partial t + \varepsilon_t\end{aligned}$$

3.3 Cointegration

The main objective of this paper is to assess not only the pairwise nature of causality among the variables but also the short-run and long-run dynamic impacts, which we tested for cointegration using two well-known approaches: the one developed by Engle and Granger (1987) and the other one by Johansen (1988).

3.4 Engel-Granger Method

The Engle-Granger test is a procedure that involves an OLS estimation of a prespecified cointegrating regression between the variables. This was followed by a unit root test performed on the regression residuals previously identified. We applied the Engle Granger to find the number of cointegration equations between the two variables.

3.5 An Error Correction Model

For interpret the vectoreuor correction model found in the different regression equations. Indeed an error correction model (ECM) can detect the dynamics of short-term and long-term variable around its stationary equilibrium value. Thus, for an adjustment, error correction requires that the sign of the coefficient of residual is negative and statistically significant.

The model error corrections read

$$\Delta x_t = \alpha_1 z_{t-1} + \text{lagged}\,(\Delta x_t, \Delta y_t) + \varepsilon_{1t} \tag{5}$$

$$\Delta y_t = \alpha_2 z_{t-1} + \text{lagged}\,(\Delta x_t, \Delta y_t) + \varepsilon_{2t} \tag{6}$$

with z_{t-1} the error correction term resulting from estimating the cointegration relationship and ε the error term stationary $|\alpha_1| + |\alpha_2| \neq 0$.

3.6 Causality Test

The causality test based on the model vector correction has the advantage of providing a causal relationship even if no estimated coefficient of lagged variables used is significant.

Thus, an error correction model after processing can be rewritten as

$$\Delta p_t = \alpha + \sum_{i=1}^{k} \lambda_i \Delta p_{t-i} + \sum_{i=1}^{k} \sigma_i \Delta y_{t-i} + \theta z_{t-1} + \varepsilon_t \tag{7}$$

$$\Delta y_t = \beta + \sum_{i=1}^{k} \phi_i \Delta y_{t-i} + \sum_{i=1}^{k} \varphi_i \Delta p_{t-i} + \psi z_{t-1} + \mu_t \tag{8}$$

From both equations, p_t does not cause y_t the sense of Granger if $\varphi_i = \psi = 0,$ *and* y_t does not cause p_t if $\sigma_i = \theta = 0$.

4 Empirical Results and Interpretation

4.1 Statistical Data Properties

The variables that we used in our application are the oil prices (p_t) and the economic growth (y_t). Figure 1 shows the evolution of real GDP and oil price in Algeria from 1982 to 2012. The real GDP is characterized by an upward trend while the price of oil is presented as an additive model. The correlation between the real GDP and oil price is 0.90.

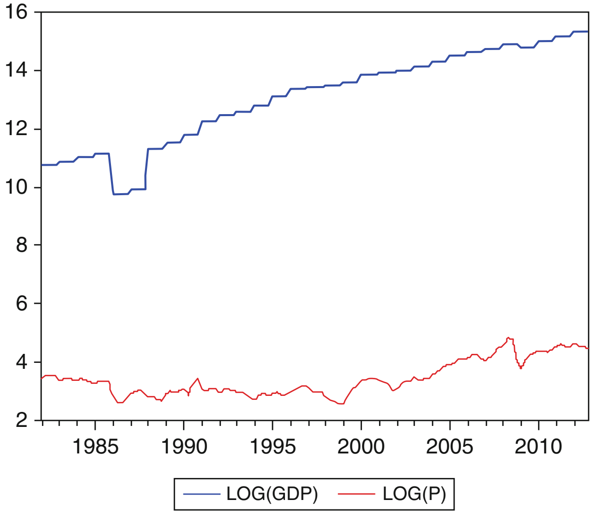


Fig. 1 Representation of log(GDP) and log(p)

Table 1 Test of stationarity for GDP

Null hypothesis (GDP) has a unit root
Exogenous: None
Lag length: 4 (automatic based on SIC MAXLAG= 12)

		t-Statistic	prob*
Augmented Dickey-Fuller test statistic		3.909510	1.0000
Test critical values:	1 % level	−2.584539	
	% level	−1.943540	
	10 % level	−1.614941	

MacKinon (1996) one-sided p-values

4.1.1 Test of Stationarity of GDP and Oil Prices

prob= 1 greater than 0.05, then we can conclude that GDP was nonstationary.
prob= 0.8905 greater than 0.05, then we can conclude that the oil prices were nonstationary.

Tables 1 and 2 present the test results for stationarity of GDP and oil price. The results showed that the two variables were nonstationary.

prob= 0.0298 less than 0.05, then we can conclude that D(GDP) was stationary.
prob= 0.000 less than 0.05, then we can conclude that D(P) was stationary.

Tables 3 and 4 present the test results for stationarity for the difference of GDP and oil price. The results showed that the two variables were stationary.

Table 2 Test of stationarity for oil price

Null hypothesis P has a unit root
Exogenous: None
Lag length: 4 (automatic based on SIC MAXLAG= 12)

		t-Statistic	prob*
Augmented Dickey-Fuller test statistic		0.837473	0.8905
Test critical values:	1 % level	−2.584539	
	5 % level	−1.943540	
	10 % level	−1.614941	

MacKinon (1996) one-sided p-values

Table 3 Test of stationarity for the difference of GDP

Null hypothesis D(GDP) has a unit root
Exogenous: None
Lag length: 3 (automatic based on SIC MAXLAG= 12)

		t-Statistic	prob*
Augmented Dickey-Fuller test statistic		−2.165458	0.0298
Test critical values:	1% level	−2.584539	
	5% level	−1.943540	
	10% level	−1.614941	

MacKinon (1996) one-sided p-values

Table 4 Test of stationarity for the difference of oil price

Null hypothesis D(P) has a unit root
Exogenous: None
Lag length: 3 (automatic based on SIC MAXLAG= 12)

		t-Statistic	prob*
Augmented Dickey-Fuller test statistic		−7.892420	0.0000
Test critical values:	1% level	−2.584539	
	5% level	−1.943540	
	10% level	−1.614941	

MacKinon (1996) one-sided p-values

4.1.2 An Error Correction Model Estimate

Table 5 presents the normalized coefficient of the variables in the model. The coefficient was correctly signed and statistically significant at 1 percent level.

4.1.3 Cointegration Tests

Table 6 presents the test results for the number of cointegrating vectors. The results showed that the trace statistic (39.69792 greater than 15.49471 and 6.580289 greater than 3.841466) suggests the presence of two cointegrations among the two variables.

Table 5 Test for stationarity of the errors

Null hypothesis ECT has a unit root
Exogenous: None
Lag length: 3 (automatic based on SIC MAXLAG= 12)

		t-Statistic	prob*
Augmented Dickey-Fuller test statistic		−2.945771	0.0035
Test critical values:	1% level	−2.584707	
	5% level	−1.943536	
	10% level	−1.614927	

MacKinon (1996) one-sided p-values

Table 6 Test of cointegration

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob**
None*	0.242930	39.69792	15.49471	0.0000
At most*	0.053795	6.580289	3.841466	0.0103

Trace test indicates 2 cointegrating equations at the 0.05 level
*Denotes rejection of the hypothesis at the 0.05 level
MacKinnon-Haug-Michelis (1999) p-values

Table 7 Test of causality

Pairwise Granger causality tests
Logs: 2

Null hypothesis:	Obs	F-Statistic	Prob
GDP does not Granger cause p	122	10.0078	0.0001
P does not Granger cause GDP		7.69307	0.0007

4.1.4 Causality Tests

Table 7 shows that the Prob statistic of the first test= 0.0001 and Prob statistic of the second test= 0.0007 less than 0.05 suggest the presence of two senses of causality among the two variables.

5 Conclusion

This paper employs an empirical analysis to examine the impacts of oil price fluctuations on the level of real economic activity in Algeria. The first step in the empirical analysis involves testing the time series characteristics of the data series using ADF test and running the pairwise Granger causality test. This was followed by applying the Johansen cointegration test and the estimation of the long-run cointegrating vectors and the number of cointegration equations equals two. The analysis was capped with the estimation of short-run error correction model (ECM).

The hypothesis of nonstationarity was rejected at first difference. The Granger pairwise causality test showed that the null hypothesis that oil prices do not Granger cause real GDP could be safely rejected at the 1 percent level. In other words, null hypothesis that GDP does not Granger cause oil price could be safely rejected at the 1 percent level and finally we find that the two variables are causes or there exists feedback causality between the oil price and the GDP.

References

1. Abosedra, S., Baghestani, H.: New evidence on the causal relationship between United States energy consumption and gross national product. J. Energy Dev. **14**(2), 285–292 (1991)
2. Altinay, G., Karagol, E.: Structural break, unit root, and the causality between energy consumption and GDP in Turkey. Energy Econ. **26**, 985–994 (2004)
3. Amaira, B.: The relationship of oil prices and economic growth in Tunisia: A vector error correction model analysis. The Rom. Econ. J. **43**, 3–22 (2012)
4. Engle, R.F., et Granger, C.W.J.: Cointegration and error correction: representation, estimation, and testing. Econometrica **55**(2), 251–76 (1987)
5. Granger, C.W.J.: Investigating causal relations by econometric models and cross spectral methods. Econometrica **36**, 424–438 (1969)
6. Granger, C.W.J. Some recent developments in a concept of causality. J. Economet. **39**, 199–211 (1988)

Solution Behavior of Heston Model Using Impression Matrix Norm

Ahmet Duran and Burhaneddin Izgi

Abstract We are interested in the behavior of solutions for several stochastic differential equations such as Heston model. We focus on the numerical solutions via Milstein method for different stock market conditions. We examine the advantages and limitations of the model. Moreover, we introduce 3-dimensional matrix norms. Furthermore, we define market impression matrix norm as an application to the 3-dimensional matrix norms. Later, we perform simulations for various parameters.

Keywords Numerical solutions of stochastic differential equations • Heston model • 3-dimensional matrix norm • Impression matrix norm

1 Introduction

Our goal in this paper is to study the behavior of solutions for several stochastic differential equations such as Heston model (see [1]). Heston model is a very useful stochastic volatility model in finance market where the evolution for the stock price volatility is described and the volatility is a random process. We focus on the numerical solutions via Milstein method (see [2] and [3]) for different stock market conditions and parameters.

A. Duran (✉) • B. Izgi
Department of Mathematics, Istanbul Technical University, Istanbul 34469, Turkey
e-mail: aduran@itu.edu.tr
web page: http://web.itu.edu.tr/aduran

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_20,

It is important to examine several variables together in order to quantify market impression and summarize large data set for this purpose. Although market price reflects all past publicly available information according to weak-form efficient-market hypothesis (EMH) (see [4]), many traders consider that prices can be overvalued or undervalued. Therefore, we would like to find a methodology for market impression in addition to market price. We believe that market impression may be expressed via several variables such as volatility, interest rate, and time, together.

The remainder of the paper is organized as follows. In sect. 2, we introduce 3-dimensional matrix norms as generalizations of the matrix norms and we prove them by using the applicable numerical linear algebra and analysis arguments (see [5, 6] and references therein). In sect. 3, we define market impression matrix norm as an application to the 3-dimensional matrix norms. Section 4 concludes the paper. Appendix includes Milstein method and Heston model.

2 3-Dimensional Matrix Norms

Matrix norms are essential parts of numerical linear algebra (see [5]) and its applications in science, engineering, and finance.

Definition. A 3-dimensional matrix norm $\| \cdot \|$ is a function from *m-by-n-by-p* complex matrices into $\Re$ that satisfies the following properties:

- $\| A \| \geq 0$ and $\| A \| = 0$ if and only if $A = 0$;
- $\| \alpha A \| = | \alpha | \| A \|$, for scalar α;
- $\| A + B \| \leq \| A \| + \| B \|$; where A and B are matrices in *m-by-n-by-p* dimensional space.

Definition. Let $A \in C^{m \times n \times p}$ then;

$$\| A \|_1 = max_j \sum_{k=1}^{p} \sum_{i=1}^{m} | a_{ij}^{(k)} | = \text{the largest absolute block-column sum.}$$

$$\| A \|_\infty = max_i \sum_{k=1}^{p} \sum_{j=1}^{n} | a_{ij}^{(k)} | = \text{the largest absolute block-row sum.}$$

Proof. Proofs are straightforward and just come from their definition.

Definition. The $p - norm$ of $A \in C^{m \times n \times p}$ is defined as follows:

$$\| A \|_p = \Big(\sum_{k=1}^{p} \sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij}^{(k)} |^p \Big)^{\frac{1}{p}}, \quad \text{for } 1 < p < \infty.$$

Proof. • $\| A \|_p \geq 0$ and $\| A \|_p = 0$ if and only if $A = 0$ (by the definition)

- $\parallel \alpha A \parallel_p = (\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid \alpha a_{ij}^{(k)} \mid^p)^{\frac{1}{p}} = (\mid \alpha \mid^p \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij}^{(k)} \mid^p)^{\frac{1}{p}} = \mid \alpha \mid \parallel A \parallel_p$
- We have to show $\parallel A + B \parallel_p \leq \parallel A \parallel_p + \parallel B \parallel_p$ where $A, B \in C^{m\times n\times p}$.

$$\begin{aligned}
\parallel A + B \parallel_p^p &= \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij}^{(k)} + b_{ij}^{(k)} \mid^p; \text{by the Minkowski inequality}\\
&\leq \sum_{k=1}^{p}\sum_{i=1}^{m}\left((\sum_{j=1}^{n} \mid a_{ij}^{(k)} \mid^p)^{\frac{1}{p}} + (\sum_{j=1}^{n} \mid b_{ij}^{(k)} \mid^p)^{\frac{1}{p}} \right)^p\\
&\leq \sum_{k=1}^{p}\left((\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij}^{(k)} \mid^p)^{\frac{1}{p}} + (\sum_{i=1}^{m}\sum_{j=1}^{n} \mid b_{ij}^{(k)} \mid^p)^{\frac{1}{p}} \right)^p\\
&\leq \left((\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij}^{(k)} \mid^p)^{\frac{1}{p}} + (\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid b_{ij}^{(k)} \mid^p)^{\frac{1}{p}} \right)^p\\
&= \left(\parallel A \parallel_p + \parallel B \parallel_p \right)^p
\end{aligned}$$

The special case (p = 2) of p-norm is the Frobenius norm of $A \in C^{m\times n\times p}$, and it can be defined as follows.

Definition. The Frobenius norm of $A \in C^{m\times n\times p}$ is defined as follows:

$$\parallel A \parallel_F = \sqrt{\sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} \mid a_{ij}^{(k)} \mid^2}.$$

Definition. Let $A \in C^{m\times n\times p}$; then the 2-norm of A is defined as follows:
$\parallel A \parallel_2 = max\big(max_{\parallel x \parallel_2 = 1; k=1,\ldots,p} \parallel A^{(k)}x \parallel_2 \big) = \sqrt{\lambda_{max}^k}$ where λ_{max}^k is the largest eigenvalue of $(A^{(k)})^* A^k$ for all k. Also, it can be defined as $\parallel A \parallel_2^2 = max(\lambda_{max}^k)$ where $\lambda_{max}^k = max_k\big(\mid (A^{(k)})^* A^k - \lambda_k I \mid = 0\big)$; (k = 1, ..., p).

Proof. • $\parallel A \parallel_2 = \sqrt{max(\lambda_{max}^k)} \geq 0$ (since all eigenvalues of $(A^{(k)})^* A^k$ are real and nonnegative) and $\parallel A \parallel_2^2 = 0 \Rightarrow A = 0$.

- $\parallel \alpha A \parallel_2^2 = max(\lambda_{\mathbf{max}}^{\mathbf{k}}) = max\big\{max_k\big(((\alpha A)^{(k)})^* \alpha A^k - \lambda_{\mathbf{k}} I\big)\big\}$ where $\lambda_{\mathbf{k}}$ is the eigenvalue of $(\alpha A)^{(k)})^* \alpha A^k$.

  Also we know from the definition that $\parallel A \parallel_2 = max\big\{max_k\big((A^{(k)})^* A^k - \lambda_k I\big)\big\}$ where λ_k is the eigenvalue of $(A^{(k)})^* A^k$:

$$(\alpha A)^{(k)})^* \alpha A^k x = \lambda_{\mathbf{k}} x$$

$$(A)^{(k)})^* \alpha^* \alpha A^k x = \lambda_{\mathbf{k}} x$$

$$
\begin{aligned}
(A)^{(k)})^{*}\alpha^{2}A^{k}x &= \leftthreetimes_{\mathbf{k}}x \\
\alpha^{2}(A)^{(k)})^{*}A^{k}x &= \leftthreetimes_{\mathbf{k}}x \\
\alpha^{2} \leftthreetimes_{k} x &= \leftthreetimes_{\mathbf{k}}x \\
\big(\alpha^{2} \leftthreetimes_{k} - \leftthreetimes_{\mathbf{k}}\big)x &= 0; \qquad (x \neq 0\, vector) \\
\leftthreetimes_{\mathbf{k}} &= \alpha^{2}\leftthreetimes_{k}
\end{aligned}
$$

Finally, $\parallel \alpha A \parallel_2^2 = max(\leftthreetimes^{\mathbf{k}}_{\mathbf{max}}) = max(\alpha^2 \leftthreetimes^{k}_{max}) = \alpha^2 max(\leftthreetimes^{k}_{max}) = \alpha^2 \parallel A \parallel_2^2 \Rightarrow \parallel \alpha A \parallel_2 = \mid \alpha \mid \parallel A \parallel_2$.

- We have to show $\parallel A + B \parallel_2 \leq \parallel A \parallel_2 + \parallel B \parallel_2$ *where* $A, B \in C^{m\times n\times p}$*and* $A^k, B^k \in C^{m\times n}$.

$$
\begin{aligned}
\parallel A + B \parallel_2 &= max\big(max_{\|x\|_2=1;k=1,\dots,p} \parallel (A^{(k)} + B^{(k)})x \parallel_2 \big) \\
&= max\big(max_{\|x\|_2=1;k=1,\dots,p} \parallel A^{(k)}x + B^{(k)}x \parallel_2 \big) \\
&\leq max\big(max_{\|x\|_2=1;k=1,\dots,p} \parallel A^{(k)}x \parallel_2 + \parallel B^{(k)}x \parallel_2 \big) \\
&\leq max\big(max_{\|x\|_2=1;k=1,\dots,p} \parallel A^{(k)}x \parallel_2 \big) + max\big(max_{\|x\|_2=1;k=1,\dots,p} \parallel B^{(k)}x \parallel_2 \big) \\
&= \parallel A \parallel_2 + \parallel B \parallel_2
\end{aligned}
$$

3 Impression Matrix Norm

In financial markets we face with dynamic large data sets. Investors are curious about real-time market impression under fluctuating volatility, interest rate, and market price changes. It is hard to handle lots of variables at the same time. Therefore, it is important to define a proxy to reflect market impression quickly. Consider a 3-dimensional matrix with time, interest rate, and stochastic volatility dimensions where matrix entries are market prices. As an application of 3-dimensional norms we define impression matrix norm.

We define impression matrix norm (IMN) as a norm of the moving matrix by the time. It is generated by evaluating the norm of the matrix at each related time subinterval. IMN of the 3-D matrix gives a good picture of all 3-D matrix data and helps us to understand and interpret 3-D matrix more easily. We perform the simulations and obtain graph in Fig. 1 by using the parameters in Table 1.

Interest rates generally change randomly at the real market. To converge the real market behavior by this aspect, we start with initial interest rate and generate 280 new random interest rates by adding a random number within $\%[-10, 10]$ to the previous interest rate at each step. Then we perform simulations and get 3-dimensional stock price expectation matrix M ($M \in C^{1000\times 100\times 280}$). If we use IMN to analyze M, then we obtain the graph in Fig. 2.

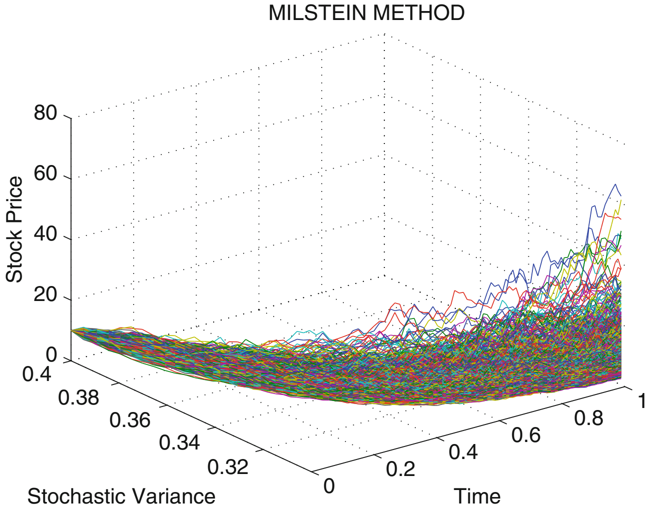
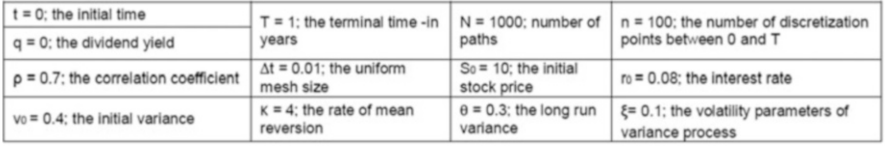


Fig. 1 One thousand realizations of simulation

Table 1 Simulation parameters



t = 0; the initial time	T = 1; the terminal time -in years	N = 1000; number of paths	n = 100; the number of discretization points between 0 and T
q = 0; the dividend yield			
ρ = 0.7; the correlation coefficient	Δt = 0.01; the uniform mesh size	S₀ = 10; the initial stock price	r₀ = 0.08; the interest rate
v₀ = 0.4; the initial variance	κ = 4; the rate of mean reversion	θ = 0.3; the long run variance	ξ= 0.1; the volatility parameters of variance process

4 Conclusions

We find that defining 3-dimensional matrix norms, summarizing large financial data set, and quantifying market impression in the presence of several variables together are useful. We obtain a proxy for time evolution of market impression value and perform simulations for various model parameters.

As a next step, we extend our approach by using several numerical solution methods for Heston stochastic volatility model and by applying to large data sets such as Borsa Istanbul-100 (BIST-100) and BIST-30, in another paper study.

Appendix

Milstein Method

Definition. Let $\overline{y}_N$ be the numerical approximation to $y(t_N)$ after N steps with constant stepsize $h = (t_N - t_0/N)$; then $\overline{y}$ is said to converge strongly to y with order p if $\exists C > 0$ (independent of h) and $\delta > 0$ such that

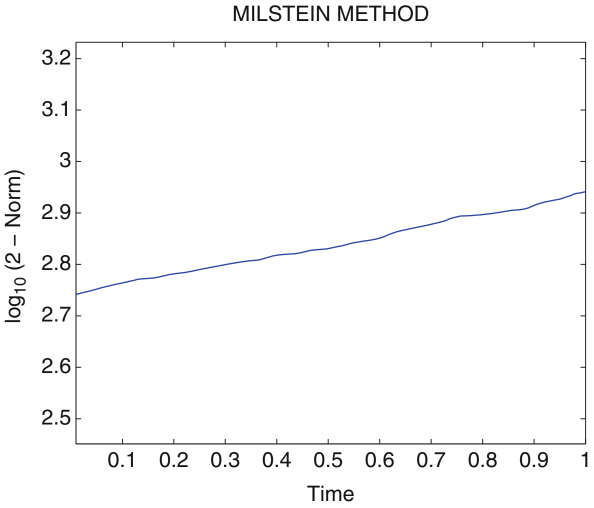


Fig. 2 Impression norm of matrix M

$$E(\|\overline{y}_N - y(t_N)\|) \leq Ch^p,\ h \in (0, \delta).$$

Lets consider the following SDEs :

$$dy = f(t, y)dt + g(t, y)dW,\ y(0) = y_0. \tag{1}$$

Milstein method has the following form for equation (1):

$$\begin{aligned}\Delta y_i &= f(t_i, y_i)\Delta t_i + g(t_i, y_i)\Delta W_i + \frac{1}{2}g(t_i, y_i)\frac{\partial g}{\partial y}(t_i, y_i)(\Delta W_i^2 - \Delta t_i)\\ \Delta t_i &= t_{i+1} - t_i\\ \Delta W_i &= W_{i+1} - W_i\end{aligned}$$

Milstein method has strong order 1 for solving SDEs. Also Brownian motion ΔW_i can be modeled as $\Delta W_i = z_i\sqrt{\Delta t_i}$ where z_i is chosen from N(0,1) standard normally random variable (see [3]).

Heston Model

In Heston's stochastic volatility model the asset price process S_t and the variance process $v_t := \sigma_t^2$ solve the following two-dimensional stochastic differential equation (see [1]):

$$dS_t = (r - q)S_t dt + \sqrt{v_t} S_t dW_1(t)$$
$$dv_t = \kappa(\theta - v_t)dt + \xi \sqrt{v_t} dW_2(t)$$

At the Black-Scholes-Merton (BSM) model (see [7]), the volatility σ was assumed to be constant. The main difference between BSM and Heston model is volatility behavior. It is stochastic and it satisfies mean-reverting property with a mean-reverting drift at the Heston model. The W_1andW_2 represent Brownian motions of asset price process and the variance process is correlated with correlation coefficient $\rho \in [-1, 1]$. Here $\xi > 0$ is the volatility parameter of the variance process, $r \geq 0$ is the risk-free interest rate, $q \geq 0$ is the dividend yield, $\kappa > 0$ is the rate of mean reversion, and $\theta > 0$ is the long-run variance level (see [1]). Stochastic volatility model of Heston (1993) is frequently used. Heston's model is derived from the CIR model of Cox, Ingersoll, and Ross (1985) for interest rates (see [8]). We choose the parameters as they satisfy the Feller condition $2\kappa\theta \geq \xi^2$ at our simulations so that non-negativity of volatility can be guaranteed (see [9]).

References

1. Heston, S.: A Closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**(2), 327–343 (1993)
2. Kloeden, P.E., Platen, E., Schurz, H.: Numerical Solution of SDE Through Computer Experiments. Springer, Berlin (2003)
3. Milstein, G.N.: Approximate integration of stochastic differential equations. Theor. Prob. Appl. **19**, 557–562 (1974)
4. Bodie, Z., Kane, A., Marcus, A.J.: Investments, 6th ed. Irwin McGraw-Hill, Boston, MA (2005)
5. Trefethen, L.N., Bau III, D.: Numerical Linear Algebra. SIAM, Philadelphia (1997)
6. Royden, H.L.: Real Analysis, 3rd edn. Macmillan, New York (1988)
7. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
8. Cox, J.C., Ingersoll, J.E., Ross, S.A.: A Theory of the term structure of interest rates. Econometrica **53**, 385–407 (1985)
9. Feller, W.: Two singular diffusion problems. Ann. Math. **54**, 173–182 (1951)

Approximations to the Solution of the Frank-Kamenetskii Equation in a Spherical Geometry

Moustafa Aly Soliman

Abstract In this paper, an approximate analytical solution for Frank-Kamenetskii equation modeling a thermal explosion in a sphere, is obtained. The approximate solution is obtained by perturbation methods in terms of small and large distance parameter. The approximate solution is compared with the numerical solution obtained from an initial value problem formulation to the original boundary value problem. The approximate solution obtained is valid for all values of the distance parameter. For the original boundary value problem and for a given Frank-Kamenetskii parameter, a nonlinear algebraic equation needs to be solved to be able to apply the approximate solution.

Keywords Frank-Kamenetskii equation • Thermal explosion • Stellar structure • Perturbation • Sphere

1 Introduction

Several theoretical studies and numerical methods were used for the study of the Frank-Kamenetskii equation [1–12] which models thermal explosion in an enclosure. The equation in a spherical enclosure occurs in the theory of stellar structure. The equation also models a non-isothermal zero order reaction in a catalyst particle. Frank-Kamenetskii [1] formulated the problem and obtained analytical solutions for the case of a slab and cylinder enclosure. More results on the case of a slab and cylinder can be found in references [13–16]. The spherical enclosure case can so far only be obtained numerically.

M.A. Soliman (✉)
Chemical Engineering Department, The British University in Egypt, El Sherouk City,
Cairo 11837, Egypt
e-mail: moustafa.aly@bue.edu.eg

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_21,

The Frank-Kamenetskii equation is given by

$$\frac{d^2u}{dr^2} + \frac{s}{r}\frac{du}{dt} = -\lambda \exp(u) \tag{1}$$

with the boundary conditions

$$u(1) = 0 \tag{2}$$

$$\frac{du}{dr} \mid_{r=0} = 0 \tag{3}$$

where s is a shape factor that takes a value of 0 for an infinite slab, 1 for an infinite cylinder, and 2 for a sphere; u is dimensionless temperature; and r is dimensionless distance.

This equation was treated by different investigators and was thus given different names such as Bratu, Liouville, Gelfand, and Frank-Kamenetskii equation.

The parameter λ is usually given the name of Frank-Kamenetskii parameter. If λ is greater than a critical value, explosion occurs and there is no solution for the equation. For λ less than the critical value, two solutions exist for the case of slab and cylinder. For the case of a sphere for different values of λ, we can have no solution, one, two, or multiple number of solutions. The solution is characterized by infinite number of solutions at $\lambda = 2$.

The aim of the present paper is to give an approximate analytical solution for equations (1–3).

In the next section, we transform the boundary value problem to an initial value problem and obtain approximate solutions for this initial value problem for large and small distance parameter using perturbation methods. Then we modify the solution obtained for large distance parameter to be also valid for the small distance parameter case. Thus we obtain one general approximate solution valid for all values of the distance parameter. Finally numerical results are presented.

2 Mathematical Development

First we change the boundary value problem to an initial value problem as follows:

Let

$$\psi = u_0 - u, u_0 = u(0) \tag{4}$$

and

$$\xi^2 = \lambda r^2 \exp(u_0) \tag{5}$$

Equations (1–3) become

$$\frac{d^2\psi}{d\xi^2} + \frac{s}{\xi}\frac{d\psi}{d\xi} = \exp(-\psi) \tag{6}$$

$$\psi(0) = 0 \tag{7}$$

$$\frac{d\psi}{d\xi}\ |_{\xi=0} = 0 \tag{8}$$

Tables for ψ against ξ for the case of a sphere are given by Chandrasekhar and Wares [16] and Horedt [17].

2.1 Approximate Solution for Large ξ

We treat equation (6) for the case of a sphere (s=2).

We notice that

$$\psi = 2\ln(\xi) - \ln(2) \tag{9}$$

satisfies equation (6) but does not satisfy initial conditions (7, 8). This solution is called singular solution [2]. This solution approaches the exact solution as ξ tends to infinity.

Define

$$\zeta = \ln(\xi) \tag{10}$$

$$\phi = \psi - 2\ln(\xi) + \ln(2) = \psi - 2\zeta + \ln(2) \tag{11}$$

Substituting equation (11) into equation (6), we obtain

$$\frac{d^2\varphi}{d\zeta} + \frac{d\varphi}{d\zeta} + 2 = 2\exp(-\varphi) \tag{12}$$

For small φ (this happens as ξ tends to ∞), we can approximate equation (12) by

$$\frac{d^2\varphi}{d\zeta^2} + \frac{d\varphi}{d\zeta} + 2\varphi = 0 \tag{13}$$

which has a solution in the form

$$\varphi = A\exp\left(-\frac{\zeta}{2}\right)\sin\left(B + \frac{\sqrt{7}}{2}\zeta\right) \tag{14}$$

or

$$\varphi = \frac{A}{\sqrt{\xi}} \sin\left(B + \frac{\sqrt{7}}{2}\ln\xi\right) \tag{15}$$

where A and B are constants to be determined from the numerical solution. φ being small, we can write

$$\varphi = \ln\left(1 + \frac{A}{\sqrt{\xi}} \sin\left(B + \frac{\sqrt{7}}{2}\ln\xi\right)\right) \tag{16}$$

This formula is similar to what was obtained by Chandrasekhar [3] and Adler [9]

Thus from equations (11) and (15), we have

$$\psi = \ln\left(\frac{\xi^2}{2}\right) + \frac{A}{\sqrt{\xi}} \sin\left(B + \frac{\sqrt{7}}{2}\ln\xi\right) \tag{17}$$

We have solved equations (6–8) numerically using DASSL routine [18] and determined the constants A and B. In fact we used the two points $(\xi, \psi) = (210650.464, 23.823), (525665.99, 25.653)$ to obtain A=−1.178, B=−0.507787. The first point corresponds to $\lambda = 2$ at which $\phi = 0$ and the second point to a turning point.

Equation (17) becomes

$$\psi = \ln\left(\frac{\xi^2}{2}\right) - \frac{1.178}{\sqrt{\xi}} \sin\left(-0.507787 + \frac{\sqrt{7}}{2}\ln\xi\right) \tag{18}$$

By definition, the value of ψ at $\lambda = 2$ is

$$\psi = \ln\left(\frac{\xi^2}{2}\right) \tag{19}$$

so that the condition for $\lambda = 2$ is

$$\sin\left(-0.507787 + \frac{\sqrt{7}}{2}\ln\xi\right) = 0 \tag{20}$$

or

$$\left(-0.507787 + \frac{\sqrt{7}}{2}\ln\xi\right) = n\pi \qquad \text{(n is an integer)} \tag{21}$$

giving

$$\xi = \exp\left\{\frac{2}{\sqrt{7}}(0.507787 + n\pi)\right\} = 1.4679(10.74909)^n \tag{22}$$

Enig [8] has shown that for any shape, the following relation is satisfied at the turning points:

$$\frac{d\psi}{d\zeta} = 2 \tag{23}$$

This means that

$$\frac{d}{d\xi}\left[\frac{1}{\sqrt{\xi}}\sin\left(-0.507787 + \frac{\sqrt{7}}{2}\ln\xi\right)\right] = 0 \tag{24}$$

giving

$$\tan\left(-0.507787 + \frac{\sqrt{7}}{2}\ln\xi\right) = \sqrt{7} \tag{25}$$

or

$$\xi = \exp\left\{\frac{2}{\sqrt{7}}(0.507787 + n\pi + 1.209429)\right\} = 3.6623(10.74909)^n \tag{26}$$

This is the condition for a turning point.

We could repeat the same analysis but defining

$$\zeta = \ln\left(1 + \frac{\xi^2}{2}\right) \tag{27}$$

We arrive at the following formula:

$$\psi = \ln\left(1 + \frac{\xi^2}{2}\right) - \frac{1.178}{\sqrt[4]{2+\xi^2}}\sin\left(-0.507787 + \frac{\sqrt{7}}{4}\ln(2+\xi^2)\right) \tag{28}$$

This formula would extend the range of applicability of equation (18).

We could further improve the accuracy of equation (28) by writing

$$\psi = \ln\left(1 + \frac{\xi^2}{2}\left(1 - \frac{1.178}{\sqrt[4]{2+\xi^2}}\sin\left(-0.507787 + \frac{\sqrt{7}}{4}\ln(2+\xi^2)\right)\right)\right) \tag{29}$$

The condition for $\lambda = 2$ is given by

$$\xi^2 = \exp\left\{\frac{4}{\sqrt{7}}(0.507787 + n\pi)\right\} - 2 \tag{30}$$

and for a turning point is given by

$$\xi^2 = \exp\left\{\frac{4}{\sqrt{7}}(0.507787 + n\pi + 1.209429)\right\} - 2 \tag{31}$$

2.2 Approximate Solution for All ξ

The following formula was synthesized from equation (29)

$$\psi = \ln\left(1 + \frac{\xi^2}{2}\left(1 - \frac{2}{3\sqrt[4]{1 + \xi^2/15}}\cos\left(\frac{\sqrt{7}}{4}\ln\left(1 + \xi^2/23.162231\right)\right)\right)\right) \tag{32}$$

Notice

$$\exp\left(\frac{4}{\sqrt{7}}\left(0.507787 + \frac{\pi}{2}\right)\right) = 23.162231$$

Equation (32) contains the main components obtained for large ψ and satisfies the small ψ approximation:

$$\psi \cong \frac{1}{6}\xi^2 - \frac{1}{120}\xi^4 \tag{33}$$

For equation (32) to approach equation (29) for large ξ, the constant pre-multiplying the cosine function (2/3) should be 0.59858

since

$$1.178/\sqrt[4]{15} = 0.59858$$

We can then suggest a general formula of the following form:

$$\psi = \ln\left(1 + \frac{\xi^2}{2}\left(1 - \frac{\left(2/3 + 0.59858{*}10^{-6}\xi^6\right)}{(1 + 10^{-6}\xi^6)\sqrt[4]{1 + \xi^2/15}}\right.\right.$$

$$\left.\left.\cos\left(\frac{\sqrt{7}}{4}\ln\left(1 + \xi^2/23.162231\right)\right)\right)\right) \tag{34}$$

3 Numerical Results and Discussion

We have chosen the DASSL FORTAN code [18] to solve equations (6–8). It uses backward differentiation formula method to solve a system of differential algebraic equations. We were able to reproduce the table of Horedt [17] up to a value of $\psi = 14.7$ and go beyond that to a value of 26. Numerical results show that equation (34) is slightly better than equation (32), but both of them are indistinguishable with the numerical solution. To solve the original problem (equations (1–3)) for a given λ we need to solve the nonlinear algebraic equation (5).

4 Conclusions

Approximate analytical solution for the Frank-Kamenetskii equation valid for all values of λ was derived. Numerical solutions of the equation show that the approximate solution is of good accuracy. The Frank-Kamenetskii parameter λ at the turning points has been evaluated to five decimal places. In catalysis, the equation represents an approximation for zero-order reaction in spherical catalyst particle. Future work will be thus to extend the approximate solution to a general order reaction.

References

1. Frank-Kamenetskii, D.A.: Diffusion and Heat Exchange in Chemical Kinetics. Princeton University Press, Princeton, NJ (1955)
2. Aris, R.: The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts, Vol. I. Oxford University Press, New York (1975)
3. Chandrasekhar, S.: An Introduction to the Study of Stellar Structure. Dover Publications, New York (1967)
4. Steggerda, J.J.: Thermal stability: an extension of Frank-Kamenetskii's theory. J. Chem. Phys. **43**, 4446–4448 (1965)
5. Hlavacek, V., Marek, M.: Modelling of chemical reactors-IX the non-isothermal zero-order reaction within a porous catalyst particle. Chem. Eng. Sci. **23**, 865–880 (1968)
6. Moise, A., Pritchard, H.O.: Newton-variational solution of the Frank-Kamenetskii thermal explosion problem. Can. J. Chem. Eng. **67**, 442–445 (1989)
7. Nazari-Golshan, A., Nourazar, S.S., Ghafoori-Fard, H., Yildirim, A., Campo, A.: A modified homotopy perturbation method coupled with the Fourier transform for nonlinear and singular Lane-Emden equations. Appl.Math. Lett. **26**, 1018–1025 (2013)
8. Enig, J.W.: Critical parameters in the poisson-boltzmann equation of steady-state thermal explosion theory. Combust. Flame **10**, 197–199 (1967)
9. Adler, J.: The spherical Liouville and associated differential equations. IMA J. Appl. Math. **76**, 817–833 (2011)
10. Bazley, N.W., Wake, G.C.: Criticality in a model for thermal ignition in three or more dimensions. Z. Angew. Math. Phys. **32**, 594–602 (1981)

11. Gustafson, K.E., Eaton, B.E.: Exact solutions and ignition parameters in the Arrhenius conduction theory of gaseous thermal explosion. Z. Angew. Math. Phys. **33**, 392–405 (1982)
12. Britz, D., Strutwolf, J., Osterby, O.: Digital simulation of thermal reactions. Appl. Math. Comput. **218**, 1280–1290 (2011)
13. Boyd, J.P.: One-point pseudospectral collocation for the One dimensional Bratu equation. Appl. Math. Comp. **217**, 5553–5565 (2011)
14. Soliman, M.A.: Rational approximation for the one-dimensional Bratu equation. International Journal of Engineering and Technology **13**(5), 24–39 (2013)
15. Harley, C., Momoniat, E.: Alternate derivation of the critical value of the Frank-Kamenetskii parameter in cylindrical geometry. J. Nonlinear Math. Phys. **15**, 69–76 (2008)
16. Chandrasekhar, S., Wares, G.W.: The isothermal function. Astrophys. J. **109**, 551–554 (1949)
17. Horedt, G.P.: Seven-digit tables of Lane-Emden functions. Astrophys. Space Sci. **126**, 357–408 (1986)
18. Petzold, L.: A description of DASSL: A differential algebraic system solver. SAND 82–8637 (1982) (September)

Blowup of Series Solutions on the Half Line

A. Boumenir

Abstract We prove the convergence of series solutions of a semilinear reaction diffusion equation on the half line with quadratic nonlinearity. We also construct a positive solution that blows up in finite time. The algorithm, which is based on algebraic operations only, is fast and can be used to approximate and extend these solutions beyond blowup.

Keywords KPP equation • Series solution • Decomposition method

1 Introduction

We are concerned with the classical one-dimensional nonlinear reaction diffusion equation defined by

$$\begin{cases} u_t = u_{xx} + au + \lambda u^2 & x \in [0,\infty), \quad t \geq 0 \\ u(0,t) + u_x(0,t) = 0 & \text{and } u(x,0) = b(x) \end{cases} \tag{1}$$

It is well known that, depending on the initial condition $b(x)$, the solution may blow up in finite time and can do so at a single point [2, 3]. This makes the use of numerical methods, such as finite differences, challenging as they might miss the singularity. To overcome these difficulties we look for a nonnumerical method and seek a solution of (1) in the form

$$u(x,t) = \sum_{n \geq 0} \lambda^{n-1} e^{n\beta t} \psi_n(x) \tag{2}$$

A. Boumenir (✉)
Department of Mathematics, Kuwait University, Kuwait
e-mail: boumenir@sci.kuniv.edu.kw

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_22,

where ψ_n are bounded continuous positive functions that satisfy a family of linear differential equations. Formula (2) extends the idea of separation of variables to nonlinear equations and ψ_n are computed recursively, using simple algebraic operations only and no numerical integration is used as in [1]. Due to the exponential coefficients in (2), the solution is expected to blow up in finite time if $\beta > 0$ and exists globally if $\beta < 0$, thus, the need to study the main properties, such as smoothness and boundedness, of the sequence $\{\psi_n\}_{n\geq 0}$. The series solutions are important from the geometrical point of view as they help us understand how singularities develop and evolve and how a solution can be extended beyond blowup [4–6]. It is obvious that blowup is a simple matter of radius of convergence of the series in (2). Solutions given by (2) are also important from the computational point of view, as they can benchmark standard numerical methods for accuracy. The truncation error can be estimated explicitly which then help produce guaranteed error bounds.

2 The Sequence ψ_n

If the series (2) is a solution of (1), then the ψ_n should verify

$$\sum_{n\geq 0}\lambda^{n-1}e^{\beta n t}\beta n\psi_n(x)=\sum_{n\geq 0}\lambda^{n-1}e^{\beta n t}\left(\psi_n''(x)+a\psi_n(x)\right)$$

$$+\sum_{n\geq 0}\lambda^{n-1}e^{\beta n t}\sum_{k=0}^{n}\psi_k(x)\psi_{n-k}(x) \tag{3}$$

and the boundary condition

$$\psi_n(0)+\psi_n'(0)=0. \tag{4}$$

Factoring out $\lambda^{n-1}e^{\beta n t}$ yields a sequence of Sturm-Liouville equations

$$\beta n\psi_n(x)=\psi_n''(x)+a\psi_n(x)+\sum_{k=0}^{n}\psi_k(x)\psi_{n-k}(x) \qquad n=0,1,\ldots. \tag{5}$$

Setting $n=0$ in (5) gives

$$\psi_0''(x)+a\psi_0(x)+\psi_0^2(x)=0 \tag{6}$$

which is the rescaled steady state, $\lim_{t\to\infty}\lambda u(x,t)=\psi_0(x)$. The sequence ψ_n, for $n\geq 2$, is defined by

$$\psi_n''(x)+(a-\beta n-2\psi_0(x))\,\psi_n(x)=-R_n(x) \tag{7}$$

where the right-hand side is made of previously computed ψ_n

$$R_n(x) = \sum_{k=1}^{n-1} \psi_k(x)\psi_{n-k}(x). \tag{8}$$

For simplicity, we select the trivial steady state $\psi_0 = 0$, which reduces (7) to

$$\psi_n''(x) + (a - \beta n)\,\psi_n(x) = -R_n(x) \tag{9}$$

with $R_1(x) = 0$, and so the first equation of the sequence (7) is simply

$$\begin{cases} \psi_1''(x) + (a - \beta)\,\psi_1(x) = 0 \qquad x > 0 \\ \psi_1(0) + \psi_1'(0) = 0. \end{cases} \tag{10}$$

If $0 < \beta - a$ then a general solution is a linear combination of $\exp\left(\pm x\sqrt{\beta - a}\right)$ and so an $L^2(0, \infty)$ solution is given by $\exp\left(-x\sqrt{\beta - a}\right)$ and for the boundary condition to hold we must have $\beta - a = 1$

$$\beta = 1 + a \tag{11}$$

and so in all cases

$$\psi_1(x) = \exp(-x).$$

In order to proceed further define the self-adjoint operator acting in $L^2(0, \infty)$ by

$$\begin{cases} A(y) = -y''(x) \quad \text{for } x \geq 0 \\ y(0) + y'(0) = 0 \end{cases}$$

whose spectrum is given by

$$\sigma_A = \{-1\} \cup [0, \infty).$$

Thus (9) becomes

$$(A + (n\beta - a))\,\psi_n(x) = R_n(x) \tag{12}$$

and in order to generate a nontrivial sequence ψ_n from (12), we need $a - n\beta \notin \sigma_A$. If $a > 0$, then condition (11) is enough to imply

$$a - n\beta < -1 \quad \text{for} \quad n \geq 2. \tag{13}$$

The solutions ψ_n can then be found recursively and explicitly, since $R_n(x)$ has a simple expression:

$$\begin{aligned}
&\psi_2''(x) + (a - 2\beta)\,\psi_2(x) = -\psi_1^2(x)\\
&\psi_3''(x) + (a - 3\beta)\,\psi_3(x) = -2\psi_1(x)\psi_2(x)\\
&\psi_4''(x) + (a - 4\beta)\,\psi_4(x) = -2\psi_1(x)\psi_3(x) - \psi_2^2(x).
\end{aligned}$$

For example, since $R_2(x) = \psi_1^2(x) = \exp\left(-2x\sqrt{\beta - a}\right) > 0$, then the next function ψ_2 is the solution of

$$\psi_2''(x) + (a - 2\beta)\,\psi_2(x) = -\exp\left(-2x\sqrt{\beta - a}\right) < 0$$

whose general solution is

$$\psi_2(x) = c_{21}\exp\left(-x\sqrt{2\beta - a}\right) + \frac{1}{3a - 2\beta}\exp\left(-2x\sqrt{\beta - a}\right),$$

while c_{21} is determined by the boundary condition $\psi_2(0) + \psi_2'(0) = 0$, i.e.,

$$c_{21} = \frac{1}{2\beta - 3a}\frac{\left(1 - 2\sqrt{\beta - a}\right)}{\left(1 - \sqrt{2\beta - a}\right)}.$$

Next we compute

$$\begin{aligned}
R_3(x) = 2\psi_1(x)\psi_2(x) &= 2c_{21}\exp\left(-x\left(\sqrt{2\beta - a} + \sqrt{\beta - a}\right)\right)\\
&+ \frac{2}{3a - 2\beta}\exp\left(-3x\sqrt{\beta - a}\right) \qquad (14)
\end{aligned}$$

which yields

$$\begin{aligned}
\psi_3(x) = c_{31}\exp\left(-x\sqrt{3\beta - a}\right) &+ c_{32}\exp\left(-x\left(\sqrt{2\beta - a} + \sqrt{\beta - a}\right)\right)\\
&+ c_{33}\exp\left(-3x\sqrt{\beta - a}\right). \qquad (15)
\end{aligned}$$

By induction it is easy to see that

$$\psi_n(x) = \sum_{k=1}^{m_n} c_{nk}\exp\left(-xa_{nk}\right) \qquad (16)$$

and the constants a_{nk} and c_{nk} are easily computed algebraically which makes the algorithm fast. In order to track down these values a_{nk} and find the properties of ψ_n we work out a specific case. For example, we can choose

$$a = 1 \quad \text{and} \quad \beta = 2 \tag{17}$$

to satisfy (11) which then yields (13). In this case the first few terms of the sequence for $n \geq 2$

$$\begin{cases} \psi_n''(x) + (1-2n)\,\psi_n(x) = -R_n(x) & x > 0 \\ \psi_n(0) + \psi_n'(0) = 0 \end{cases} \tag{18}$$

are

$$\begin{aligned} \psi_1(x) &= \exp(-x) \\ \psi_2(x) &= \frac{1}{(-1+\sqrt{3})}\exp(-\sqrt{3}x) - \exp(-2x) \\ \psi_3(x) &= \frac{\left(-1+\sqrt{3}\right)}{\left(8-5\sqrt{3}\right)}\exp(-(1+\sqrt{3})x) \\ &\quad + \frac{1}{2}\exp(-3x) - \frac{(6\sqrt{3}-11)}{(8-5\sqrt{3}+5\sqrt{5}\sqrt{3}-8\sqrt{5})}\exp(-\sqrt{5}x), \end{aligned} \tag{19}$$

and the solution looks like

$$u(x,t) = \sum_{n\geq 1} e^{2nt}\lambda^{n-1}\psi_n(x) = \psi_1(x)\exp(2t) + \psi_2(x)\lambda\exp(4t) + \psi_3(x)\lambda^2\exp(6t) + \dots \,. \tag{20}$$

To prove its convergence, it is enough to find a bound on the functions ψ_n.

Proposition 1. *For $n \geq 1$ and $x > 0$ we have $0 < \psi_n(x) < 1$ and ψ_n is a decreasing function of x.*

Proof. We use induction. It is certainly true from (19) for $n = 1, 2, 3$. Assume it is true up to $n-1$, then we deduce from (8),

$$1 - n \leq -R_n(x) \leq 0$$

out of which follows that 1 is an upper solution since

$$\begin{cases} 0 - (2n-1)\,1 \leq -R_n(x) \\ 1 \geq 0 \end{cases}$$

and that 0 is lower solution

$$\begin{cases} 0-(2n-1)\,(0) \geq -R_n(x) \\ 0=0. \end{cases}$$

Thus $0<\psi_n(x)<1$. We also use induction to show that $\psi_n'(x)<0$. Obviously $\psi_1(x)=\exp(-x)$ is decreasing and if it is true up to $n-1$, then $\psi_n'(x)$ satisfies, see (18)

$$\begin{cases} \psi_n'''(x)+(1-2n)\,\psi_n'(x)=-R_n'(x)>0 & x>0 \\ \psi_n'(0)=-\psi_n(0)<0. \end{cases}$$

By the maximum principle, it follows that 0 is an upper solution and so $\psi_n'(x)<0$ for $x>0$.

The fact that ψ_n are bounded, $0<\psi_n(x)<1$, implies that when $\left|e^{2t}\lambda\right|<1$, the series converges for $0\leq t<T$ where $T=-\ln\left(\sqrt{\lambda}\right)$ and

$$u(x,t)=\sum_{n\geq 1}\lambda^{n-1}e^{2nt}\psi_n(x)\leq u(0,t). \tag{21}$$

We next show that it is a solution by examining its partial sums. □

Proposition 2. *Let* $0<\lambda<1$, $a=1$ *and* $\beta=2$, *then the solution of (1) is defined by (20) which exists for* $t\in[0,T)$.

Proof. We need to examine the partial sums for $t\in[0,T)$ and $x\geq 0$

$$u(k)=\sum_{n=1}^{n=k}\lambda^{n-1}e^{2nt}\psi_n(x).$$

We have that $u_t(k)=\sum_{n=1}^{n=k}2n\lambda^{n-1}e^{2nt}\psi_n(x)$, $u_{xx}(k)+u(k)=\sum_{n=1}^{n=k}\lambda^{n-1}e^{2nt}\left[\psi_n''(x)+\psi_n(x)\right]$ and

$$\tilde{R}(k)=\sum_{n=1}^{n=k}\lambda^{n-1}e^{2nt}R_n\,(x)$$

Use (18) and proposition 1 to see that

$$\left|\psi_n''(x)+\psi_n(x)\right|=|2n\psi_n(x)-R_n(x)|\leq 2n+n-1<3n$$

$$|R_n(x)|=\sum_{k=1}^{n-1}\psi_k(x)\psi_{n-k}(x)\leq n-1$$

Thus $u_t(k), u_{xx}(k) + u(k)$ and $\tilde{R}(k)$ converge uniformly, as $k \to \infty$, on any compact $[0, T - \varepsilon] \times [0, N]$, for any ε and N where $T > \varepsilon > 0$ and $N < \infty$. Moreover it follows from (5) that

$$u_t(k) = u_{xx}(k) + u(k) + \tilde{R}(k) \quad \text{for all } k \geq 1$$

and thus at the limit we obtain (3) which means that the series defined by (20) is a solution to (1). The boundary condition at $x = 0$ is easily seen to hold and $u(x, 0) = \sum_{n\geq 1} \lambda^{n-1} \psi_n(x)$ which is continuous on $(0, \infty)$ since $0 < \psi_n(x) < 1$.

□

3 Blowup

A sufficient condition for (11) and (13) to hold is simply

$$\beta = a + 1 > 0. \tag{22}$$

To show that the solution blows up in finite time it is enough to show that the integral

$$h(t) = \int_0^\infty u(x, t)\psi_1(x)\, dx \to \infty \quad \text{as } t \to T_b < \infty$$

since $\psi_1(x) = \exp(-x)$. Next multiplying equation (1) by $\psi_1(x)$ leads to

$$\int_0^\infty u_t(x, t)\psi_1(x)dx = \int_0^\infty u_{xx}(x, t)\psi_1(x)dx + a\int_0^\infty u(x, t)\psi_1(x)dx + \lambda \int_0^\infty u^2(x, t)\psi_1(x)dx \tag{23}$$

and $u_x(0, t) + u(0, t) = 0$ yields

$$\int_0^\infty u_{xx}(x, t)\ \psi_1(x)dx = \int_0^\infty u(x, t)\psi_1(x)dx.$$

Now since $\int_0^\infty \psi_1(x)dx = 1$, Jensen's inequality yields

$$\int_0^\infty u^2(x, t)\psi_1(x)dx \geq \left(\int_0^\infty u(x, t)\psi_1(x)dx\right)^2. \tag{24}$$

Combine (24) and (23) to see that h satisfies the inequality

$$h'(t) \geq \beta h(t) + \lambda h^2(t).$$

Thus if $h(0) > 0$ we have blow up in finite time, T_b say, and

$$T_b < \frac{1}{\beta} \ln \left(1 + \frac{\beta}{\lambda h(0)} \right). \tag{25}$$

It is also easy to see from (21) that if blowup occurs at a single point, then it has to be at $x = 0$, which is on the boundary. The nature of the singularity depends only on the convergence of the series:

$$u(x, T_b) = \sum_{n \geq 1} \lambda^{n-1} e^{n\beta T_b} \psi_n(x)$$

Summarizing the above we have

Proposition 3. *Assume that* $0 < \lambda < 1$*, and (22) holds, then the solution defined by (2) blows up in finite time.*

Proof. Since $u(x, 0) = \sum_{n \geq 1} \lambda^{n-1} \psi_n(x) > 0$ then $h(0) > 0$ and so the solution blows up before T_b, (25). As for the location of the blowup point, observe that ψ_n are decreasing, and so $\psi_n(x) < \psi_n(0)$ which means that $u(x, t) < u(0, t)$. Thus, if the solution blows up at a single point, it has to be at $x = 0$. □

References

1. Boumenir, A.: Power series solutions for the KPP equation. Numer. Algor. **43**(2), 177–187 (2006)
2. Friedman, A., Mcleod, J.B.: Blow-up of positive solutions of semilinear heat equations. Indiana Univ. Math. J. **34**, 425–447 (1985)
3. Fujita, H.: On the blowing up of solutions to the Cauchy problem J. Fac. Sci. Univ. Tokyo, Sect. IA, Math. **18**, 109–124 (1966)
4. Galaktionov, V.A.: On new exact blow-up solutions for nonlinear heat conduction equations, with source and applications. Diff. Integr. Equat. **3**, 863–874 (1990)
5. Galaktionov, V.A., Vazquez, J.L.: Continuation of blowup solutions of nonlinear heat equations in several space dimensions. Comm. Pure Appl. Math. **50**(1), 1–67 (1997)
6. Galaktionov, V.A., Svirshchevskii, S.R.: Exact solutions and invariant subspaces of nonlinear partial differential equations in mechanics and physics. Chapman Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman Hall/CRC, Boca Raton, FL (2007)

Two Models of Subdiffusion Processes: When Are They Similar?

T. Kosztołowicz and K.D. Lewandowska

Abstract We study two models which describe subdiffusive processes. Subdiffusion is defined by the relation $\left\langle (\Delta x)^2(t) \right\rangle = D_\alpha t^\alpha$, where $\left\langle (\Delta x)^2(t) \right\rangle$ denotes the mean square displacement, α is a subdiffusion parameter which obeys $0 < \alpha < 1$ and D_α is a subdiffusion coefficient. The first model consists of a nonlinear partial differential equation with derivatives of a natural order obtained from a Sharma–Mittal nonadditive entropy, whereas the second model is based on a linear partial differential equation with a fractional time derivative which is derived from the continuous time random walk formalism. We obtain the fundamental solutions for both models. Next, we assume two agreement conditions. According to the first one the fundamental solutions for both model fulfill the relation which defines subdiffusion. The second agreement condition presumes the equality of the first passage time distributions. On the basis of these agreement conditions we answer the question when the considered models are similar.

Keywords Subdiffusion • Fractional derivative • Nonadditive entropy

T. Kosztołowicz (✉)
Institute of Physics, Jan Kochanowski University, ul. Świętokrzyska 15, 25-406 Kielce, Poland
e-mail: tadeusz.kosztolowicz@ujk.edu.pl
web page: http://www.ujk.edu.pl/strony/tadeusz.kosztolowicz/

K.D. Lewandowska
Department of Radiological Informatics and Statistics, Medical
University of Gdańsk, ul. Tuwima 15, 80-210 Gdańsk, Poland
e-mail: kale@gumed.edu.pl

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings
in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_23,

1 Introduction

Subdiffusion can be defined as a transport process in which the mean square displacement of a random walker fulfills the following relation [1,2]:

$$\left\langle (\Delta x)^2(t) \right\rangle = D_\alpha t^\alpha \ , \tag{1}$$

where $\alpha \in (0,1)$ is a subdiffusion parameter and D_α denotes a subdiffusion coefficient given in the units of m^2/t^α. When $\alpha = 1$ we are dealing with normal diffusion. The mean–square displacement can be calculated according to the following formula:

$$\left\langle (x - x_0)^2 \, (t) \right\rangle = \int (x - x_0)^2 \, G_0(x, t; x_0) dx \ , \tag{2}$$

where G_0 denotes the fundamental solution to a subdiffusion equation for a spatially unlimited system with the initial condition

$$G_0(x, 0; x_0) = \delta(x - x_0) \ , \tag{3}$$

$\delta(x)$ denotes the Dirac–delta function and x_0 is the initial position of a random walker. Subdiffusion can occur in media with a complex internal structure as, for example, media with a fractal geometry, porous media, gels, or biological membranes [1–5]. If a random walker moves in a medium in which, for example, narrow tubes are present, its movement can be hindered so much that the mean waiting time of random walker to take its next step is infinity, whereas the length of steps has finite moments [1–5].

Subdiffusion can be described in many different ways using, for example, the fractional Brownian motion, the continuous time random walk (CTRW) formalism which provides a linear subdiffusion equation with a fractional time derivative [1], the generalized Langevin equation, the generalized master equation, or nonlinear subdiffusion equations with ordinary derivatives obtained on the base of nonadditive entropies [6–11, 18]. All of these models provide the relations (1). Another problem concerns the physical interpretation of subdiffusive models. Some of them have a stochastic interpretation. In other cases physical interpretation can be unclear or unknown. For example, in the case of the model based on the CTRW formalism, the stochastic interpretation of subdiffusion process is satisfactorily simple, whereas in the case of the model based on nonadditive entropies, the stochastic interpretation is not so obvious (at least in our opinion).

Below, we will present two models describing subdiffusive processes. One of them will be a fractional model derived from the CTRW formalism and the second one will be a Sharma–Mittal model based on the Sharma–Mittal nonadditive entropy. We will assume the agreement conditions between these models in such a way that the fundamental solutions to the subdiffusion equations for both models provide the relation (1) and the first passage time (FPT) distributions will be equal. We will find dependencies between the fractional model parameters and the Sharma–Mittal ones for which these model will be similar.

2 Sharma–Mittal Entropy

Sharma–Mittal entropy is defined as [6]

$$S[P] = \frac{1-\left(\int P^r dx\right)^{(q-1)/(r-1)}}{q-1}\ , \tag{4}$$

where $q, r > 0$, $q, r \neq 1$ and P is the probability density function of finding a random walker at point x at time t.

Sharma–Mittal entropy is one of nonadditive entropies. Entropy S can be called the nonadditive entropy when satisfies the following relation for two statistically independent systems A and B and for $q \neq 1$

$$S(A+B) = S(A) + S(B) + (1-q)S(A)S(B)\ . \tag{5}$$

Nonlinear Fokker–Planck equation for the nonadditive entropies can be obtained in the following way. Firstly, a subdiffusion flux is assumed to be a form [6]

$$J = QM(P)\frac{\partial}{\partial x}\frac{\delta S}{\delta P}\ , \tag{6}$$

where Q is the fluctuation strength, $M(P)$ is a function of P (the form of this function should be assumed) and $\delta S/\delta P$ denotes a functional derivative of entropy with respect to the probability. After replacing the flux in the continuity equation

$$\frac{\partial J(x,t)}{\partial x} = -\frac{\partial P(x,t)}{\partial t}\ , \tag{7}$$

with the formula (6) and for $M(P) = P$, we obtain the following nonlinear Fokker–Planck equation for nonadditive entropies:

$$\frac{\partial P(x,t)}{\partial t} = -Q\frac{\partial}{\partial x}P^2\frac{\partial}{\partial x}\frac{\partial S}{\partial P}\ . \tag{8}$$

Equation (8) is a nonlinear partial differential equation with derivatives of a natural order.

For the Sharma–Mittal entropy (4) Eq. (8) takes the following form [6, 11, 12]:

$$\frac{\partial P(x,t)}{\partial t} = Q\left(\int \left[P\left(x,t\right)\right]^r dx\right)^{(q-r)/(r-1)}\frac{\partial^2\left[P(x,t)\right]^r}{\partial x^2}\ . \tag{9}$$

The fundamental solution to Eq. (9) reads [6, 12]

$$G_{0SM}(x,t;x_0) = D_{SM}(t)\left[\left\{1-\frac{C_{SM}(t)}{2}(r-1)(x-x_0)^2\right\}_+\right]^{\frac{1}{r-1}} \ , \tag{10}$$

where $r > 1/3$ and $r \neq 1$ and $\{z\}_+ = \max\{z,0\}$, the subscript SM stands for the Sharma–Mittal model and

$$D_{SM}(t) = \left[\frac{1}{2r(1+q)QK_{r,q}|z_r|^2 t}\right]^{\frac{1}{1+q}} \ , \tag{11}$$

$$C_{SM}(t) = 2(z_r D_{SM}(t))^2 \ , \tag{12}$$

$$z_r = \begin{cases} \sqrt{\frac{\pi}{r-1}}\frac{\Gamma(r/(r-1))}{\Gamma((3r-1)/(2(r-1)))} \ , & r > 1 \ , \\ \sqrt{\pi} \ , & r = 1 \ , \\ \sqrt{\frac{\pi}{1-r}}\frac{\Gamma((1+r)/2(1-r))}{\Gamma(1/(1-r))} \ , & 1/3 < r < 1 \ , \end{cases} \tag{13}$$

$$K_{r,q} = \begin{cases} \left(\frac{3r-1}{2r}\right)^{\frac{q-r}{1-r}} \ , & r \neq 1 \ , \\ \left(\sqrt{\mathrm{e}}\right)^{1-q} \ , & r = 1 \ . \end{cases} \tag{14}$$

3 The Continuous Time Random Walk Formalism

Within a framework of the continuous time random walk in which the probability density of finding a random walker at a time t at a point x, $P(x,t)$, depends on the probability density of the waiting time of the random walker to take its next step $\omega(t)$ and the probability density of jump length $\lambda(x)$. When $\omega(t)$ and $\lambda(x)$ are independent then the probability density $P(x,t)$ reads in terms of the Laplace transform $\mathcal{L}\{\omega(t)\} \equiv \hat{\omega}(s) \equiv \int_0^\infty e^{-st}f(t)dt$ and the Fourier transform $\mathcal{F}\{\lambda(x)\} \equiv \hat{\lambda}(k) \equiv \int_{-\infty}^{\infty} e^{ikx}f(x)dx$ [1]

$$\hat{P}(k,s) = \frac{1-\hat{\omega}(s)}{s}\frac{1}{1-\hat{\omega}(s)\hat{\lambda}(k)} \ . \tag{15}$$

As far as we know, an inverse transform of the above equation in the most general case has not been found yet, with the exception of a few very special cases. For this reason, Eq. (15) is usually considered within the limit of small values of s and k. For subdiffusion, it is assumed that the first moment of $\omega(t)$ (the average value) is infinite and that the Laplace transform $\hat{\omega}(s)$ has the following form for small values of s

$$\hat{\omega}(s) \cong 1 - \tau^\alpha s_\alpha \ , \tag{16}$$

where subdiffusion parameter α obeys $0 < \alpha < 1$ and τ_α is a positive parameter, whereas all the moments of the natural order of $\lambda(x)$ are finite. Assuming that $\lambda(t)$ is a symmetric function, the Fourier transform $\hat{\lambda}(k)$ has the following form for small values of k

$$\hat{\lambda}(k) \cong 1 - \sigma^2 \frac{k^2}{2} , \tag{17}$$

where σ^2 is the second moment of $\lambda(x)$. Computing the inverse Laplace and Fourier transforms leads to $\omega(t)$ being proportional to $1/t^{1+\alpha}$ (therefore $\omega(t)$ is referred to as a heavy-tailed distribution) and $\lambda(x)$ in the form of Gauss distribution.

Substituting Eqs. (16) and (17) into Eq. (15) we obtain

$$\hat{G}(k,s) = \frac{1}{s + \tilde{D}_\alpha s^{1-\alpha} k^2} , \tag{18}$$

where $\tilde{D}_\alpha = \sigma^2/\tau_\alpha$. Transforming Eq. (18) to the form

$$s\hat{P}(k,s) - 1 = -s^{1-\alpha} k^2 \tilde{D}_\alpha \hat{P}(k,s) \tag{19}$$

and using the following inverse Fourier and Laplace transforms

$$\mathcal{F}^{-1}\left\{k^2 \hat{P}(k,t)\right\} = -\frac{\partial^2 P(x,t)}{\partial x^2} , \qquad \mathcal{F}^{-1}\{1\} = \delta(x) , \tag{20}$$

and

$$\mathcal{L}^{-1}\left\{s\hat{G}(x,s;0) - \delta(x)\right\} = \frac{\partial G(x,t;0)}{\partial t} , \qquad \mathcal{L}^{-1}\left\{s^{1-\alpha}\hat{P}(x,s)\right\} = \frac{\partial^{1-\alpha} P(x,t)}{\partial t^{1-\alpha}} , \tag{21}$$

where $d^\alpha f(t)/dt^\alpha$ denotes the Riemann–Liouville fractional time derivative which is defined for $\alpha \in (0,1)$ as [13]

$$\frac{d^\alpha f(t)}{dt^\alpha} = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_0^t dt' \frac{f(t')}{(t-t')^\alpha} , \tag{22}$$

the following subdiffusion equation is obtained:

$$\frac{\partial P(x,t)}{\partial t} = \tilde{D}_\alpha \frac{\partial^{1-\alpha}}{\partial t^{1-\alpha}} \frac{\partial^2 P(x,t)}{\partial x^2} , \tag{23}$$

where

$$\tilde{D}_\alpha = \frac{\Gamma(1+\alpha) D_\alpha}{2} . \tag{24}$$

Equation (23) is a linear partial differential equation with a fractional time derivative.

The fundamental solution to Eq. (23) reads [14]

$$G_{0F}(x,t;x_0)=\frac{1}{2\sqrt{\tilde{D}_\alpha}}f_{\alpha/2-1,\alpha/2}\left(t;\frac{|x-x_0|}{\sqrt{\tilde{D}_\alpha}}\right)\,, \tag{25}$$

where the subscript F stands for the fractional model and the function $f_{\nu,\beta}(t;a)$ is defined as for $a>0$

$$f_{\nu,\beta}(t;a)=\frac{1}{t^{1+\nu}}\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(-k\beta-\nu)}\left(-\frac{a}{t^\beta}\right)^k\,. \tag{26}$$

The above function can be also expressed by the Fox H-function [15].

4 Comparison of the Models

Both of the considered models are assumed to describe subdiffusion which is defined by the relation (1) in this paper. Therefore, the first condition for a similarity between these models will be Eq. (1) which should be fulfilled by both models. As the second condition we have chosen an equality of first passage time distributions.

4.1 The First Agreement Condition

As previously mentioned, the fundamental solution for the fractional model $G_{0F}(x,t;x_0)$ (see Eq. (25)) fulfills the definition of subdiffusion (1). Now, let us determine conditions which should be fulfilled by the fundamental solution (25) in order for the Sharma–Mittal model to also satisfy the relation (1).

Function (10) provides the formula

$$\left\langle (x-x_0)^2(t)\right\rangle=\frac{2}{3r-1}\frac{1}{C_{SM}(t)}\,. \tag{27}$$

Comparing (1) with (27) and taking into account consideration presented in the paper [16] we obtain

$$q=\frac{2}{\alpha}-1\,, \tag{28}$$

$$Q=\frac{\alpha[2D_\alpha(3r-1)]^{1/\alpha}}{4rK_{r,2/\alpha-1}|z_r|^{2(1-1/\alpha)}}\,, \tag{29}$$

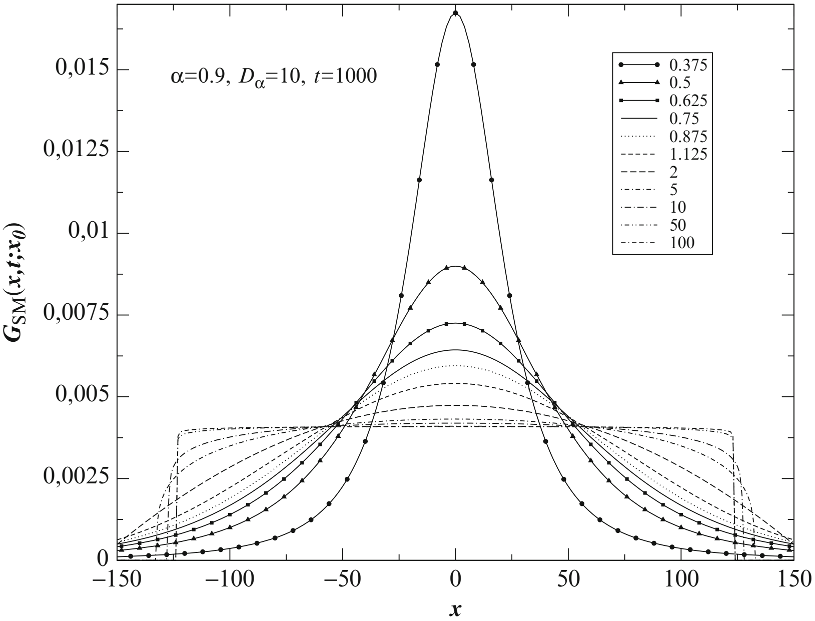


Fig. 1 The fundamental solutions for the SM model, here $x_0 = 0$. Values of the parameter r are given in the legend

where $r > 1/3$. Thus, the fundamental solution (10) fulfills the relation (1) only if its form is as follows:

$$G_{0SM}(x,t;x_0) = \frac{1}{\sqrt{2D_\alpha(3r-1)t^\alpha}|z_r|}\left[\left\{1 - \frac{(r-1)(x-x_0)^2}{2D_\alpha(3r-1)t^\alpha}\right\}_+\right]^{\frac{1}{r-1}}. \quad (30)$$

The fundamental solution (30) has different properties depending on the value of the parameter r. The example fundamental solutions are presented in Fig. 1 for different values of r. It can be noticed that the fundamental solutions for $r > 1$ have finite supports.

The fundamental solution for the fractional model, $G_{0F}(x,t;x_0)$, see Eq. (25), is controlled by the two parameters α and D_α. We have found the dependencies between the Sharma–Mittal model parameters (q and Q) and the fractional model parameters (α and D_α); see Eqs. (28) and (29), respectively. Therefore, the fundamental solution for the Sharma–Mittal model (30) is controlled by the three parameters α, D_α and r; two of them are common for both models.

4.2 The Second Agreement Condition

The first passage time is defined as the time that the random walker takes to reach a target located in x_M for the first time, from a starting point x_0. FPT is a random variable which is described by a probability density, $F(t; x_0, x_M)$. We can calculate $F(t; x_0, x_M)$ from the following formula:

$$F(t; x_0, x_M) = -\frac{dP(t; x_0, x_M)}{dt} \; , \tag{31}$$

for $t > 0$ and $F(t; x_0, x_M) = 0$ for $t \le 0$, where $P(t; x_0, x_M)$ denotes the probability of finding the particle which started from x_0 in the system with a fully absorbing wall located at x_M at time t (in the following we assume that $x_0 < x_M$)

$$P(t; x_0, x_M) = \int_{-\infty}^{x_M} G_{\mathrm{abs}}(x, t; x_0) dx \; , \tag{32}$$

$G_{\mathrm{abs}}(x, t; x_0)$ is the fundamental solution for subdiffusive system with a fully absorbing wall located at $x = x_M$. The fundamental solution $G_{\mathrm{abs}}(x, t; x_0)$ can be found through the means of the method of images and reads for $x, x_0 < x_M$

$$G_{\mathrm{abs}}(x, t; x_0) = G_0(x, t; x_0) - G_0(x, t; 2x_M - x_0) \; , \tag{33}$$

where $G_0(x, t; x_0)$ and $G_0(x, t; 2x_M - x_0)$ are the fundamental solutions for unrestricted system.

We were assumed that the second agreement condition is the equality of the FPT distributions, thus

$$F_{SM}(t; x_0, x_M) = F_F(t; x_0, x_M) \; . \tag{34}$$

Farther calculations we will make over a long time limit which we can estimate as $t \gg \max\{t_{SM}, t_F\}$, where $t_{SM} = \left[\frac{(x_M - x_0)^2}{D_\alpha |(3r-1)/(1-r)|}\right]^{1/\alpha}$ and $t_F = \left[\frac{|x_M - x_0| \Gamma(1-\alpha/2)}{\sqrt{2\tilde{D}_\alpha} \Gamma(1-\alpha)}\right]^{2/\alpha}$, respectively. Taking into account the formulae and calculations presented in the paper [16] we obtain

$$\frac{1}{\Gamma(1-\alpha/2)} \frac{1}{\sqrt{\Gamma(1+\alpha)}} = \frac{\sqrt{2}}{\sqrt{3r-1} |z_r|} \; . \tag{35}$$

The agreement condition (35) is the same for both cases of the fundamental solution for the Sharma–Mittal model ($1/3 < r < 1$ and $r > 1$) [16]. The numerical solution to Eq. (35) has a good approximation in the following form [16]:

$$r = 3.008\alpha^5 - 5.471\alpha^4 + 3.768\alpha^3 - 0.869\alpha^2 + 0.101\alpha + 0.463 \; . \tag{36}$$

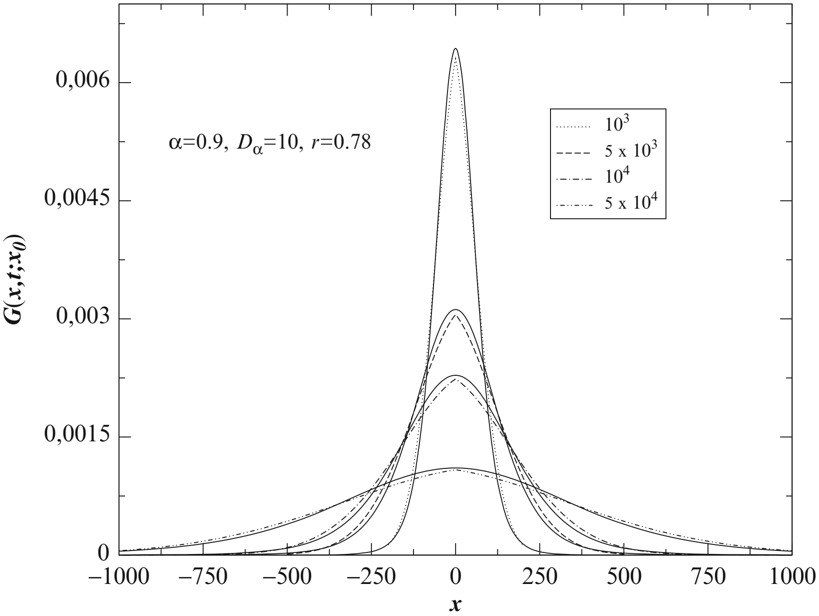


Fig. 2 Comparison between the fundamental solutions for the Sharma–Mittal model (*solid lines*) and the fractional model (*dashed lines*) for different times given in the legend

The satisfactory similarity of the fundamental solutions for the fraction model and the Sharma–Mittal model for r given by Eq. (36) can be observed in Fig. 2. The comparison between the fundamental solutions for the fractional model and the Sharma–Mittal model can only be done for $1/3 < r < 1$ because in this case the fundamental solution for the Sharma–Mittal model has a finite support like the fundamental solution for the fractional model. It should be noticed here that the fact that the fundamental solutions for both models are very similar does not stand for the equivalence of these models even though these models fulfill the relations (1) and (34).

5 Final Remarks

In this paper we have studied two models for subdiffusive processes: the Sharma–Mittal model and the fractional model. We have found the conditions under which the fundamental solutions for both models are similar, Eqs. (1) and (34). We have also determined the dependencies between the parameters of the fractional model (α and D_α) and the Sharma–Mittal model (q, Q and r), Eqs. (28), (29) and (35).

The equality of the FPT distributions over the long time limit (see Eq. (34)) gives an opportunity for extracting the parameters of the models from experimental data. The parameters of the fractional model, namely the subdiffusion parameter

α and the subdiffusion coefficient D_α, can be extracted from the experimental data by means of different methods such as a time evolution of a near-membrane layer [4, 5] or a time evolution of a reaction front in a subdiffusive system with chemical reactions [17]. The subdiffusion parameters α and D_α are the same for both models; therefore, we can calculate the values of the rest parameters by means of the relation (34).

References

1. Metzler, R., Klafter, J.: The random walk's guide to anomalous diffusion: A fractional dynamics approach. Phys. Rep. **339**, 1–77 (2000)
2. Metzler, R., Klafter, J.: The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J. Phys. A **37**, R161–R208 (2004)
3. Bouchaud, J., Georges, A.: Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. Phys. Rep. **195**, 127–293 (1990)
4. Kosztołowicz, T., Dworecki, K., Mrówczyński, S.: Measuring subdiffusion parameters. Phys. Rev. E **71**, 041105 (2005)
5. Kosztołowicz, T. , Dworecki, K., Mrówczyński, S.: How to measure subdiffusion parameters. Phys. Rev. Lett. **94**, 170602 (2005)
6. Frank, T.D.: Nonlinear Fokker–Planck Equations. Fundamental and Applications. Springer, Berlin (2005)
7. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics. Springer, NY (2009)
8. Stariolo, D.A.: The Langevin and Fokker–Planck equations in the framework of a generalized statistical mechanics. Phys. Lett. A **185**, 262–264 (1994)
9. Plastino, A.R., Plastino, A.: Non–extensive statistical mechanics and generalized Fokker–Planck equation. Physica A **222**, 347–354 (1995)
10. Carslaw, H.S., Jaeger, J.C.: Conduction of Heat in Solids. Clarendon Press, Oxford (1989)
11. Frank, T.D.: A Langevin approach for the microscopic dynamics of nonlinear Fokker–Planck equations. Physica A **301**, 52–62 (2001)
12. Frank, T.D., Daffertshofer, A. Exact time–dependent solutions of the Rényi Fokker–Planck equation and the FokkerPlanck equations related to the entropies proposed by Sharma and Mittal. Physica A **285**, 351–366 (2000)
13. Podlubny, I.: Fractional differential equations. Academic Press, San Diego (1999)
14. Kosztołowicz, T.: Phase–space diffusion in a system with a partially permeable wall. J. Phys. A **31**, 1943–1948 (1998)
15. T. Kosztołowicz, From the solutions of diffusion equation to the solutions of subdiffusive one. J. Phys. A **37**, 10779–10789 (2004)
16. Kosztołowicz, T., Lewandowska, K.D.: First passage time for subdiffusion: the nonextensive entropy approach versus the fractional model. Phys. Rev. E **86**, 021108 (2012)
17. Kosztołowicz, T., Lewandowska, K.D.: Time evolution of the reaction front in a subdiffusive system. Phys. Rev. E **78**, 066103 (2008)
18. Kosztołowicz, T., Lewandowska, K.D.: Conciliating the nonadditive entropy approach and the fractional model formulation when describing subdiffusion. Centr. Eur. J. Phys. **10**, 645–651 (2012)

Analysis of Customers' Impatience in M/M/1 Queues with Server Subject to Random Breakdowns and Exponential Vacations

Rehab F. Khalaf

Abstract In this paper we consider M/M/1queuing systems with server's vacations and random breakdowns. Customers are impatient, where customers' impatience is due to an absentee of the server upon arrival. This absentee is because either the server is on vacation or it is under repair. The mean number of customers in the system and the total waiting time the customer spends in the system have been derived in this work as a very common and important performance measures.

Keywords M/M/1 queue • Server vacation • Random breakdowns • Customers' impatience

1 Introduction

Queuing systems with server interruptions can be used in modeling numerous real-world queuing situations have arisen in systems such as manufacturing systems, communication systems, and production-inventory systems. There is now a growing interest in the analysis of queuing systems with impatient customers. This is due to the potential application of such systems in many related areas (cf. [1]).

Queuing systems with server vacations and/or random system breakdowns have been studied by numerous researchers, for instance, [2–4]. In their work [5–8] investigated a batch arrival queuing system with a Bernoulli scheduled vacation and random system breakdowns. In addition, [9] studied a queuing system with four different main server interruptions and a standby that ever replaces the main server during any potential stop.

R.F. Khalaf (✉)
Department of Computer Science, Faculty of Mathematics and Computer Science, University of M'sila, M'sila, Algeria
e-mail: gasmi_a@yahoo.fr

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_24,

The impatience phenomenon has been studied under various assumptions by many authors. The pioneering work of [10, 11] seems to be the first to analyze queuing systems with impatient customers by considering the unlimited buffer M/M/c queue and assuming that each individual customer stays in the queue as long as his waiting time does not exceed an exponentially distributed impatience time. Furthermore, [12] analyzed queuing systems with server vacations, where each arriving customer who finds no servers on duty because it is on vacation activates an independent random impatience timer. If a server does not show up by the time the timer expires, the customer abandons the queue.

In this work we study the queuing system where the customers' impatience is due to the absence of servers upon arrival. This absence of the server is either because of the vacation or because of random breakdowns. If an arriving customer sees no server present in the system, he/she may abandon the queue if no server shows up within some time. Such a model, representing frequent behavior by waiting customers in service systems, has not been treated before in the literature.

2 The Model

The underlying process is a M/M/1 queue with multiple server ([12]). The customers arrive to the system according to a Poisson distribution with arrival rate λ. Service times are exponentially distributed with mean service rate. Also the vacation times are exponentially distributed with parameter γ. The system may breakdown at random, and breakdowns are assumed to occur according to a Poisson stream with mean breakdown rate $\alpha > 0$. Once the system breaks down, the required repairs start immediately. The duration of repairs follows an exponential distribution with repair rate $\theta > 0$.

We consider the impatience of a customer by noting that whenever a customer arrives to the system and realizes that the server is on vacation or under repair, he/she activates an "impatience timer" T, T which is exponentially distributed with parameter ξ. The impatience time T is independent of the queue size at that moment. If the server returns from his vacation before the time T expires (and starts providing service), the customer stays in the system until his/her service is completed. However, if T expires while the server is still on vacation or still under repairs, the customer leaves the queue, never to return.

3 Balance Equations

We start by letting L denote the total number of customers in the system, and we let J denote the number of working servers. When the server is on vacation or under repair, this means that $J = 0$, while $J = 1$ implies that the server is active. The pair (J, L) defines a continuous-time Markov process with transition-rate diagram

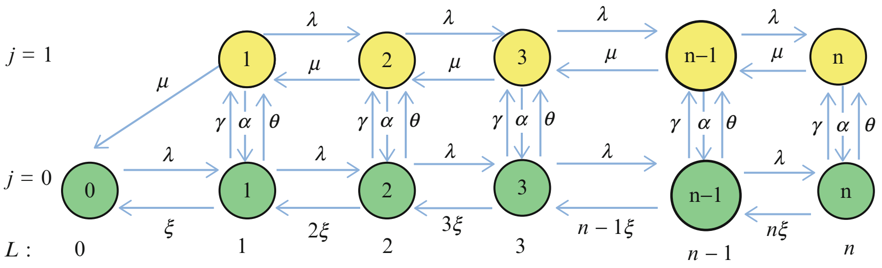


Fig. 1

as illustrated in Figure 1. Let $Pjn = P\{J = j, L = n\}(j = 0, 1; n = 0, 1, 2, \ldots)$ denote the (steady state) system state probabilities. The set of balance equations is given as

$$j = 0 \begin{cases} n = 0 & \lambda P_{00} = \mu P_{11} + \xi P_{01} \\ n \geqslant 1 & (\lambda + n\xi + \gamma + \theta) \, P_{0n} = (n + 1) \, \xi P_{0n+1} + \lambda P_{0n-1} + \alpha P_{1n} \end{cases} \tag{1}$$

$$j = 1 \begin{cases} n = 1 & (\lambda + \mu + \alpha) \, P_{11} = \mu P_{12} + (\gamma + \theta) \, P_{01} \\ n \geqslant 2 & (\lambda + \mu + \alpha) \, P_{1n} = \mu P_{1n+1} + \lambda P_{1n-1} + (\gamma + \theta) \, P_{0n} \end{cases} \tag{2}$$

Define the probability generating functions (PGFs)

$$G_0(z) = \sum_{n=0}^{\infty} P_{0n} z^n, \qquad G_1(z) = \sum_{n=1}^{\infty} P_{1n} z^n.$$

Then by multiplying each equation for n in equation (2) by z^n , summing over n and rearranging terms, we get

$$G_1(z) = \frac{z(\gamma + \theta)G_0(z) - z((\gamma + \theta)P_{00} + \mu P_{11})}{((1 - z)(z\lambda - \mu) + z\alpha)} \tag{3}$$

In a similar manner from equation (1) we obtain

$$(1 - z)\xi G_0'(z) = (\lambda - \lambda z + \gamma + \theta)G_0(z) - ((\gamma + \theta)P_{00} + \mu P_{11}) - \alpha G_1(z), \tag{4}$$

where $G_0'(z) = \frac{d}{dz} G_0(z)$. In the next section we will derive the mean number of customers in the system when the server works and the mean number of costumers when the server does not work.

4 Derivation of $E(L_0)$ and $E(L_1)$

We begin by setting $A = (\gamma + \theta)P_{00} + \mu P_{11}$. The probability that the server is working is defined by

$$G_1(1) = \sum_{n=1}^{\infty} P_{1n} = P_1$$

From equation (3) we obtain

$$P_{1\cdot} = \frac{(\gamma + \theta)P_{0\cdot} - A}{\alpha} \tag{5}$$

Clearly the probability that the server does not work (on vacation or under repair) is

$$G_0(1) = \sum_{n=1}^{\infty} P_{0n} = P_{0\cdot}$$

Obviously $P_{0\cdot} + P_{1\cdot} = 1$, so from equation (5) we get

$$P_{1\cdot} = \frac{(\gamma + \theta) - A}{(\alpha + \gamma + \theta)} \tag{6}$$

Then

$$P_{0\cdot} = \frac{\alpha + A}{(\alpha + \gamma + \theta)}. \tag{7}$$

The mean number of customers in the system when the server does not work (on vacation or under repair) is given by $E(L_0) = G_0'(1) = \sum_{n=0}^{\infty} nP_{0n}$. So using L'Hospital's rule on equation (4), we get

$$E(L_0) = \lim_{z=1} G_0'(z) = \lim_{z=1} \frac{-\lambda G_0(z) + (\gamma + \theta)G_0'(z) - \alpha G_1'(z)}{-\xi}$$

Implying that

$$E(L_0) = \frac{\lambda P_{0\cdot} + \alpha E(L_1)}{(\xi + \gamma + \theta)} \tag{8}$$

To find the mean number of customers in the system when the server works $E(L_1) = G_0'(1)$, from equation (3), we get

$$E(L_1) = G_1'(1) = \frac{\alpha(\gamma+\theta)E(L_0) + (\lambda-\mu)((\gamma+\theta)P_{0\cdot} - A)}{\alpha^2}. \tag{9}$$

Substituting the value of $E(L_1)$ in equation (8) we get

$$E(L_0) = \frac{P_{0\cdot}(\alpha\lambda + (\lambda-\mu)(\gamma+\theta) - (\lambda-\mu)A}{\alpha\xi} \tag{10}$$

Then

$$E(L_1) = \frac{P_{0\cdot}(\alpha\lambda(\gamma+\theta) + (\xi+\gamma+\theta)(\lambda-\mu)(\gamma+\theta)) - (\xi+\gamma+\theta)(\lambda-\mu)A}{\alpha^2\xi} \tag{11}$$

If we define S to be the total sojourn time of customers in the system, measured from the moment of arrival until departure, either after completion of service or as a result of abandonment. By little's law

$$E(S) = \frac{E(L_0) + E(L_1)}{\lambda} \tag{12}$$

So using equations (10) and (11) we can get $E(S)$.

5 Conclusion

We have introduced and analyzed in this paper a new type of impatience behavior in which customers become impatient (and may leave the system) when the server goes on vacation or the server is broken down and under repair. This is in contrast with previously studied impatience behavior, which did not consider server breakdown and where customers may become impatient when the number of customers or the amount of workload queued in front of them is large. We derived explicit expressions for the PGF of the number of customers (conditioned on the server state) in the system. The closed forms to find the mean number of customers in the systems as well as the mean sojourn time are given also in this work.

References

1. Gans, N., Koole, G., Mandelbaum A.: Telephone call centres: Tutorial, review, and research prospects. Manuf. Serv. Oper. Manage. **5**, 79–141 (2003)
2. Kulkarni, V.G., Choi, B.D.: Retrial queues with server subject to breakdowns and repairs. Queueing Syst. **7**(2), 191–209 (1990)
3. Madan, K.C., Abu Al-Rub, A.Z.: On a single server queue with optional phase type server vacations based on exhaustive deterministic service and a single vacation policy. Appl. Math. Comput. **149**, 723–734 (2004)

4. Maraghi, F.A., Madan, K.C., Darby-Dowman, K.: Bernoulli schedule vacation queue with batch arrivals and random system breakdowns having general repair time distribution. Int. J. Oper. Res. **7**(2), 240–256 (2010)
5. Khalaf, R.F., Madan, K.C., Lucas, C.A.: An M[X] /G/1 queue with bernoulli schedule, general vacation times, random breakdowns, general delay times and general repair times. Appl. Math. Sci. **5**(1), 35–51 (2011a)
6. Khalaf, R.F., Madan, K.C., Lucas, C.A.: An M[X]/G/1 queue with bernoulli schedule general vacation times, general extended vacations, random breakdowns, general delay times for repairs to start and general repair times. J. Math. Res. **3**(4), (2011b) (In Press).
7. Khalaf, R.F., Madan, K.C., Lucas, C.A.: On a batch arrival queuing system equipped with a stand-by server during vacation periods or the repairs times of the main server. J. Probab. Stat. Article ID 812726 (2011c). doi:10.1155/2011/812726
8. Khalaf, R.F., Madan, K.C., Lucas, C.A.: On an M[X]/G/1 queueing system with random breakdowns, server vacations, delay times and a standby. Int. J. Oper. Res **15**(1), 30–47 (2012)
9. Khalaf, R.F., Queueing systems with four different main server's interruptions and a stand-by server. Int. J. Stat. Probabilit. Accepted. (2013)
10. Palm, C.: Methods of judging the annoyance caused by congestion. Tele **4**, 189–208 (1953)
11. Palm, C.: Research on telephone traffic carried by full availability groups. Tele **1**, 107 (1957)
12. Altman, E., Yechiali, U.: Analysis of customers impatience in queues with server vacations. Queueing Syst. **52**, 261–279 (2006)

Exchange Curve and Coverage Analysis Tools for Better Inventory Management: A Case Study

S. Bhattacharya and S. Sarkar (Mondal)

Abstract In this paper, we explore how historical demand information can be used to forecast future demand and how these forecast affect the inventory management system. Forecasting, which have a long-term perspective of operations, is typically based on demand for the goods and services it offers, compared to the cost of producing them. It is used to determine the direction of future trends by using some historical data. We develop a novel and useful yet very simple methodology known as optimal policy curve or exchange curve for planning inventory by using these forecasted demands. Using this curve, reduction of average investment in inventories in the form of cycle stocks/total stocks or number of orders per year or both as desired are done. Controlling the inventory at aggregate levels is analyzed by using a scientific analysis, known as coverage analysis. We verify the idea of exchange curve and coverage analysis using a real-life problem for which data are collected from Durgapur Steel Plant, Durgapur, India, and find its optimal ordering policy and coverage for each raw material. It is found that the results obtained from economic order quantity (EOQ) model are same as the results calculated on the basis of exchange curve.

Keywords Demand forecast • Exchange curve • Coverage analysis • Optimal ordering policy

1 Introduction

The term “inventory” refers to the stock of production that a firm is offering for sale and the components that make up the production. Control of inventory, which

S. Bhattacharya (✉) • S. Sarkar (Mondal)
Department of Mathematics, National Institute of Technology, Durgapur, India
e-mail: sandipamca@gmail.com; seemasarkarmondal@yahoo.co.in

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_25,

typically represents 45 % to 90 % of all expenses for business, is needed to ensure that the business has the right goods on hand to avoid stock-outs, to prevent spoilage, and to provide proper accounting. The aim of inventory management is to hold inventories at the lowest possible cost. Starr and Miller (1962) [1] determined trade-offs between two performance measures: (i) number of orders per year (workload) and (ii) average investment in inventory in the case of multiple items. The coverage analysis of Murdoch (1965) [2] discussed a simple and powerful approach of controlling the inventory at aggregate levels. According to Chopra and Meindl (2004) [3], forecasts of future demand are essential to the inventory manager for his decision-making process. They described several methods to forecast demand and estimate forecast accuracy. Dutta (2007) [4] suggested better policies and measures to be adopted based on scientific analysis and mathematical modeling. According to Tsou et al. (2011) [5], inventory management involves trade-offs between conflicting objectives such as cost minimization and service level maximization. They draw an exchange curve of cost and service which is useful in determining the best customer service possible for the given investment in inventory management. From this literature survey, we predict the forecasted demand from the historical demand information and see how these forecasts affect the inventory management system. We first calculate the inventory investment and the number of orders. Using these data a curve, known as exchange curve or optimal policy curve, has been plotted followed by a scientific analysis, known as coverage analysis controlling the inventory at aggregate levels.

2 Definitions and Notations

Exchange curve is a device which is most useful to the executives to generate this curve under the circumstances where it is difficult or even impossible to obtain satisfactory estimates of the relevant costs. This curve can be plotted using total inventory investment (T.I.) along one axis and total number of orders (T.O.) along the other axis perpendicular to the former and is based on the assumption of EOQ which is used for trading off cycle stock investment versus total number of orders per year. Such graph demonstrates the nonlinear relationship between increasing inventory and service level under some constraints. Moreover, it is useful for planning aggregate inventory level in an organization. If, for item i (i =1, 2, . . . , n)

Ordering cost	$= \mathrm{C}_o$
Purchase cost	$= \mathrm{C}_i$
Holding cost in percent	$= \mathrm{C}_h$
Annual demand	$= \mathrm{D}_i$
Optimal order quantity	$= \mathrm{Q}_i$

then, $\text{T.I.} = \sum_{i=1}^{n} \frac{Q_i C_i}{2} = \sum_{i=1}^{n} \frac{(\sqrt{\frac{2C_o D_i}{C_h C_i}})C_i}{2} = \sum_{i=1}^{n} \sqrt{\frac{C_o D_i C_i}{2C_h}} = \sqrt{\frac{C_o}{C_h}} \frac{1}{\sqrt{2}} \sum_{i=1}^{n} \sqrt{D_i C_i}$

and T.O. $= \sum_{i=1}^{n} \frac{D_i}{Q_i} = \sum_{i=1}^{n} \frac{D_i}{\sqrt{\frac{2C_o D_i}{C_h C_i}}} = \sum_{i=1}^{n} \sqrt{\frac{C_h D_i C_i}{2C_o}} = \sqrt{\frac{C_h}{C_o}} \frac{1}{\sqrt{2}} \sum_{i=1}^{n} \sqrt{D_i C_i}$

$$Therefore,\ \ T.I. \times T.O. = \frac{1}{2}(\sum_{i=1}^{n} \sqrt{D_i C_i})^2 \tag{1}$$

$$and\ \frac{T.I.}{T.O.} = \frac{C_o}{C_h} \tag{2}$$

The above two equations (1) and (2) lead to the development of the exchange curve which is in the shape of a hyperbola and accordingly any point on the hyperbolic curve is given by the ratio $\frac{T.I.}{T.O.}$ or $\frac{C_o}{C_h}$. Thus, the desired operating points can be found out from the curve.

Coverage analysis is a simple but powerful approach of controlling the inventory at aggregate levels. In particular this approach aims at reducing the investment on inventory for a given total number of orders per year by appropriate adjustment of the total number of orders among different groups of items. The coverage of an item is defined as the ratio of average stock level to annual usage of the item.

3 Real-Life Example

We establish the idea of exchange curve and coverage analysis using a real-life problem for which data are collected from Durgapur Steel Plant, Durgapur, India. We collect the data for four years (2009, 2010, 2011, 2012) which are shown in Tables 1 and 2. The goal is to select the most appropriate forecasting method and then use to forecast demand for the next year.

Here i represents the item and in our case i = 12 and ordering cost per order (C_o) = Rs. 10,300, 9,900, 10,700, and 12,900 and holding cost (C_h) in percentage = 9 %, 8.6 % , 7.5 %, and 7 %, i.e., 0.09 , 0.086, 0.075, and 0.07 according to the years 2012, 2011, 2010, and 2009, respectively (Table 3).

♯1. Solution for exchange curve for the year 2012

(A) Nonoptimal case:

Total Cost = Rs. (247200 + 5773.72) = Rs. 2,52,973.72 ∼ Rs. 2,52,974

(B) Optimal Case:

(i) Optimal ordering policy for EOQ model:

Optimal inventory investment = $Y_i = \sqrt{\frac{2D_i C_o C_i}{C_h}}$

Total cost = Rs. (24031.20 + 24081.24) = Rs. 48112.44 ∼ Rs. 48112

Therefore, total cost decreases by an amount of Rs. (2,52,974−48112) = Rs. 204862 or about 80.9 % in comparison to nonoptimal case (Table 4).

Table 1 Year 2012 and 2011

		D_i in MT		Unit cost (C_i) in Rs.		Orders per year		Lead time in month	
SL no.	Raw materials	2012	2011	2012	2011	2012	2011	2012	2011
1.	Coking coal	52	40	668.73	437.74	4	3	3.5	0.5
2.	Noncoking coal	15	21	298.21	497.63	2	2	0.5	1.5
3.	Calcium ammoniate	11	15	295.81	267.42	1	1	2	1.5
4.	Hot metal	47	51	564.13	650.18	3	3	4	1
5.	Iron ore	18	22	245.72	265.27	2	2	0.5	0.5
6.	Coal	14	10	158.34	126.83	1	1	1	1
7.	Silicon manganese	39	30	328.15	322.20	2	3	2.5	0.5
8.	Ferro manganese	32	35	534.51	413.14	3	2	1.5	1
9.	Aluminum	10	16	540.14	578.17	1	2	0.5	0.5
10.	Coke	21	21	553.55	427.82	2	2	1	1
11.	Limestone	20	12	116.18	176.18	2	1	0.5	0.5
12.	Dolomite	12	24	283.23	566.46	1	2	0.5	0.5

Table 2 Year 2010 and 2009

		D_i in MT		Unit cost (C_i) in Rs.		Orders per year		Lead time in month	
SL no.	Raw materials	2010	2009	2010	2009	2010	2009	2010	2009
1.	Coking coal	20	35	377.14	425.17	2	2	1	1
2.	Noncoking coal	18	20	298.21	366.15	2	1	1.5	1
3.	Calcium ammoniate	10	10	254.62	220.21	1	1	4	0.5
4.	Hot metal	30	38	420.00	508.34	2	2	0.5	1.5
5.	Iron ore	25	21	310.17	272.24	2	1	1.5	1.5
6.	Coal	12	16	175.23	182.54	1	1	2.5	5.5
7.	Silicon manganese	32	31	421.25	364.27	2	2	0.5	0.5
8.	Ferro manganese	29	36	400.10	400.00	2	2	1	1
9.	Aluminum	15	20	530.00	620.00	2	2	0.5	0.5
10.	Coke	17	15	580.23	530.15	2	1	0.5	4.5
11.	Limestone	22	23	610.00	640.00	2	1	0.5	1
12.	Dolomite	12	18	370.48	583.21	1	2	2	0.5

(ii) Restricted the total number of orders at 24

We want to determine optimal inventory X_i
to minimize total inventory investment (T.I.) $= \sum_{i=1}^{n} \frac{X_i C_i}{2}$
subject to, total number of orders (T.O.) $= \sum_{i=1}^{n} \frac{D_i}{X_i} = 24$
We form Lagrangian function L $= \sum_{i=1}^{n} \frac{X_i C_i}{2} + \lambda \left(\sum_{i=1}^{n} \frac{D_i}{X_i} - 24\right)$

The necessary conditions for optimum X_i are as follows:

$$\frac{\partial L}{\partial X_i} = 0 \Rightarrow \frac{C_i}{2} - \frac{\lambda D_i}{X_i^2} = 0 \tag{3}$$

Table 3 Calculations of total cost

SL. No.	Raw materials	Orders per year	Avg. inv. (=$\frac{D_i}{2}$) in MT	Avg. inv. invst. AI (=$\frac{D_i C_i}{2}$) in Rs	Ordering cost (orders per year×C_o) in Rs.	Holding cost (AI ×C_h) in Rs.
1.	Coking coal	4	26	17386.98	41,200	1564.83
2.	Noncoking coal	2	7.5	2236.58	20,600	201.29
3.	Calcium ammoniate	1	5.5	1626.96	10,300	146.43
4.	Hot metal	3	23.5	13257.06	30,900	1193.14
5.	Iron ore	2	9	2211.48	20,600	199.03
6.	Coal	1	7	1108.38	10,300	99.75
7.	Silicon manganese	2	19.5	6398.92	20,600	575.90
8.	Ferro manganese	3	16	8552.16	30,900	769.69
9.	Aluminum	1	5	2700.7	10,300	243.06
10.	Coke	2	10.5	5812.28	20,600	523.10
11.	Limestone	2	10	1161.80	20,600	104.56
12.	Dolomite	1	6	1699.38	10,300	152.94
Total		24		64152.68∼ 64152	247,200	5773.72

Table 4 Calculation of optimal ordering policy

SL. No.	Raw materials	Orders per year	Avg. inv. (=$\frac{D_i}{2}$) in MT	Avg. inv. invst. AI (=$\frac{D_i C_i}{2}$) in Rs	Ordering cost (orders per year×C_o) in Rs.	Holding cost (AI ×C_h) in Rs.
1.	Coking coal	89215.32	44607.66	0.39	4014.69	4014.69
2.	Noncoking coal	31997.72	15998.86	0.14	1439.90	1439.90
3.	Calcium ammoniate	27290.73	13645.36	0.12	1236.00	1228.08
4.	Hot metal	77902.40	38951.20	0.34	3505.60	3505.60
5.	Iron ore	31817.70	15908.85	0.14	1431.80	1431.80
6.	Coal	22525.36	11262.68	0.10	1030.00	1013.64
7.	Silicon manganese	54122.88	27061.44	0.24	2435.53	2435.53
8.	Ferro manganese	62569.87	31284.94	0.27	2815.64	2815.64
9.	Aluminum	35161.35	17580.67	0.15	1582.26	1582.26
10.	Coke	51582.27	25791.14	0.22	2266.00	2321.20
11.	Limestone	23061.79	11530.90	0.10	1037.78	1037.78
12.	Dolomite	27891.55	13945.77	0.12	1236.00	1255.12
Total					24031.20	24081.24

$$and\ \frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{n} \frac{D_i}{X_i} - 24 = 0 \tag{4}$$

Solving equations (3) and (4), we get, $X_i = \sqrt{\frac{2\lambda D_i}{C_i}}$ and $\lambda = \frac{(\sum_{i=1}^{n}\sqrt{D_i C_i})^2}{2\times(24)^2}$ (using Table 5).

Table 5 Calculations of average inventory investment

SL. No.	Raw materials	Optimal inv. invst. (Y$_i$) in Rs.	Avg. inv. invst. AI ($=\frac{Y_i}{2}$) in Rs.	$\sqrt{D_i C_i}$	No. of orders per year ($=\frac{D_i C_i}{Y_i}$)
1.	Coking coal	8689.97	4344.98	186.48	4.00
2.	Noncoking coal	3116.60	1558.30	66.88	1.44
3.	Calcium ammoniate	2657.13	1328.57	57.02	1.22
4.	Hot metal	7587.88	3793.94	162.83	3.49
5.	Iron ore	3098.90	1549.45	66.50	1.43
6.	Coal	2193.93	1096.96	47.08	1.01
7.	Silicon manganese	5271.86	2635.93	113.13	2.43
8.	Ferro manganese	6094.35	3047.17	130.78	2.80
9.	Aluminum	3424.63	1712.32	73.49	1.58
10.	Coke	5024.41	2512.20	107.82	2.31
11.	Limestone	2246.12	1123.06	48.20	1.03
12.	Dolomite	2716.78	1358.39	58.30	1.25
Total			26061.27∼26061	1118.51	23.99∼24

Therefore, $\mathrm{Y}_i = \mathrm{X}_i \mathrm{C}_i = \sqrt{2\lambda}\sqrt{D_i C_i} = 46.6\sqrt{D_i C_i}$

If the total number of orders per year remains at 24, then inventory investment is found to be Rs. 26,061. Hence reduction in inventory investment is Rs.(64,152−26,061) = Rs. 38,091 or about 59.4 % in comparison to nonoptimal situation.

(iii) Restricted the average inventory investment at Rs. 64,152.
In this case, to minimize total number of orders (T.O.) $= \sum_{i=1}^{n} \frac{D_i}{X_i}$
under the restriction that total inventory investment (T.I.) $= \sum_{i=1}^{n} \frac{X_i C_i}{2} = 64152$,
we follow the similar method used in **(ii)** in which the optimum number of orders is obtained as 10. Therefore, the total number of orders per year has been reduced by (24−10) = 14, i.e., about 58.3 % in comparison to nonoptimal situation.

(iv) Reducing the average inventory investment at Rs. 40,000.
To optimize total cost = $10300 \sum_{i=1}^{n} \frac{D_i}{X_i} + 0.09 \sum_{i=1}^{n} \frac{X_i C_i}{2}$
subject to, T.I. $= \sum_{i=1}^{n} \frac{X_i C_i}{2} = 40{,}000$, the optimum total cost is obtained as Rs. 322900.
Hence, the increase in total cost is Rs.(322900−252974) = Rs. 69926 or about 21.6 % in comparison to nonoptimal situation.
Similar computations have been done for different T.I. and T.O. which are shown in the following Table 6.
where * denotes the optimal value.
Using the above data, the following exchange curve is plotted.

In Fig.1, point A corresponds the current company policy. Company's current policy can be improved by considering the following two situations:

Table 6 Calculations of (T.I.)×(T.O.)

Policy	Total inventory investment (T.I.)	Total number of orders (T.O.)	(T.I.) × (T.O.)
1.	80,000*	8	6,40,000
2.	12,510	50*	6,25,500
3.	6,252	100*	6,25,200
4.	64,152	10*	6,41,520
5.	26,061*	24	6,25,464
6.	20,000*	32	6,40,000
7.	50,000	12*	6,00,000

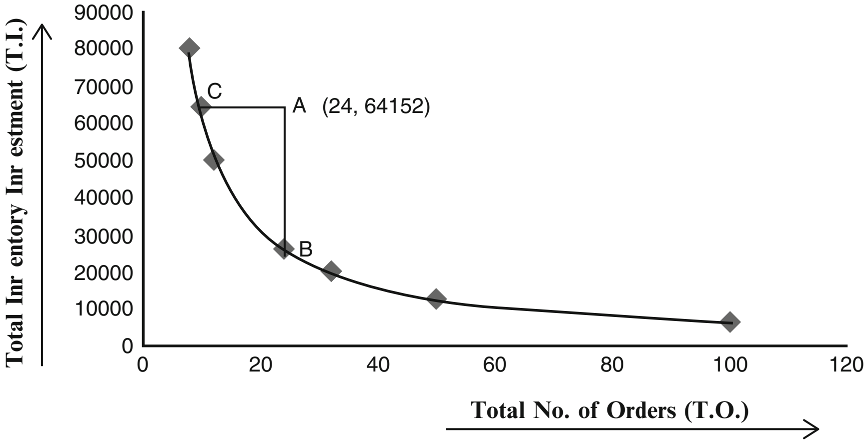


Fig. 1 Exchange curve

(a) Keeping the number of orders at 24, we can reduce the inventory investment to Rs. 26,061, which can be obtained by drawing a straight line parallel to the T.I. axis from point A. The line meets the curve at B whose coordinate is (24, 26061).

(b) Keeping the inventory fixed at Rs. 64,152, we can reduce the total number of orders to 10 which is determined by drawing a straight line parallel to the T.O. axis from point A. This line meets the curve at C whose coordinate is (10, 64152).

The result of T.I. × T.O. computed in the last column of Table 6 is compared with the result obtained using the relation T.I. × T.O.= $\frac{1}{2}(\sum_{i=1}^{n}\sqrt{D_i C_i})^2$. This gives the constant product of T.I. and T.O., i.e., T.I.× T.O. = 6,25,532.31, which is approximately equal to the computed results.

♯2. Solution for coverage analysis for the year 2012:

In order to determine coverage, we use the following notations:

Table 7 Calculations of reorder level (ROL)

SL no.	Raw materials	D_i in MT per year	$D_{im}(=\frac{D_i}{12})$ per mth in MT	Q_i per mth	Orders per mth $(=\frac{D_{im}}{Q_i})$	T $(=\frac{Q_i}{D_{im}})$	LT in mth (given)	ROL per mth in MT
1.	Coking coal	52	4.33	1.10	3.94	0.25	3.5	17.32
2.	Noncoking coal	15	1.25	0.72	1.52	0.66	0.5	0.62
3.	Calcium ammoniate	11	0.92	0.76	1.21	0.10	2	1.84
4.	Hot metal	47	3.92	1.54	2.54	0.39	4	15.68
5.	Iron ore	18	1.50	0.82	1.83	0.55	0.5	0.75
6.	Coal	14	1.17	0.84	1.39	0.72	1	1.17
7.	Silicon manganese	39	3.25	1.32	2.46	0.40	2.5	8.12
8.	Ferro manganese	32	2.67	1.02	2.62	0.38	1.5	4.00
9.	Aluminum	10	0.83	0.65	1.28	0.78	0.5	0.42
10.	Coke	21	1.75	1.02	1.72	0.58	1	1.75
11.	Limestone	20	1.67	0.83	2.01	0.50	0.5	0.84
12.	Dolomite	12	1.00	0.76	1.32	0.76	0.5	0.50

Lot size (Q) $= \sqrt{\frac{2D_i C_o}{C_h}}$

Lead time = LT Reorder level (ROL) = ($\frac{D_i}{12}$ *per* mth) ×Ã- (LT in mth.), where mth. denotes month

Cycle length = T

Buffer stock (BS) = Average Demand per month during $\sqrt{(T+LT)}$

Standard deviation of demand (σ_D) $= \sqrt{\frac{D_i^2}{12} - (\frac{D_i}{12})^2} = 14$

95 % of Service level (z) = 1.65 (by normal distribution)

Safety stock (SS) = z $\sigma_D \sqrt{(T+LT)}$ during $\sqrt{(T+LT)}$

Average inventory level (I_i) $= \frac{Q_i}{2} + \text{SS}$

Coverage (G_i) $= \frac{I_i}{D_i}$

The coverage is calculated using Tables 7 and 8.

Using Tables 7 and 8, we plot the curve of coverage analysis, taking T.O. in horizontal axis and coverage in vertical axis. Every point on this curve represents the ratio of average inventory level to the annual usage for the collected data item (Fig. 2).

Forecast for the year 2013:

We consider the 4-point moving average for each item. We make the forecast for period 5 using the last 4 periods for which we assume the current period to be t= 4. We estimate the level L_t in period 4 from the following relation using 4-point moving average.

$L_t = \frac{D_t + D_{t-1} + \ldots + D_{t-N+1}}{N}$, where N = 4 in our case and F_{t+1} (forecasting at the period t+1) = L_t and $F_{t+N} = L_{t+N-1}$.

Table 8 Calculations of coverage (G_i)

SL no.	Raw materials	Buffer stock (BS) in MT	Safety stock (SS) in MT	Avg. inv. level (I_i) in MT	Coverage (G_i)
1.	Coking coal	8.40	44.81	45.36	10.48
2.	Noncoking coal	1.34	24.72	25.13	20.10
3.	Calcium ammoniate	1.33	33.50	33.88	36.83
4.	Hot metal	8.18	48.28	49.05	12.51
5.	Iron ore	1.53	23.56	23.97	15.98
6.	Coal	1.53	30.26	30.68	26.22
7.	Silicon manganese	5.53	39.27	39.93	12.29
8.	Ferro manganese	3.66	31.65	32.16	12.04
9.	Aluminum	0.94	26.10	26.42	31.83
10.	Coke	2.20	29.10	29.61	16.92
11.	Limestone	1.67	23.10	23.52	14.08
12.	Dolomite	1.12	25.87	26.25	26.25

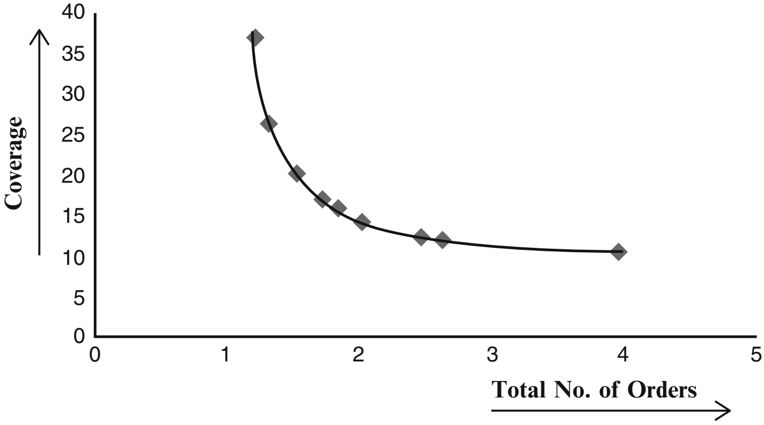


Fig. 2 Coverage curve

We obtain the level and the forecasted demand, unit cost, orders per year, and lead time per month for period 5 for each item as shown in Table 9.

In this case $D_{i4}, D_{i3}, D_{i2} and \ \ D_{i1}$ represent demand of ith item for the years 2012, 2011, 2010, and 2009, respectively. Ordering cost per order (C_o) = Rs.$\frac{(10300+9900+10700+12900)}{4}$ = Rs. 10,950 and holding cost in percentage (C_h) = $\frac{(0.09+0.086+0.075+0.07)}{4} = 0.080$.

We develop an exchange curve and coverage curve by using this forecasted data set of items and we see that the results obtained from the exchange curve and coverage curve are same as that obtained from the results calculated by the method given in 3.

Table 9 Calculations of forecasted demand (F_5), unit cost, orders per year, LT per month

SL no.	Raw materials	Estimated level L_4= $(\frac{D_{i4}+D_{i3}+D_{i2}+D_{i1}}{4})$= F_5	Unit cost (C_i) in Rs.	Orders per year	LT per month
1.	Coking coal	37	477.20	3	1.5
2.	Noncoking coal	18	365.05	2	1.1
3.	Calcium ammoniate	12	259.46	1	2
4.	Hot metal	42	535.66	2	1.8
5.	Iron ore	22	273.35	2	1
6.	Coal	13	160.74	1	2.5
7.	Silicon manganese	33	358.97	2	1
8.	Ferro manganese	33	436.94	2	1.1
9.	Aluminum	15	567.08	2	0.5
10.	Coke	18	522.94	2	1.8
11.	Limestone	19	385.59	2	0.6
12.	Dolomite	16	450.84	2	0.9

4 Discussions and Conclusions

According to Table 6, in the year 2012, the total number of orders (T.O.) = 24, at the current operating point A at which the total stock value (TS) is 64,152. If the management desires to keep T.O. at 24, he can reduce TS to 25,000 in the coming years (as shown in Fig. 1). Hence the possible reduction in TS value of inventory for the same service level is Rs. $(64,152 - 25,000)$ = Rs. 39,152. We calculate the savings for different stock investments with respect to the desired operating point B. Total savings thus calculated is Rs. 1, 52,530 and the possible reduction in TS value of inventory for the same service level is about 23.7 %, but in the year 2013 the total savings = Rs. 2, 23,119 and the possible reduction in TS value of inventory for the same service level is about 25.1 %.

The optimal policy curve shows how orders and inventory investment can be traded one for the other. The executive can quickly converge on the optimal point on the curve for the company without having had to convert his knowledge into the form of carrying costs and ordering costs. The curve shows that the average saving against each item was found to be around 20–25 % with respect to the desired operating point. This curve finds its optimal ordering policy for each raw material. Optimal policy curve or an exchange curve can be considerably more useful than simply as a geometrical analogy to an already completed mathematical argument. It shows exactly that how orders and inventory investment can be traded one for other. It is the most valuable tool in the frequently occurring and difficult cases where satisfactory estimates of the relevant costs are not available.

Coverage analysis can be measured as a scientific analysis for planning and controlling stock levels so that the inventory investment can be optimized from a financial point of view.

Our future plan is to investigate various planning models that will allow inclusive planning of coverage. The idea will also investigate the nature of the causes of coverage and how the appropriate measurement and control may be applied. We can extend our knowledge for the optimal policy for any possible combination of values of ordering cost (C_o) and holding cost (C_h) for a given set of data items.

References

1. Starr, M.K., Miller, D.W.: Inventory Control: Theory and Practice. Prentice Hall, Englewood Cliffs, NJ (1962)
2. Murdoch, J.: Coverage Analysis New Technique for Optimizing the Stock Ordering Policy. Proceeding from one day conference held at Cranfield, UK (1965)
3. Chopra, S., Meindl, P.: Supply Chain Management Strategy, Planning and Operation. Prentice-Hall of India, Delhi (2004)
4. Dutta, S.K.: Better inventory management through exchange curves. The Icfai J. Oper. Manage. **VI**(3), 18 (2007)
5. Tsou, C.S. et al.: Estimating exchange curve For inventory management through evolutionary multi-objective Optimization. Afr. J. Bus. Manage. **5**(12), 4847–4852 (2011)

Modelling Combustion Process in Circulating Fluidised Bed Boiler: A Fuzzy Graph Approach

Razidah Ismail, Sumarni Abu Bakar, and Nurul Syazwani Yusof

Abstract Integration of fuzzy graph and autocatalytic set theory plays an important role in the emergence of a new concept of fuzzy autocatalytic set (FACS). This concept was successfully used to describe the chemical reactions in the clinical waste incineration process. Therefore, this paper aimed to extend the application of FACS in modelling combustion process in circulating fluidised bed boiler (CFB). Fifteen important chemical substances known as species are represented as nodes, V, and its catalytic relationships between nodes are represented by the edges, E, in the graph $G_F(V, E)$. The membership value of fuzzy edge connectivity between two nodes in the graph is calculated using material chart balance based on simulated data. The fuzzy graphical model of the combustion process provides more information on the strength of connection of its catalytic relationship between species. Analysis of dynamics in the combustion process gives reasonable and equitable results in terms of sequence of depleting species over time and the end products as compared to the real process. Some characteristics related to the graph dynamics of the combustion process in CFB are also highlighted.

Keywords Autocatalytic set • Directed graph • Fuzzy modelling • Perron–Frobenius eigenvalue

1 Introduction

Circulating fluidised bed boiler (CFB) is a device used to generate high-pressure steam which is then used to spin turbine in a steam turbine to produce electricity. The steam is generated by burning fossil fuels, namely coal and limestone, in a

R. Ismail (✉) • S. Abu Bakar • N.S. Yusof
Faculty of Computer and Mathematical Sciences, Universiti Teknologi
MARA (UiTM), Shah Alam, Selangor, 40450, Malaysia
e-mail: razidah@tmsk.uitm.edu.my; sumarni@tmsk.uitm.edu.my

A.R. Ansari (ed.), *Advances in Applied Mathematics*, Springer Proceedings in Mathematics & Statistics 87, DOI 10.1007/978-3-319-06923-4_26,

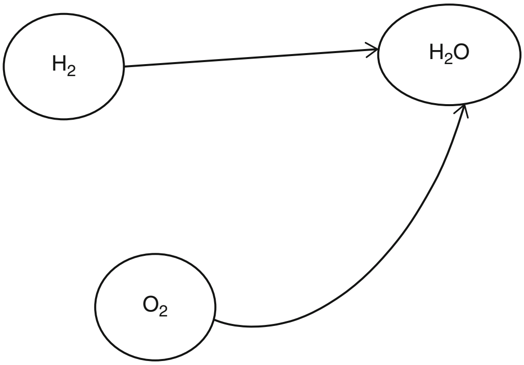


Fig. 1 Relationship of H_2, O_2 and H_2O as a graph

furnace which operates under a special hydrodynamic condition [1, 2]. The burning or combustion process involves complex interactions of chemical substances that exist during the process. Due to this reason, not many researchers had attempted to model the combustion process in CFB [3]. Recently, a crisp graphical model, $G(V, E)$, and its modification, $G_M(V, E)$, which is based on graph theoretical approach were developed [4–6]. In these models, V denotes a set of chemical substances or species that has significant role in the combustion process and E is a set of links or edges which described interactions among the species according to autocatalytic set (ACS) concept. The species are then represented as node in the graph and edge between two nodes is constructed if there is a catalytic relationship among the nodes. For example, H_2, O_2 and H_2O are identified as species for reaction equation of $H_2 + O_2 \rightarrow H_2O$. The formations of species H_2O are being catalysis by H_2 and O_2. Thus the graph representing these relationships is depicted in Fig. 1. In graph theoretic perspective, ACS is a subgraph where each of whose nodes has at least one incoming link from a node belonging to the same subgraph [7, 8]. A link from vertex j to vertex i indicates that species j catalyses the production of species i. Figure 2 illustrates the crisp graphical ACS model with 15 species and 46 edges.

However dynamical nature of the combustion process is not well explained from both of these models. For example, end product of the combustion process which is determined through Graph Dynamical Algorithm [4] reveals a similar output, namely carbon dioxide (CO_2) and carbon monoxide (CO). The value of 0 or 1 which was given to the link of a crisp graph could have contributed to such results. This has motivated further exploration to modify the model using fuzzy graph approach specifically fuzzy autocatalytic set (FACS) of fuzzy graph Type-3.

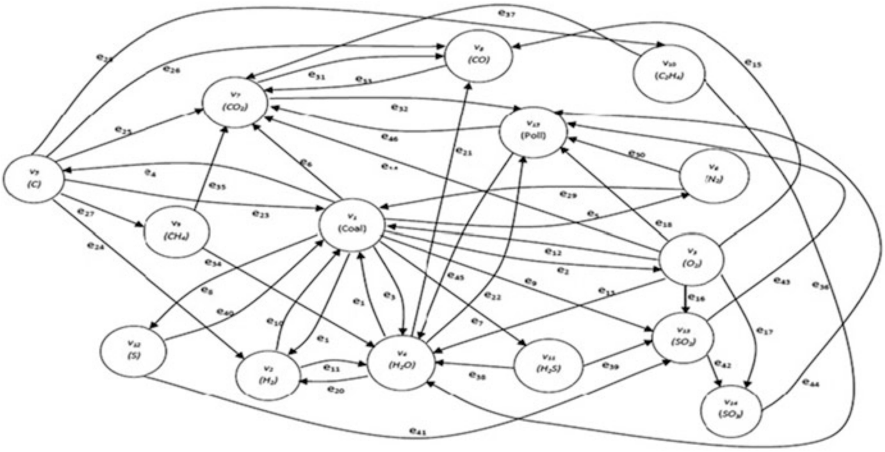


Fig. 2 Graphical ACS model of combustion process in CFB

2 Fuzzy Autocatalytic Set

FACS is a concept emerged from integration of autocatalytic set (ACS) and fuzzy graph theory. A clinical waste incineration process in Malacca, Malaysia, was successfully modelled using FACS of fuzzy graph Type-3 [9] whereby its theoretical foundation was first laid in [10]. Meanwhile, fuzzy graph of Type-3 involved crisp vertex and edge, but the edge has fuzzy connectivity in which it involved fuzzy heads and fuzzy tails [11]. The formal definition of FACS and fuzzy edge connectivity are stated in Definitions 1 and 2 respectively. The membership value that constitutes to the entry of adjacency matrix of fuzzy graph is stated in Definition 3.

Definition 1 **([10]).** FACS was defined as a subgraph where each of whose nodes has at least one incoming link with membership value $\mu(e_i) \in (0, 1), \forall e_i \in E$.

Definition 2 **([10]).** The fuzzy head $h(e_i)$ and the fuzzy tail $t(e_i)$ are functions of e_i such that $h : E \to [0, 1]$ and $t : E \to [0, 1]$ for $e_i \in E$. Fuzzy edge connectivity is a tuple $(t(e_i), h(e_i))$ and the set of all fuzzy edge connectivity is denoted as $C = \{(t(e_i), h(e_i)) : e_i \in E\}$. The membership value of fuzzy edge connectivity is denoted as $\mu(e_i) = \min\{t(e_i), h(e_i)\}$.

Definition 3 **([9]).** The entries for adjacency matrix of FACS of fuzzy graph Type-3; $C_{F_{ij}}$, is

$$C_{F_{ij}} = \begin{cases} 0 & \text{for } i = j \text{ and } e_i \notin E \\ \mu(e_i) \in (0, 1] & \text{for } i \neq j \text{ and } e_i \in E \end{cases}. \tag{1}$$

These definitions are referred to in the analysis of the proposed graphical FACS model of the combustion process in CFB.

3 Determination of Membership Value of Fuzzy Edge Connectivity

Aforementioned, Fig. 2 is a crisp graphical model $G_M(V, E)$ with 15 chemical substances or species and 46 edges. Thus, $V = \{v_1, v_2, \ldots, v_{15}\}$ is the set of nodes for the graph where v_1 = coal, v_2 = hydrogen (H_2), v_3 = oxygen (O_2), v_4 = water (H_2O), v_5 = carbon (C), v_6 = nitrogen (N_2), v_7 = carbon dioxide (CO_2), v_8 = carbon monoxide (CO), v_9 = methane (CH_4), v_{10} = ethylene (C_2H_4), v_{11} = hydrogen sulphide (H_2S), v_{12} = sulphur (S), v_{13} = sulphur dioxide (SO_2), v_{14} = sulphur trioxide (SO_3) and v_{15} = pollution. Based on the catalytic relationship between the species, 46 edges are identified which are denoted by set E where $E = \{e_i = (v_r, v_k); r, k = 1, 2, \ldots, 15$ and $r \neq k\}$ for $i = 1, \ldots, 46$.

The rationale made for every edge of this particular graph is published in [5]. Calculation of the fuzzy membership value for each edge in Fig. 2 is based on the physical measurement method [9]. This method is used in determining fuzzy value in many applications of fuzzy logic and is commonly used in chemical industry and engineering where huge amount of raw information are obtained from experiments. Here, the membership value for every edge in the graph is calculated using material chart balance based on 100 kg of coal [12] as in Table 1.

The weight (in kg) in column input of Table 1 is the product of mol. vol and its relative atomic mass (RAM) and relative molecular mass (RMM). Next, membership value for each edge denoted by $\mu(e_i)$ for $i = 1, 2, \ldots, 46$ is determined through percentage of composition of coal, assumptions, ratio and calculation of material chart balance. The derivation of fuzzy membership value for edges leaving node v_1 is shown in Table 2, whereas the explanation for each edge for the rest of these nodes in the set of V can be found in [6].

For edges e_1, e_2, e_4 and e_8 the calculation of membership values are based on the total input weight of solid fuel (coal) and air which is 1,073.883 kg. Thus, the

Table 1 Material chart balance for 100 kg of coal

Input				Output			
	Mol.vol	RAM	Weight (kg)	Flue gas	Mol.vol	RMM	Weight (kg)
(A) Fuel							
C	5.916	12	70.99	CO_2	5.916	44	260.304
H_2	2.5312	2	5.062	H_2O	2.75	18	49.5
H_2O	0.2188	18	3.938	SO_2	0.0375	64	2.4
S	0.0446	32	1.2	N_2	27.2028	28	761.678
(B) Air							
O_2	7.2191	32	231.0112				
N_2	7.2191	28	760.4311				
1,073.883				1,073.882			

Table 2 Fuzzy membership value for edges leaving node v_1

Input of $H_2 = 5.062$ kg (thus, $\mu(e_1) = 0.0047$)	Output of $H_2O = 49.5$ kg (thus, $\mu(e_3) = 0.0461$)
Input of $O_2 = 231.0112$ kg (thus, $\mu(e_2) = 0.2151$)	Output of $N_2 = 761.678$ kg (thus, $\mu(e_5) = 0.7093$)
Input of $C = 70.99$ kg (thus, $\mu(e_4) = 0.0661$)	Output of $CO_2 = 260.304$ kg (thus, $\mu(e_6) = 0.2424$)
Input of $S = 1.2$ kg (thus, $\mu(e_8) = 0.0011$)	Output of $SO_2 = 2.4$ kg (thus, $\mu(e_9) = 0.0022$)

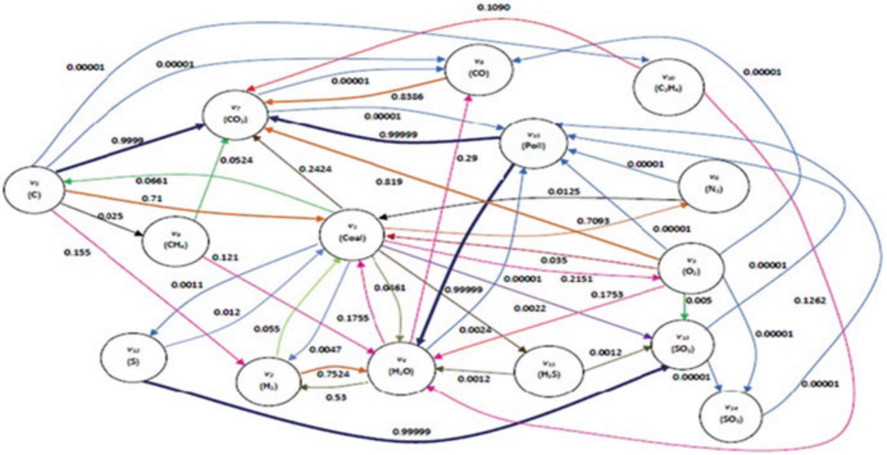


Fig. 3 Graphical FACS model of combustion process in CFB

membership value for e_1, e_2, e_4 and e_8 is calculated using the ratio of the weight of chemical substance (species), namely H_2, O_2, C and S with respect to total input weight. For edges e_3, e_5, e_6 and e_9, the fuzzy membership values are based on the ratio of the output weight of flue gas, namely H_2O, N_2, CO_2 and SO_2 with total product of flue gas which is 1,073.882 kg.

Now, every edge in $G_M(V, E)$ has various strengths of connection which is determined by membership values. The greater the membership value, the stronger is the connection between two species in the graph. The new model, $G_F(V, E)$, is shown in Fig. 3, where different thicknesses of the link are used to differentiate the membership value that represents the strength of connection between two vertices.

Thus, the graphical FACS model is represented as $G_F(V, E)$ where $V = \{v_1, v_2, \ldots, v_{15}\}$ denotes the species involved in the combustion process and $E = \{\mu(e_1), \mu(e_2), \ldots, \mu(e_{46})\}$ denotes the membership value of fuzzy edge connectivity between the species.

4 Graph Dynamics of $G_F(V, E)$

Subsequently, the adjacency matrix of $G_F(V, E)$ is obtained as shown in matrix C_F. The entry of adjacency matrix is a fuzzy membership value which represents strength of connection between two species, whereas in $G_M(V, E)$, the entry of value 1 of its adjacency matrix is only representing the existence of connection between two species [6]. This adjacency matrix C_F is a type of non-negative matrix which can be related to Perron–Frobenius Theorem [13].The graphical model $G_F(V, E)$ obtained is a static graph. The actual dynamical nature of the combustion process can only be investigated if an assumption is made to the species. In this study, assumption is made where all the species are thought of as "living" on the nodes of the graph while undergoing evolution process after certain time t. Real combustion process in CFB is dynamic in nature.

$$C_F = \begin{bmatrix}
0 & 0.0055 & 0.035 & 0.1755 & 0.71 & 0.0125 & 0 & 0 & 0 & 0 & 0 & 0.012 & 0 & 0 & 0 \\
0.0047 & 0 & 0 & 0.53 & 0.155 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2151 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0.0461 & 0.7524 & 0.1753 & 0 & 0 & 0 & 0 & 0 & 0.1214 & 0.1262 & 0.0012 & 0 & 0 & 0 & 0.99999 \\
0.0661 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.7093 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2424 & 0 & 0.8195 & 0 & 0.99999 & 0 & 0 & 0.8386 & 0.0524 & 0.1090 & 0 & 0 & 0 & 0 & 0.99999 \\
0 & 0 & 0.00001 & 0.29 & 0.00001 & 0 & 0.00001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.025 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.00001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0024 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0011 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0022 & 0 & 0.00519 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0012 & 0.99999 & 0 & 0 & 0 \\
0 & 0 & 0.00001 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.00001 & 0 & 0 \\
0 & 0 & 0.00001 & 0.00001 & 0 & 0.00001 & 0.00001 & 0 & 0 & 0 & 0 & 0 & 0.00001 & 0.00001 & 0
\end{bmatrix}.$$

This situation leads to further investigation on the evolution process of the graph in a long run. In order to explore this process, graph dynamics of $G_F(V, E)$ after certain time t is presented, based on Perron–Frobenius eigenvalue (PFE) of the adjacency matrix and the Graph Dynamic Algorithm [4].

Table 3 shows the comparison of graph dynamics for $G_M(V, E)$ and $G_F(V, E)$ in terms of end product of the combustion process and the sequence of depleted species. The end product of combustion for $G(V, E)$ is carbon dioxide (CO_2) and carbon monoxide (CO) while the end product for $G_F(V, E)$ is hydrogen (H_2) and water (H_2O). According to Bruner [14], the by-products of any combustion process include CO_2, water and pollution. Here, both graphs give significant by-product. In the formation of water, hydrogen (H_2) is needed while carbon monoxide (CO) is a type of flue gas which is considered as pollution. The sequence of depleted species for $G_M(V, E)$ and $G_F(V, E)$ is different at certain time t. For example, in sequence of depleted species shown in Table 3, oxygen (O_2) is depleted at $n = 6$ in $G_M(V, E)$, but it is depleted when $n = 9$ in $G_F(V, E)$.

This result is reasonable to graph dynamics of $G_F(V, E)$ as compared to $G_M(V, E)$ since oxygen (O_2) is needed to complete the combustion. Therefore, the result of $G_F(V, E)$ is equitable to describe the real process in CFB.

Table 3 Comparison of graph dynamics between $G_M(V, E)$ and $G_F(V, E)$

	$G_M(V, E)$	$G_F(V, E)$
Sequence of depleted species	1. Methane (CH_4)	1.Ethylene (C_2H_4)
	2. Ethylene (C_2H_4)	2. Sulphur trioxide (SO_3)
	3. Sulphur trioxide (SO_3)	3. Pollution
	4. Carbon (C)	4. Methane (CH_4)
	5. Sulphur (S)	5. Sulphur (S)
	6. Oxygen (O_2)	6. Hydrogen sulphide (H_2S)
	7. Hydrogen sulphide (H_2S)	7. Sulphur dioxide (SO_2)
	8. Nitrogen (N_2)	8. Carbon (C)
	9. Sulphur dioxide (SO_2)	9. Oxygen (O_2)
	10. Coal	10. Coal
	11. Hydrogen (H_2)	11. Nitrogen (N_2)
	12. Water (H_2O)	12. Carbon monoxide (CO)
	13. Pollution	13. Carbon dioxide (CO_2)
End product of combustion	Carbon dioxide (CO_2) and Carbon monoxide (CO)	Water (H_2O) and Hydrogen (H_2)

5 Conclusion

The crisp graph of the combustion process in CFB has been refined to graphical FACS model, $G_F(V, E)$, using the fuzzy graph approach. In this model, all the edges linking the vertices or species have fuzzy membership value between 0 and 1. Analysis on the dynamical behaviour of combustion process by using Graph Dynamic Algorithm through graph updates reveals that the end product of the combustion is water (H_2O) and hydrogen (H_2). Since water (H_2O) is part of by-product of any combustion, therefore this result is also reasonable to describe the real process in CFB, but in terms of sequence of depleted species, $G_F(V, E)$ gives significant result as compared to $G_M(V, E)$. Even though the membership value of fuzzy edge connectivity between two nodes in the graph is calculated using material chart balance based on simulated data since real data is almost impossible to find, the graph $G_F(V, E)$ is found to be more realistic to describe the real combustion process in CFB as compared to $G_M(V, E)$. It is hoped that the fuzzy graph approach utilised in this work will provide more opportunities for future research especially in applying FACS to optimisation problem in other complex systems such as optimisation of human resources which could be related to enhance the performance of an organisation or in optimisation of variables identified in incineration-related industries for reducing air pollution and energy saving. It could be possibly applied in the identification of important materials (waste) to be recycled in recycling industries so that it could eventually reduce land waste (garbage and trash).

Acknowledgements This work was supported financially by UiTM Fundamental Research Grant Scheme (FRGS). The authors would like to thank the Research Management Institute, Universiti Teknologi MARA, Malaysia, for the financial support through the Fundamental Research Grant Scheme, ref: 600-RMI/FRGS 5/3 (8/2013). Their gratitude is also extended to anonymous reviewers for their valuable comments and suggestions.

References

1. Basu, P., Fraser, S.A.: Circulating Fluidized Bed Boilers. Reed Publishing, New York (1991)
2. Basu, P.: Combustion of coal in circulating fluidized bed boilers: a review. Chem. Eng. Sci. **54**, 5547–5557 (1999)
3. Huilin, L., Guangbo, Z., Rushan, B., Yangjin, C., Gidapow, D.: A coal combustion model for circulationg fluidized bed boilers. Fuel, **79**(2), 165–172 (2000)
4. Sumarni, A.B., Noor Ainy, H., Fatin, H.O., Shafawati, I., Fudzla, A.M.: Autocatalytic set of chemical reactions of circulating fluidized bed boiler. In: IEEE Proceedings of International Conference on System Engineering and Technology, 978-1-4673-2374-1/12/2012 (2012)
5. Sumarni, A.B., Ismail, R., Noor Ainy, H.: Graph dynamics representation of chemical reaction of a boiler. In: IEEE Business Engineering and Industrial Applications Colloquium, pp. 906–910 (2013)
6. Nurul Syazwani, Y., Sumarni, A.B., Ismail, R.: Modified graphical autocatalytic set model of combustion process in circulating fluidized bed boiler. In: Proceedings of 21st National Symposium of Mathematical Sciences, Penang, Malaysia, 6–8 Nov 2013
7. Jain, S., Krishna, S.: Autocatalytic sets and the growth of complexity in an evolutionary model. http://www.arxiv.org/abs/nlin.AO/9809003 (1998). Accessed 24 Dec 2011
8. Jain, S., Krishna, S.: Graph theory and the evolution of autocatalytic networks. In: Bornholdt, S., Schuster, H.G. (eds.) Handbook of Graphs and Network: From the Genome to the Internet, pp. 355–395. Wiley-VCH Verlag GmbH and Co., Weinheim (2003)
9. Sabariah, B., Tahir, A., Rashid, M.: Fuzzy edge connectivity relates the variable in clinical waste incineration process. Matematika **25**(1), 31–38 (2009)
10. Tahir, A., Sabariah, B., Anuar, K.K.: Modeling a clinical incineration process using fuzzy autocatalytic set. J. Math. Chem. **47**(4), 1263–1273 (2010)
11. Blue, M., Bush, B., Puckett, J.: Unified approach to Fuzzy graph problems. Fuzzy Sets Syst. **125**, 355–368 (2002)
12. Chattopadhyay, P.: Boiler Operation Engineering: Question and Answer, 2nd edn. Tata Mc Graw Hill, New York (2000)
13. Seneta, E.: Non-Negative Matrices: An Introduction to Theory and Applications. George Allen and Unwin Ltd., London (1973)
14. Bruner, C.R.: Handbook of Incineration System. McGraw-Hill, New York (1991)