Proceedings of the
14th International Conference on
Computational and Mathematical Methods
in Science and Engineering
Costa Ballena, Rota, Cádiz (Spain)
July 3rd-7th, 2014

CMMSE 2014

VOLUME II
Editor: J. Vigo-Aguiar

Associate Editors
I. P. Hamilton, J. Medina, P. Schwerdtfeger, W. Sprößig,
M. Demiralp, E. Venturino, V.V. Kozlov, P. Oliveira
Proceedings of the 2014 International Conference on Computational and Mathematical Methods in Science and Engineering

Costa Ballena (Rota), Cádiz, Spain

July 3-7, 2014

Editor:
J. Vigo-Aguiar

Associate Editors
I. P. Hamilton, J. Medina, P. Schwerdtfeger, W. Sprößig, M. Demiralp, E. Venturino, V.V. Kozlov, P. Oliveira
Preface

One year more, it is our great pleasure to present the proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE 2014), at Rota, Cádiz (Spain), July 3rd-7th, 2014. These proceedings, comprised of the extended abstracts and the papers accepted to the conference, are of significant interest and contain original and substantial analyses of computational and mathematical methodologies. The proceedings have five volumes, the first four correspond to the articles typeset in LaTeX and the fifth to articles typeset in Word.

CMMSE 2014 continues with the same philosophy of being a great forum where researchers from several disciplines of applied mathematics discuss about the new advances and open problems. We hope that during the session the usual and desirable exchange of ideas, comments and suggestions leading to the improvement and deepening of the papers presented to allow further development of the research occurs. We also hope that the developed activity narrows and renews the links between participants.

This year we have achieved a new record in the number of symposiums and the quality of the accepted papers is also very high. The first one, high-performance computing, considers new large-scale problems that arise in fields like bioinformatics, computational chemistry, and astrophysics. The second symposium address analytical, numerical and computational aspects of partial differential equations in life and materials science. Computational finance is a session focuses on solving problems related to asset pricing, trading and risk analysis of financial assets that have no analytic solutions under realistic assumptions and thus require computational methods to be resolved. A forum for discussion of the growing impact of new technologies on teaching and the development of new tools to increase learning efficiency is provided in the symposium: new educational methodologies supported by new technologies offers. The symposium on mathematical models and information-intelligent transport systems researches in the field of flow-modelling of particles with motivated behaviour in complex networks, applied to traffic flows, pedestrian flows, ecology, etc. Special utility has this symposium in the traffic regulation in Moscow. The seventh symposium studies computational methods for linear and nonlinear optimization and numerical methods for solving nonlinear problems is given in another session. Bio-mathematics studies both theoretical and practical applications of population dynamics, eco-epidemiology, epidemiology of infectious diseases and molecular and antifenic evolution. The tenth symposium presents recent methodological developments in function approximation, multiway array decompositions, ODE and PDE solutions: applications from dynamical systems to quantum and statistical dynamics. The applications of fractional derivatives in sciences are considered in the twelfth symposium. The applied mathematics and computer science symposium focuses on new methods, technologies and applications of computer science and mathematics. Fractional Calculus from a theoretical viewpoint is considered in analytical and numerical methods for fractional differential equations. Obtaining a consistent description of the transition from small clusters to the liquid or solid state is a major challenge in computational chemistry and physics and will be addressed in the symposium: from clusters to the solid state. Crypto & codes aims to provide a forum there researchers can exchange the latest results, trends and open problems in the areas of cryptography and coding theory. Computational methods for fluid flow uses numerical methods and algorithms to solve and analyze the mathematical models that govern fluid
flows. Various numerical solution methods for large linear systems are presented and discussed in this session. The enormous potential of fixed point theory, which is needed in mathematics, engineering, chemistry, biology, economics, computer science, and other sciences, justify the great interest in fixed point theory in various abstract spaces and related applications. Finally, special sessions cover topics related to industrial mathematics, computational discrete mathematics and the numerical solution of differential equations.

We would like to thank the plenary speakers for their outstanding contributions to research and leadership in their respective fields, including physics, chemistry and engineering. We would also like to thank the special session organizers and scientific committee members, who have played a very important part in setting the direction of CMMSE 2014. Finally, we would like to thank the participants because, without their interest and enthusiasm, the conference would not have been possible.

We cordially welcome all participants. We hope you enjoy the conference.

Costa Ballena, Rota, Cádiz (Spain), July 15th, 2014

I. P. Hamilton, J. Vigo-Aguiar, J. Medina,
P. Schwerdtfeger, W. Sprößig, M. Demiralp,
E. Venturino, V.V. Kozlov, P. Oliveira
CMMSE 2014 Mini-symposia

<table>
<thead>
<tr>
<th>Session Title</th>
<th>Organizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Performance Computing (HPC)</td>
<td>Paula Oliveira &amp; J.A. Ferreira</td>
</tr>
<tr>
<td>P.D.E.'S in Life and Material Sciences</td>
<td>Luis Ortiz Gracia &amp; Carlos Vázquez Cendón</td>
</tr>
<tr>
<td>New Educational Methodologies Supported by New Technologies</td>
<td>Valerii V. Kozlov &amp; Andreas Schadschneider &amp; Alexander P. Buslaev</td>
</tr>
<tr>
<td>Mathematical Models and Information-Intelligent Systems on Transport</td>
<td>Maria Teresa Torres Monteiro</td>
</tr>
<tr>
<td>Computational Methods for Linear and Nonlinear Optimization</td>
<td>Juan R. Torregrosa &amp; A. Cordero</td>
</tr>
<tr>
<td>Computational Finance</td>
<td>Ezio Venturino, Nico Stollenwerk &amp; Maíra Aguiar</td>
</tr>
<tr>
<td>Numerical Methods for Solving Nonlinear Problems</td>
<td>Metin Demiralp &amp; Alper Tunga &amp; Burcu Tunga</td>
</tr>
<tr>
<td>Bio-mathematics</td>
<td>Jesús Medina &amp; Manuel Ojeda-Aciego</td>
</tr>
<tr>
<td>Recent Methodological Developments in Function Approximation, Multiway Array</td>
<td>Luisa Morgado</td>
</tr>
<tr>
<td>Decompositions, ODE and PDE Solutions: Applications From Dynamical Systems to Quantum and Statistical Dynamics</td>
<td></td>
</tr>
<tr>
<td>Mathematical Models for Computer Science</td>
<td>Ian Hamilton &amp; Peter Schwerdtfeger</td>
</tr>
<tr>
<td>Analytical and numerical methods for fractional differential equations</td>
<td>Antonio F. Roldán López de Hierro &amp; Juan Martínez Moreno</td>
</tr>
<tr>
<td>From clusters to the solid state</td>
<td>Zhenquan Li</td>
</tr>
<tr>
<td>Computational methods for fluid flow</td>
<td></td>
</tr>
<tr>
<td>Fixed Point Theory in various abstract spaces and related applications</td>
<td></td>
</tr>
</tbody>
</table>
Acknowledgements

We would like to express our gratitude to Universidad de Cádiz, our sponsor, for its assistance.

We also would like to thank all of the local organizers for their efforts devoted to the success of this conference:

- Jesús Medina Moreno, Universidad de Cádiz, Spain
- Juan Carlos Díaz Moreno- Universidad de Cádiz, Spain
- María Eugenia Cornejo Piñero - Universidad de Cádiz, Spain
- Eloisa Ramírez Poussa - Universidad de Cádiz, Spain

CMMSE 2014 Plenary Speakers

- Carlos Vázquez Cendón - University of A Coruña, Spain
- Peter Schwerdtfeger - University in Auckland, New Zealand
- Prof. Dr. Wolfgang Sprößig - TU Bergakademie Freiberg, Germany
- Maira Aguiar - University Lisbon, Portugal
- Metin Demiralp - Istanbul Technical University, Turkish
- Ezio Venturino - University of Torino, Italy
- Valery V. Kozlov - Russia Academy of Sciences, Russia
- Paula Oliveira - University of Coimbra, Portugal
Contents:

Volume I

The Ikebe-Schulz algorithm for inverting Hessenberg matrices.
Abderramán Marrero J. ................................................................. 1

Inversion of infinite tridiagonal matrices.
Abderramán Marrero J., Tomeo V., Torrano E. ...................... 5

Aceto L., Magherini C., Novati P. .................................................. 17

If “football fever could be a dose of dengue”, the Simon Hay fever should have given a dose of samba.
Aguiar M., Rocha F., Stollenwerk N. ........................................... 20

Borda-type algorithms to aggregate partial rankings.

An extension of parallel dynamical systems over graphs.
Aledo J.A., Martinez S., Valverde J.C. ..................................... 32

A numerical study of a nonlocal degenerate parabolic problem.
Almeida R.M.P., Antontsev S.N., Duque J.C.M. ..................... 37

On the characterization of almost strictly totally negative matrices.
Alonso P., Peña J.M., Serrano M.L. ......................................... 49

High Dimensional Model Representation in Image Processing.
Altin E.M., Tunga B. ................................................................. 55

A parallel algorithm for secure multicast.

A new Runge-Kutta-Nyström pair for the numerical solution of periodic initial value problems.
Anastassi Z.A., Kosti A.A. ......................................................... 71

Evaluation of time-series of satellite reflectance data for land delimitation using clustering algorithms.
Arango R.B., Díaz I., Campos A.M., Combarro E.F., Canas E.R. .. 77
A hybrid numerical method for a two-dimensional second order hyperbolic equation.
Araújo A., Neves C., Sousa E. .................................................................................. 88

Fitted parametrized spline curves help Mapmakers to define roads.
Ariza-López F.J., Barrera D., Reinoso J.F. ................................................................. 100

Time-Aware Multi-threaded Genetic Algorithm for Accelerating a Forest Fire Spread Forecast System.
Artés T., Cencerrado A., Cortés A., Margalef T. ....................................................... 103

Dissolution and bulk erosion in viscoelastic materials: numerical study.
Azhdari E., Ferreira J.A., de Oliveira P., da Silva P.M. ............................................. 115

A comparison of interval estimation methods in partially non-regular log-exponential models.
Barranco-Chamorro I., Jiménez-Gamero M.D., Alba-Fernández M.V. .................... 127

On spline-based differential quadrature.
Barrera D., González P., Ibáñez F., Ibáñez M. J. ......................................................... 131

Resolution of parabolic and hyperbolic PDEs using interpolating transient PS-splines.
Barrera D., González P., Palomares A., Pasadas M. ................................................... 137

Approximation of Multivariate Functions via Fluctuationlessness Theorem by Using Nested Taylor Decomposition.
Baykara N.A., Gürvit E. ............................................................................................. 144

Sensitivity analysis of a linear unbranched chemical process with n steps.
Bayón L., Otero J.A., Ruiz M.M., Suárez P.M., Tasis C. ............................................ 151

Improving an autotuning engine for 3D Fast Wavelet Transform on GPU.
Bernabé G., Cuenca J., García L.P., Giménez D. ....................................................... 158

Competition among invasive and native species: the case of European and mountain hares.
Berruti A., La Morgia V., Venturino E., Zappala S. ..................................................... 170

On how far mosquitos matter in describing dengue fever epidemiology

The role of stiffness in the proliferation of brain tumors.
Branco J.R., Ferreira J.A., de Oliveira P. ................................................................. 197

Accurate and Efficient Electronic Structure Modelling of Organic Molecular Crystals.
Brandenburg J.G., Grimme S. .................................................................................... 209

On methodological aspects of traffic theory.
Buslaev A.P., Gorodnichev M.G. ................................................................................ 220

A computational study on the quickest path problem with energy constraints.
Calvete H.I., del-Pozo L., Iranzo J.A. ........................................................................ 228

Pricing fixed-rate mortgages under jump-diffusion models for the house value.
Calvo-Garrido M.C., Vázquez C. .............................................................................. 238
On the effective thermal conductivity for a transversely isotropic two phase composite material.
Calvo-Jurado C., Parnell W.J. ............................................................249

Bifurcations of the roots of a 6-degree symmetric polynomial coming from the fixed point operator of a class of iterative methods.
Campos B., Cordero A., Magreñan A., Torregrosa J.R., Vindel P. .................................253

Optimizing the Performance of Financial Applications on Heterogeneous Architectures.
Castillo E., Camarero C., Borrego A., Bosque J.L. .........................................................265

Improved methodology for high-quantiles (VaR) estimator.
Castillo J. del, Padilla M., Serra I. .................................................................277

Malcev algebras and combinatorial structures.
Ceballos M., Núñez J., Tenorio A.F. .............................................................282

Accelerating HEVC using GPU-based heterogeneous platforms.

Cicek Y., Tanoglu G. ..................................................................................300

A high order in space uniformly convergent method for parabolic singularly perturbed reaction-diffusion systems.
Clavero C., Gracia J.L. ..................................................................................310

Cocaña-Fernández A., Ranilla J., Sánchez L. .........................................................318
Contents:

Volume II

A logic-based approach to compute a direct basis from implications.
Cordero P., Enciso M., Mora A., Ojeda-Aciego M., Rodríguez-Lorenzo E. ...................... 331

A new parametric class of iterative methods for solving nonlinear systems.
Cordero A., Feng L., Magreñán A.A., Torregrosa J.R. ................................................... 340

A class of bi-parametric families of iterative methods for nonlinear systems.
Cordero A., Maimó J.G., Torregrosa J.R., Vassileva M.P. ........................................... 350

On generalization of the variants of Newton's method for solving nonlinear equations.
Cordero A., Torregrosa J.R. ......................................................................................... 364

Adjoint triples versus extended-order algebras.
Cornejo M.E., Medina J., Ramírez-Poussa E. ............................................................... 375

An study for the Microwave Heating of a Half-Space through Lie symmetries and conservation laws.
de la Rosa R., Gandarias M.L., de los Santos M. ......................................................... 385

Error analysis in the reconstruction of a convolution kernel in a semilinear parabolic problem.
De Staelen R.H., Slodicka M. ...................................................................................... 396

A hybrid algorithm for the split generalized equilibrium and the system of variational inequality problems.
Deepho J., Kumam W. ................................................................................................. 399

A mathematical model of a single population with habitat fragmentation in progress.
Del-Valle R., Córdova-Lepe F. ....................................................................................... 429

Weighted Tridiagonal Matrix Enhanced Multivariance Products Representation of Finite Interval Data.
Demiralp E. .................................................................................................................. 441

Tridiagonal Matrix Enhanced Multivariance Product Representation (TMEMPR) for Matrix Decomposition.
Demiralp E., Demiralp M. ............................................................................................. 446
NNMFPACK: a versatile approach to an NNMF parallel library.

High-Order compact scheme for pricing variance swaps.
Dilloo M.J., Tangman D.Y. .................................................................466

A fourth order accurate finite difference solution of a multipoint nonlocal problem for the Laplace equation.
Dosiyev A.A. ..........................................................................................480

Improving the Solution of Band Linear Systems on Hybrid CPU+GPU Platforms.
Dufrechou E., Ezzatti P., Quintana-Ortí E.S., Remón A. ................................485

The influence of plotting positions on the correlation coefficient based on a Normal Q-Q Plot
Estudillo-Martínez M.D., Castillo-Gutiérrez S., Lozano-Aguilera E. .....................492

Multiresolution analysis for two-dimensional interpolatory schemes on uniform grids.
Fernández L., Fortes M.A., Rodríguez M.L. ..................................................495

Fractional modelling of Pennes’ bioheat equation using distributed order differential equations.

The effect of reversible binding sites on drug release from drug eluting stents.
Ferreira J.A., Naghipoor J., de Oliveira P. .......................................................519

The decay of solutions of a wave equation with memory.
Ferreira J.A., de Oliveira P., Pena G. ..................................................................531

Non-Fickian tracer transport in porous media.
Ferreira J.A., Pinto, L. .....................................................................................543

Modeling and Simulating Colonic Cell Renewal Disruption.
Figueiredo I.N., Leal C., Romanazzi G. .............................................................555

Cooperative solution of cryptarithmic problems.
Fontanori J.F. ..............................................................................................559

Filling holes with volume constraints.
Fortes M.A., González P., Palomares A., Pasadas M. .........................................568

Inverse-free recursive multiresolution algorithms for a data approximation problem
Fortes M.A., Raydan M., Sajo-Castelli A.M. .....................................................573

Computation of highly accurate binding energies of large clusters.
Friedrich J., Anacker T. ...................................................................................579

The Holling-Tanner predation model with a special weak Allee effect on prey
Gallego-Berrío L., González-Olivares E. .........................................................585

Solving the HP Protein Folding Problem by an Evolutionary Algorithm.
García-Martínez J.M., Garzón E.M., Cecilia J.M., Pérez-Sánchez H., Ortigosa P.M. 597
Extraneous attractors for Chebyshev’s method.
García-Olivo M., Gutiérrez J.M., Magreñán A. ................................................................. 602

Melting simulations of gallium clusters: transitions between the low temperature bulk phases.
Gaston N., Steenbergen K.G. ............................................................................................... 614

Comparing total and partial connections in a patched population model.
Gazzola C, Venturino E. ........................................................................................................ 617

A Fuzzy Representation of Vehicle Trajectories using Motion Data from H264/AVC Video.
Giralt J., Moreno-García J., Jiménez-Linares L., Del Castillo E., Rodríguez-Benítez L. ........... 635

Characteristic times for multiscale diffusion of active ingredients in coated textiles.
Goessens T., Constales D. ..................................................................................................... 647
Contents:

Volume III

Partial Separation of Unknowns in Cubic Array Decomposition.
Göksu G., Demiralp M. ................................................................. 650

The Parallel Conjugate Gradient Method using Unified Parallel C.
González-Domínguez J., Marques O.A., Martín M.J., Touriño J. ....................... 657

A Three Dimensional Discrete Model of a Nonlinear Hanging String with a Tip Mass.
González-Santos G., Vargas-Jarillo C. ................................................... 667

Reconstruction of an unknown Dirichlet boundary condition in a nonlinear parabolic problem containing Volterra operators.
Grimmonprez M., Slodicka M. .................................................................. 680

Numerical Integration of Bivariate Functions by Using Fluctuationlessness Theorem via Nested Taylor Decomposition.
Gürvit E., Baykara N.A. ....................................................................... 689

Existence and uniqueness of solutions for a class of coupled systems of fractional differential equations with integral boundary conditions.
Harjani J., Rocha J., Sadarangani K. ......................................................... 696

Heterogeneous catalysis on subnanometer-sized Pt alloys: Performance predictions for the conversion of propane to propene.
Hauser A.W., Bell A.T., Head-Gordon M. ............................................... 704

Crystal structure prediction using evolutionary algorithms.
Hermann A. ....................................................................................... 708

A fast method for computing the inverse of symmetric block arrowhead matrices.
Hołuobowski W., Trawiński T. ............................................................... 716

Derivative free iterative methods for approximating multiple roots.
Hueso J.L., Martínez E., TeruelC. .............................................................. 727

Exchange Corrections to the Random Phase Approximation - Why SOSEX is better than it should be.
Hummel F., Kresse G. ........................................................................ 737
Adaptive stability conditions for graph placement problem.
Ivanko E. ........................................................................................................... 739

Multi-probe three-dimensional placement planning for liver cryosurgery: comparison of different optimization methods.
Jaberzadeh A., Essert C. .................................................................................... 743

A Synthetic cell-based model for blood clot modelling.
Janela J., Pavlova J., Sequeira A., Fasano A. ..................................................... 755

Testing for the symmetric component in skew-symmetric distributions.
Jiménez-Gamero M.D., Alba-Fernández M.V., Jodrá P., Chalco-Cano Y. ................ 763

GPU implementation of Jacobi method for data arrays that exceed GPU-dedicated memory size.
Kochurov A., Golovashkin D. ........................................................................... 770

Two-Grid Finite Difference Scheme for Nonlinear Problems in Mathematical Finance.
Koleva M.N., Vulkov L.G. .................................................................................. 777

Operator Splitting Scheme for a Generalized Leland's Model.
Koleva M.N., Vulkov L.G. .................................................................................. 781

Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) Studies: Decomposing the Planarly Unfolded Three-way Arrays.
Korkmaz Özay, E., Demiralp M. ......................................................................... 785

The variational spline method for solving Troesch's problem.
Kouibia A., Pasadas M., Belhaj Z. ....................................................................... 794

Monotonic Walk of Particles on a Chainmail and Colored Matrices.
Kozlov V.V., Buslaev A.P., Tatashev A.G., Yashina M.V. ................................. 801

Strategic Optimisation in the Internet Search Market: Effect of User Dynamics.
Kudryashova N. .................................................................................................... 806

Sensitivity analysis of a mesh refinement method using the numerical solutions of 2D driven cavity flow.
Lal R., Li Z. ......................................................................................................... 817

Accuracy analysis of an adaptive mesh refinement method using benchmarks of 2-D steady incompressible lid-driven cavity flows.
Li Z., Wood R. .................................................................................................... 829

Scalable Consistency for Large-Scale Multiple Sequence Alignments.
Llados J., Guirado F., Cores F. ............................................................................. 840

A note on the use of a new family of anomalies in the study of the orbital motion.
López Orti J.A., Agost Gómez V., Barreda Rochera M. ........................................ 852

Exponential Trichotomy of Discrete Dynamical Systems and Applications.
Luminita Sasu A., Sasu B. .................................................................................. 856

Collocation solutions of a class of nonlinear Volterra integral equations.
Malindzisa H.S., Khumalo M. ........................................................................... 860
Human mobility and measles.
Marguta R., Parisi A. ................................................................. 868

An ensemble method for time series forecasting with simple exponential smoothing.
Martínez F., Frias M.P., Pérez M.D., Rivera A.J., del Jesús M.J. ........................................ 871

Multidimensional coincidence point results for a pair of weakly compatible mappings in fuzzy metric spaces.
Martínez-Moreno J., Roldán A., Roldán A., Cho Y.J. ......................................................... 877

Determining P optimum points to construct calibration estimators of the distribution function.
Martínez S., Rueda M.M., Martínez H., Arcos A. ............................................................... 885

A meshfree numerical method for the time-fractional diffusion equation.
Martins N.F., Morgado M. L., Rebel M. ................................................................. 892

Epidemiological models in semiclassical approximation: an analytically solvable model as test case.
Mateus L., Masoero D., Rocha F., Aguiar M., Skwara U., Ghaffari P., Zambrini J.C., Stollenwerk N. ................................................................. 904

An optimization approach to select portfolios of electricity generation projects - the Portuguese case.
Matos E., Monteiro M.T.T., Ferreira P., Cunha J. ................................................................. 918

Subpicture Parallel Approaches of HEVC Video Encoder.
Migallón H., Piñol P., López-Granado O., Malumbres M.P. ................................................. 927

Knots, Links and Coils in Clusters of Dipolar Particles.
Miller M.A., Farrell J.D., Chakrabarti D., Wales D.J. ............................................................ 939

Evaluation of a Portable RX Implementation on Heterogeneous Platforms.
Molero J.M., Garzón E M., García I., Quintana-Ortí E.S., Plaza A. ........................................ 947

Morgado L.F., Morgado M.L. ............................................................................................ 951

Approximation curves of spectral type: An application to signal processing.
Navascués M.A., Sebastián M.V., Ruiz C., Iso J.M. ............................................................. 957

A New Look to the Delta-Gamma Approach.
Ortiz-Gracia L., Oosterlee C.W. .......................................................................................... 964

Mercury Melting Simulations: Impact of Relativistic Effects.
Pahl E., Calvo F., Schwerdtfeger P. ...................................................................................... 968

DRBEM solution of natural convective heat transfer with the Brinkman-Forchheimer-extended Darcy model.
Pekmen B, Tezer-Sezgin M. ............................................................................................. 970
Contents:

Volume IV

A fixed point approach to the stability of radical quartic functional equation in fuzzy Banach spaces.
Phiangsungnoen S., Kumam P., Kumam W. .................................................................980

Towards an Improved Method of Dense 3D Object Reconstruction in Structured Light Scanning.
Portalés C., Morillo P., Orduña J.M. .............................................................................992

A novel optimal control scheme for Aedes mosquito reduction management.
Putra K.W., Goetz T. .........................................................................................................1002

Introducing weight functions to construct Interval Valued Fuzzy Relations.
Quirós P., Alonso P., Díaz I., Montes S. ...........................................................................1016

A Distance-Frequency Classification Algorithm.
Ralescu A., Díaz I. .............................................................................................................1025

Ramiro C., Vidal A.M., González A. ..................................................................................1037

The use of Newton’s method in vector form for solving nonlinear scalar equations.
Ramos H. .............................................................................................................................1044

Numerical solution of the reaction-wave-diffusion equation with distributed order in time.
Rebelo M., Morgado M.L. ....................................................................................................1057

Travelling waves solutions for a sixth-order Boussinesq equation.
Recio E., Gandarias M.L., Bruzón M.S. .............................................................................1069

Modelling a temperature dependent mosquito population.
Rocha F., Yang H.M., Mateus L., Aguiar M., Braumann C., Stollenwerk N. .................1076

Seasonality effects on Dengue.
Rodrigues H.S., Monteiro M.T.T., Torres D.F.M. .................................................................1084

Gause type prey-predator model with a generalized rational non-monotonic functional response.
Rojas-Palma A., González-Olivares E. .............................................................................1092
Coincidence point theorems on metric spaces via simulation functions and consequences.
Roldán A., Roldán C., Martínez-Moreno J., Karapinar E. .........................................................1104

Applications of Countable Fuzzy Numbers to Approximation Theory, Ranking and Fuzzy Regression Theory.
Roldán A., Roldán C., Martínez-Moreno J., Aguilar C. ...............................................................1116

A Fuzzy Statistical Regression Model with Application to Economic Data Analysis.
Roldán C., Roldán A., Martínez-Moreno J., Alfonso G. ...............................................................1124

Nonlinear self-adjointness for a Fisher equation with variable coefficients.
Rosa M., Bruzón M.S., Gandarias M.L. .........................................................................................1129

Use of randomized response techniques when data is obtained from two frames.
Rueda M.M., Arcos A., Cobo B. ......................................................................................................1137

Multinominal logistic estimation in dual frame surveys.
Rueda M.M., Arcos A., Molina D., Ranall M.G. ..........................................................................1149

On the application of general solutions of under-determined linear systems to the power flow problem.
Salgado R.S., Moraes G.R. ............................................................................................................1161

Time Scaling Effects on a Reduced Blood Coagulation Model.
Santos R.F., Sequeira A. ................................................................................................................1173

Efficient and accurate treatment of weak pairs in local CCSD(T) calculations.
Schütz M. ........................................................................................................................................1182

Playing with Hexagons and Pentagons: Topological and Graph Theoretical Aspects of Fullerenes.
Schwerdtfeger P., Ori O., Wirz L., Avery J. .....................................................................................1184

On stochastic models of vector borne diseases.
Skwara U., Rocha F., Aguiar M., Stollenwerk N. ..........................................................................1187

Chaos and noise in population biology: modelling dengue fever and data analysis, multi-strain dynamics, mosquitoes and climate.
Stollenwerk N., Rocha F., Mateus L., Skwara U., Ghaffari P., Aguiar M. .................................1194

On cluster flow model on multi-lane supporters.
Strusinskiy P.M. ............................................................................................................................1208

Impact of Heston model parameters by means of Uncertainty Quantification.
Suárez-Taboada M., Witteveen J.A.S., Oosterlee C.W., Grzelak L.A. ........................................1218

The Pricing of Quanto Options under Dynamic Correlation.
Teng L., Ehrhardt M., Günther M. .................................................................................................1228

A Fast Multi-Domain Lattice-Boltzmann Solver on Heterogeneous (Multicore-GPU) Architectures.
Valero-Lara P. ...............................................................................................................................1239

Parallel Approaches for Immersed-Boundary Method (Solid-Fluid Interaction).
Valero-Lara P. ...............................................................................................................................1251
The well-posedness of a mathematical model for an intermediate state between type-I and type-II superconductivity.
Van Bockstal K., Slodicka M. .................................................................1263

A Green Job Scheduling Policy for Heterogeneous Clouds.
Vilaplana J., Mateo J., Teixidó I., Solsona F. .................................................................1269

Exploiting multi-GPU systems using the Heterogeneous Programming Library.
Viñas M., Bozkus Z., Fraguela B.B., Andrade D., Doallo R. ................................................1280

CUBLAS-aided long vector algorithms.
Vorotnikova D.G., Golovashkin D.L. .................................................................1292

The Structures of Ruthenium Clusters.
Waldt E., Hehn A., Ahrichs R., Kappas M.D., Schooss D. ................................................1300

On model of totally connected flows on ringed net.
Yaroshenko A. .................................................................1302
## Contents:

### Volume V

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural convection in porous media within slender boxes: a numerical solution based on network method.</td>
<td>Cánovas M, Alhama I., Alhama F.</td>
<td>1309</td>
</tr>
<tr>
<td>Numerical simulation of Nusselt-Rayleigh correlation in Bénard cells.</td>
<td>Cánovas M., García G., Alhama I.</td>
<td>1315</td>
</tr>
<tr>
<td>Towards modelling complex mesoscale molecular environments.</td>
<td>de Jong W.A., Lin L., Yang C., Shen H., Oliker L.</td>
<td>1320</td>
</tr>
<tr>
<td>Improving data sharing on Infiniband Networks.</td>
<td>Diaz A.F., Ortega J., Ortiz A., Garay G., Prieto A.</td>
<td>1326</td>
</tr>
<tr>
<td>Magnetic properties of supported Pdn (n=1-7, 13) particles.</td>
<td>Gantassi O., Menakbi C., Guesmi H., Mineva T.</td>
<td>1335</td>
</tr>
<tr>
<td>Trap state amelioration via complementary L-type ligands for acetate and phosphonate passivated Cd-rich CdSe nanocrystals.</td>
<td>Hamilton I.</td>
<td>1339</td>
</tr>
<tr>
<td>Quantum oscillation phenomena induced by strong nonadiabatic perturbation.</td>
<td>Kitamura H.</td>
<td>1350</td>
</tr>
<tr>
<td>Molybdenum Carbide Nanoparticles as Catalysts for Hydrogenation Reactions, between the Cluster and Bulk Perspectives.</td>
<td>Liu X., Salahub D.R.</td>
<td>1359</td>
</tr>
<tr>
<td>Particle-based parallel fluid simulation in three-dimensional scene with implicit surfaces.</td>
<td>Nakata S., Sakamoto Y.</td>
<td>1367</td>
</tr>
<tr>
<td>Optimization of Wolbachia-based strategies to control of the Aedes aegypti mosquito population.</td>
<td>Rafikov M, da Silva Tchilian R.</td>
<td>1377</td>
</tr>
<tr>
<td>Solving a Long-Distance Routing Problem using Ant Colony Optimization.</td>
<td>Royo B., Sicilia J.A., Oliveros M.J., Larrodé E.</td>
<td>1393</td>
</tr>
</tbody>
</table>
Virtual Reality technology used as a didactic tool in bridge construction.
Sampaio Z., Viana L. .................................................................1405

Degenerate Ising model for atomistic simulation of crystal-melt interfaces.
Schebarchov D., Tim P. Schulze, Shaun C. Hendy. ......................................1414

Attribute selection for Naïve Bayes classifiers.
Tallón-Ballesteros A.J., Benavides-Vallejo J.E. ..............................................1417

Acceleration of Shallow Water Equation with Space-Time Conservation Element and Solution Element Method using Multiple Graphic Processing Units.
Tseng T.I, Kuo F.A. ..................................................................................1424

Applying Sun Tzu’s military strategy for small-medium business management.
Tu K.J. ........................................................................................................1433

Enterprises’ merging and acquisition to make knowledge transfer through project management.
Tu K.J. ........................................................................................................1439

Towards a Proof-of-Concept for Acoustic Localisation of Coronary Artery Stenoses.

A modified algorithm of FVIM using fractional power series expansion method for nonlinear fractional differential equations.
Yin F., Tian T., Song J., Zhu M. ......................................................................1451

Yin F., Tian T., Song J., Zhu M. ......................................................................1464

Approaches to Optimizing Signal Timing for Traffic bottleneck roads.
Yuan S., Zhao X., An Y. ................................................................................1475
Volume II
A logic-based approach to compute a direct basis from implications

P. Cordero¹, M. Enciso¹, A. Mora¹, M. Ojeda-Aciego¹ and E. Rodríguez-Lorenzo¹

¹ Universidad de Málaga, Andalucía Tech, Spain,
emails: pcordero@uma.es, enciso@lcc.uma.es, amora@ctima.uma.es, aciego@uma.es, estrellarodlor@ctima.uma.es

Abstract

Formal Concept Analysis is an emergent area in the topic of data analysis based on lattice theory. In this framework, a context is defined as the relation between a set of objects and a set of attributes and from here it is possible to extract relevant knowledge. One of the important topics is to study the implications between the attributes considered. In a context, some equivalent sets of implications can be computed using different techniques. We are studying the direct optimal basis, which enables us to compute the closure of a set of attributes in just one iteration. A Prolog method has been implemented that computes a direct basis from a set of implications.

Key words: Formal Concept Analysis, Implications, Basis, Logic, Prolog

1 Preliminaries

Formal Concept Analysis (FCA) considers a formal context as the relationship between a set of objects and a set of attributes. Formally, this relation can be defined as follows:

Definition 1.1 Let \( G \) be a set of objects, \( M \) a set of attributes, and \( I \subseteq G \times M \) a binary relation between \( G \) and \( M \), then the triple \( K = (G, M, I) \) is called a formal context.

In this triple, \( I \) is a binary relation between \( G \) and \( M \) such that, for \( o \in G \) and \( a \in M \), \( o I a \) means that the object \( o \) has the attribute \( a \). Two mappings are defined:

- \( (\cdot)' : 2^G \to 2^M \) is defined for all \( A \subseteq G \) as \( A' = \{ m \in M \mid g I m \text{ for all } g \in A \} \).
• \((\cdot)': 2^M \to 2^G\) is defined for all \(B \subseteq M\) as \(B' = \{g \in G \mid g \ I \ m \text{ for all } m \in B\}\).

These two mappings are closure operators and their fixpoints are the so-called formal concepts.

**Definition 1.2** A formal concept is a pair \((A, B)\) such that \(A \subseteq G, \ B \subseteq M, \ A' = B\) and \(B' = A\). Consequently, \(A\) and \(B\) are closed sets of objects and attributes, respectively called extent and intent.

If the following partial ordering
\[(A_1, B_1) \leq (A_2, B_2)\] if and only if \(A_1 \subseteq A_2\) (or equivalently \(B_1 \supseteq B_2\))
is considered in the set of all the formal concepts, then this set has the structure of a complete lattice, the concept lattice associated to the context.

In FCA, the study of attribute implications is an important topic that allows us to capture all the information which can be deduced from a context. In fact, these implications summarize the semantics of the context. They are expressions of the form \(A \to B\), where \(A\) and \(B\) are sets of attributes. A context satisfies the implication \(A \to B\) if every object that has all the attributes from \(A\) also has all the attributes from \(B\).

**Definition 1.3** An (attribute) implication of a formal context \(K = (G, M, I)\) is defined as a pair \((A, B)\), written \(A \to B\), where \(A, B \subseteq M\). Implication \(A \to B\) holds (is valid) in \(K\) if \(A' \subseteq B'\).

One can find in the literature a lot of works focused on how to compute the attribute implications, and specially on how to improve the implications obtained [9]. A basis of a set of implications is a basis of least cardinality. Some approaches compute the Duquenne-Guigues basis [7], focusing on the minimality of the cardinality of the set of implications. In [5], we have developed a method to remove redundant attributes in this kind of minimal basis. In [6] a new logic-based method to obtain a basis with minimal size in the left-hand side of the implications was proposed.

We emphasize the survey proposed by Bertet and Monjardet in [3]. In this work, the authors studied five unit implicational systems obtained by different authors in different fields and show that these formalisms are, in fact, identical. Between these definitions we outline the following:

**Definition 1.4** A set of implications, \(\Sigma\), is said to be minimal or non-redundant if, for all \(X \to Y \in \Sigma\), the set \(\Sigma \setminus \{X \to Y\}\) is not equivalent to \(\Sigma\).

**Definition 1.5** A set of implications, \(\Sigma\), is called minimum set if \(|\Sigma| \leq |\Sigma'|\) for all set of implications \(\Sigma'\) equivalent to \(\Sigma\).
Definition 1.6 A set of implications, \( \Sigma \), is said to be optimal if \( ||\Sigma|| \leq ||\Sigma'|| \) for all set of implications \( \Sigma' \) equivalent to \( \Sigma \), where the size of \( \Sigma \) is defined as

\[
||\Sigma|| = \sum_{X \rightarrow Y \in \Sigma} (|X| + |Y|)
\]

\( \Sigma = \{X_0 \rightarrow Y_0, \ldots X_n \rightarrow Y_n\} \) be a set of implications, it is said a left-minimal basis if there does not exist a \( X_i \rightarrow Y_i \) and a subset \( X'_i \subseteq X_i \) such that \( \Sigma \setminus \{X_i \rightarrow Y_i\} \cup \{X'_i \rightarrow Y_i\} \) is equivalent to \( \Sigma \).

One of the main problems in FCA is the computation of the closure of a set of attributes. Interesting approaches to this problem can be seen in [2,3,8].

Definition 1.7 A set of implications, \( \Sigma \), is said to be direct if the computation of the closure with respect to \( \Sigma \) of any set \( X \) of attributes requires just one iteration, that is, a unique traversal of the set of implications.

Obviously, the direct-optimal property is the combination of the directness and optimality properties and a method to obtain a direct-optimal basis is proposed in [2].

In this paper, we introduce a Prolog implementation based on the Simplification Logic proposed in [4] to compute a direct optimal basis from a set of implications. The use in FCA of the logic programming paradigm, via the Prolog language, is explored and proposed as a framework to develop fast prototypes in which the methods based in logic are implemented in a direct way.

2 Simplification logic and closures

In [4], the Simplification Logic for Functional Dependencies (i.e. implications) called \( \text{SL}_{FD} \), was proposed. This logic is equivalent to well-known Armstrong’s Axioms [1]. The main difference is that \( \text{SL}_{FD} \) avoids the use of transitivity, and is guided by the idea of simplifying the set of functional dependencies by efficiently removing redundant attributes.

\( \text{SL}_{FD} \) considers reflexivity as an axiom scheme, together with the following inferences rules:

[Ref] \[
A \supseteq B \\
A \rightarrow B
\]

[Frag] \[
A \rightarrow B \cup C \\
A \rightarrow B
\]

[Comp] \[
A \rightarrow B, C \rightarrow D \\
A \cup C \rightarrow B \cup D
\]

[Simp] \[
A \rightarrow B, C \rightarrow D \\
A \cup (C \setminus B) \rightarrow D
\]

If we have a set of implications \( \Sigma \) and a set of attributes \( A \), the closure of \( A \) in \( \text{SL}_{FD} \) is defined as the maximum set of attributes \( A^+ \) such that \( \Sigma \vdash A \rightarrow A^+ \).

Theorem 2.1 Let \( K = (G, M, I) \) be a formal context and \( \Sigma \) a basis for \( K \). For all \( A \subseteq M \), the equality \( A^+ = A'' \) holds.
We introduced an efficient algorithm to compute the closure of a set of attributes which improves the classical closure algorithms in [8], where the interested reader can find all the details of the algorithm, which was implemented in Prolog as the predicate `closure`. We show an example about the execution in Prolog with this predicate.

**Example 2.1** Let $\Omega = a, b, c, d, e, f$ a set of attributes. We would like to compute the closure of $A = \{a\}. \Sigma = \{ab \rightarrow c, bd \rightarrow d, de \rightarrow f, ce \rightarrow f\}$.

? closure([a],[ implication([a, [b,c]]), implication([a,b], [e]), implication([c], [f]), implication([c,e], [g])]).

% the result is
? ([a,b,c,e,g]).

True.

The set of attributes $\{a,b,c,e,g\}$ is the closure of $\{a\}$.

3 Computing a direct optimal basis

In this section, we show how the $\text{SL}^\text{FD}$ is the main foundation of the new method to compute the direct optimal basis from a set of implications. As we stated before, the best choice to make the implementation is Prolog, since the method is logic-based.

The input of the Prolog program is a set of attributes $M$ and a set of implications $\Sigma$ over the attributes in $M$. The output is the direct optimal basis equivalent to this set of implications.

The main predicate of the method developed is

\[
\text{directoptimalSL}(\text{ImplicationsInput}, \text{DOBasisOutput})
\]

which receives the set of implications from the input file `ImplicationsInput`, and renders the direct optimal basis, which equivalent to the input, in the output file `DOBasisOutput`.

There are two main operations in the method:

- **applyCompositionSimplification** is a predicate which applies exhaustively the rule composition-simplification to any pair of implications to obtain a direct basis.

- **applySimplification** is a predicate that applies exhaustively the rules of $\text{SL}^\text{FD}$ to remove redundancy in the direct basis obtained in the previous step in order to get a direct optimal basis.

Below, we outline the description of the Prolog method `applySimplificationLogic` which calls these two main predicates.
applySimplificationLogic:-
    fixPoint_Non,
    applyCompositionSimplification,
    removeRedundancy, !.

The predicate applyCompositionSimplification is applied exhaustively until the fix-
point is reached. To begin with, the method collects all the implications in a List, called
ListImplications, and invokes the predicate compositionsimplificationrule that ap-
plies the rule to any pair of implications.

applyCompositionSimplification:-
    fixpoint(no),
    fixPoint_Yes,
    findall([implication(X,Y)], implication(X,Y), ListImplications),
    compositionsimplificationrule(ListImplications),
    applyComposition,
    applyCompositionSimplification.

applyCompositionSimplification.

compositionsimplificationrule([]):- !.

compositionsimplificationrule([implication(X,Y)|Rest]):-
    compositionsimplification([X,Y],Rest),
    compositionsimplificationrule(Rest), !.

compositionsimplification([_,_,[],[]]):- !.

compositionsimplification([implication(A,B),[implication(C,D)|Rest]]):-
    cs(implication(A,B), implication(C,D)),
    compositionsimplification(implication(C,D), Rest),
    compositionsimplification(implication(A,B), Rest).

where the predicate cs applies the following rule derived from the $SL_{FD}$:

$$cs(implication(A, B), implication(C, D)) = implication(A \cup C \setminus B, D \setminus (A \cup B))$$

When an implication is added in certain step of the previous algorithm, the flag
fixpoint takes the value false in order to repeat the method again.

The second step is the exhaustive application of the rules of $SL_{FD}$ to remove redun-
dancy [4]. It has been implemented with the predicate applySimplification.
applySimplification:-
    lrSimplificationRule,!,
    applySimplification.
applySimplification.

% lrSimplification Rule: X → Y, Z → U to X → Y, Z-Y → U-Y
% Z-Y → U-Y if X included in Z
% Z → U another case

Example 3.1 In this example, we will compute the direct basis of the following set of implications stored in the file ganter.txt:

implication([a],[b]).
implication([a],[c]).
implication([d],[b]).
implication([c],[b]).
implication([a,b,c,d],[e]).
implication([a,b,c,d],[g]).
implication([a,b,c,e],[d]).
implication([a,b,c,e],[g]).

We call the Prolog predicate:

directoptimalSL('ganter.txt','Output_ganter.txt').

and it renders the direct optimal basis as follows:

-- INPUT --
:- dynamic implication2/2.
→ Preparing the input for Simplification:
→ Reduction:
→ Composition:
    implication([a,b,c,e],[g]) FD removed
    implication([a,b,c,e],[d]) FD removed
    implication([a,b,c,e],[d,g]) FD added
→ Composition:
    implication([a,b,c,d],[g]) FD removed
    implication([a,b,c,d],[e]) FD removed
    implication([a,b,c,d],[e,g]) FD added
→ Composition:
    implication([a],[c]) FD removed
    implication([a],[b]) FD removed
    implication([a],[b,c]) FD added

-- Implications composed and reduced:
:- dynamic implication/2.
implication([c], [b]).
implication([d], [b]).
implication([a, b, c, e], [d, g]).
implication([a, b, c, d], [e, g]).
implication([a], [b, c]).

** First Step **
→ Equivalence - Composition + Simplification:
  implication([c], [b]) + implication([a, b, c, e], [d, g]) |---
  implication([a, c, e], [d, g]) FD added

→ Equivalence - Composition + Simplification:
  implication([c], [b]) + implication([a, b, c, d], [e, g]) |---
  implication([a, c, d], [e, g]) FD added

→ Equivalence - Composition + Simplification:
  implication([d], [b]) + implication([a, b, c, e], [d, g]) =
  implication([a, c, d, e], [g]) not added implication([a, c, e], [d, g])

→ Equivalence - Composition + Simplification:
  implication([a], [b, c]) + implication([a, b, c, e], [d, g]) |---
  implication([a, c, d, e], [g]) FD added

→ Equivalence - Composition + Simplification:
  implication([a], [b, c]) + implication([a, b, c, d], [e, g]) |---
  implication([a, c, d], [e, g]) FD added

** Second Step **
...........
...........

*** END Composition + Simplification (A direct basis) **

implication([c], [b]).
implication([d], [b]).
implication([a, b, c, e], [d, g]).
implication([a, b, c, d], [e, g]).
implication([a], [b, c]).
implication([a, c, e], [d, g]).
implication([a, c, d], [e, g]).
implication([a, e], [d, g]).
implication([a, d], [e, g]).

*** BEGIN Simplification: Removing redundancy **
→ Equivalence - Simplification and Simplification + Axiom:
  implication([c], [b]) + implication([a, b, c, e], [d, g]) |---
  implication([a, b, c, e], [d, g]) Implication removed
  implication([a, c, e], [d, g]) yet exist
→ Equivalence - Simplification and Simplification + Axiom:
iplication([c], [b]) + implication([a,b,c,d],[e,g]) |---
implication([a,b,c,d],[e,g]) Implication removed
implication([a,c,d],[e,g]) yet exist

→ Equivalence - Simplification and Simplification + Axiom:
iplication([a],[b,c]) + implication([a,c,e],[d,g]) |---
implication([a,c,e],[d,g]) Implication removed
implication([a,e],[d,g]) yet exist

→ Equivalence - Simplification and Simplification + Axiom:
iplication([a],[b,c]) + implication([a,c,d],[e,g]) |---
implication([a,c,d],[e,g]) Implication removed
implication([a,d],[e,g]) yet exist

** OUTPUT: DIRECT OPTIMAL BASIS

implication([c], [b]).
implication([d], [b]).
implication([a], [b, c]).
implication([a, e], [d, g]).
implication([a, d], [e, g]).

4 Conclusions

We have presented a Prolog implementation of a novel method to compute the direct optimal basis from a set of implications. The soundness and correctness of the method will be included in a extended version of this paper. As future work, we are planning a comparison with other methods in the literature.

Acknowledgements

Supported by grants TIN12-39353-C04-01 and TIN2011-28084 of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

CORDERO, ENCISO, MORA, OJEDA-ACIEGO, RODRÍGUEZ-LORENZO


A new parametric class of iterative methods for solving nonlinear systems

Alicia Cordero¹, Licheng Feng², Á. Alberto Magreñán³ and Juan R. Torregrosa¹

¹ Instituto de Matemática Multidisciplinar, Universitat Politècnica de València
² Department of Mathematics, Northwestern Polytechnical University
³ Departamento de Matemáticas, Universidad Internacional de La Rioja

emails: acordero@mat.upv.es, fenglicheng1992@gmail.com, alberto.magrenan@gmail.com, jrtorre@mat.upv.es

Abstract

A new class of fourth-order iterative schemes for solving nonlinear equations and systems is proposed. The convergence analysis is established, obtaining the same order of convergence for any value of the parameter. Finally, some numerical tests are made in order to check the robustness of the methods and the real dynamical behavior on specific 2-dimensional systems is analyzed, studying their stability depending on the parameter.

Key words: nonlinear systems of equations, iterative methods, basins of attraction

MSC 2000: 65H05, 65H10

1 Introduction

It is a well-known fact that problems in science and engineering usually involve nonlinear equations or systems, \( F(x) = 0 \), where \( F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n, n \geq 1 \). The most of times an analytical solution of these kind of problems is not available and, in these cases, we must use iterative methods in order to estimate their solutions.

In the scalar case (\( n = 1 \)), the efficiency of an iterative scheme for solving nonlinear equations is usually measured by means of the efficiency index, defined by Ostrowski in [8] as \( I = p^\frac{1}{d} \), where \( p \) is the order of convergence of the method and \( d \) is the number of functional evaluations per step. In order to get optimal schemes, in the sense of Kung-Traub’s conjecture [6], we must draw on
A NEW PARAMETRIC CLASS OF ITERATIVE METHODS FOR SOLVING NONLINEAR SYSTEMS

multipoint iterative schemes. Many of them are very useful for solving nonlinear equations but they are not applicable to nonlinear systems.

In this work, we present a parametric family of iterative procedures for solving the nonlinear multidimensional equation \( F(x) = 0 \). In case of \( n = 1 \), its iterative expression is

\[
y_k = x_k - \frac{2 f(x_k)}{3 f'(x_k)},
\]

\[
x_{k+1} = x_k - \left( 1 - \frac{3 u_k(1 + \beta u_k)}{4 1 + u_k(\beta + \frac{3}{2})} \right) \frac{f(x_k)}{f'(x_k)},
\]

(1)

where \( u_k = \frac{f(y_k) - f(x_k)}{f'(x_k)} \) and \( \beta \) is an arbitrary parameter. Let us observe that, for \( \beta = 0 \), this class of schemes includes the known Jarratt’s scheme [5]. It can be proved that the local order of convergence of the elements of the family is four and so, all of them are optimal methods.

We will compare the proposed schemes with the well-known Newton’s method, whose iterative expression is

\[
x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots
\]

We will also use the following iterative scheme, designed by Sharma et al. in [10], with the following iterative expression

\[
y^{(k)} = x^{(k)} - \frac{2}{3} \left[ F'(x^{(k)}) \right]^{-1} F(x^{(k)})
\]

\[
x^{(k+1)} = x^{(k)} - \frac{1}{2} \left[ -I + \frac{9}{4} \left[ F'(x^{(k)}) \right]^{-1} F'(y^{(k)}) + \frac{3}{4} \left[ F'(x^{(k)}) \right]^{-1} F'(y^{(k)}) \right] \left[ F'(x^{(k)}) \right]^{-1} F(x^{(k)}).
\]

In Section 2, the analysis of convergence is made and some specific elements of the family are presented. In Section 3, the dynamical and numerical behavior of the members of the family is analyzed on some test systems. The aim will be to choose the best elements of the class, in terms of stability and reliability and it will be made by using dynamical and numerical tools.

2 Convergence analysis

In the following result the local order of convergence of the proposed class of methods (1), extended to nonlinear systems of equations, is analyzed. We will show that, under the standard conditions, fourth-order of convergence is reached for any real value of the parameter \( \beta \).

An interesting aspect of this class of method is that it can be directly extended to nonlinear systems \( F(x) = 0 \), as only first derivatives of the nonlinear functions appear in the denominators of the iterative scheme (as it happens with Jarratt’s method). Then, the corresponding iterative expression is
\[ y^{(k)} = x^{(k)} - \frac{2}{3} [F'(x^{(k)})]^{-1} F(x^{(k)}), \]
\[ x^{(k+1)} = x^{(k)} - \left( I - \frac{3}{4} \left[ I + \left( \frac{3}{2} + \beta u^{(k)} \right) \right]^{-1} \left( u^{(k)} + \beta u^{(k)^2} \right) \right) [F'(x^{(k)})]^{-1} F(x^{(k)}), \]

where \( u^{(k)} = [F'(x^{(k)})]^{-1} (F'(y^{(k)}) - F'(x^{(k})) \). It is important to remark that this extension is made holding the order of convergence, as it is stated in the following result.

**Theorem 1** Let \( \alpha \in D \) be a zero of a sufficiently differentiable function \( F : D \in \mathbb{R}^n \rightarrow \mathbb{R}^n \) in a convex set \( D \) with nonsingular Jacobian in \( \alpha \), and let also \( x^{(0)} \) be an initial approximation close enough to \( \alpha \). For any real value of parameter \( \beta \), the scheme defined in (2) provides fourth order of convergence, whose error equation is given by
\[ e^{(k+1)} = \left( 1 - \frac{8}{3} \beta - C_3 C_2 + \frac{1}{9} C_4 \right) e^{(k)^3} + O(e^{(k)^5}), \]

where \( C_k = (1/k!)[F'(\alpha)]^{-1} F^{(k)}(\alpha), k = 2, 3, \ldots \) and \( e^{(k)} = x^{(k)} - \alpha \).

In terms of the efficiency index, all the members of the family have the same index, \( I = 4 \frac{1}{2n^2 + n} \), which is better than the one of Newton’s scheme (see Figure 1).

![Figure 1: Efficiency index of Newton’s method and proposed family](image-url)
• when $\beta = -\frac{3}{2}$, a simpler iterative expression is obtained, whose second step is
\[
x^{(k+1)} = x^{(k)} - \left( I - \frac{3}{4} \left( u^{(k)} + \beta u^{(k)^2} \right) \right) \left[ F'(x^{(k)}) \right]^{-1} F(x^{(k)}).
\]

• Also when $\beta = 0$, a simplex iterative expression is found, corresponding to classical Jarratt’s method,
\[
x^{(k+1)} = x^{(k)} - \left( I - \frac{3}{4} \left[ I + \frac{3}{2} u^{(k)} \right]^{-1} u^{(k)} \right) \left[ F'(x^{(k)}) \right]^{-1} F(x^{(k)}).
\]

• If $\beta = \frac{3}{8}$, the error equation is simpler, depending only of the value of the derivatives of the nonlinear function at the solution. In this case, the iterative expression is
\[
x^{(k+1)} = x^{(k)} - \left( I - \frac{3}{4} \left[ I + \frac{15}{8} u^{(k)} \right]^{-1} \left( u^{(k)} + \frac{15}{8} u^{(k)^2} \right) \right) \left[ F'(x^{(k)}) \right]^{-1} F(x^{(k)}).
\]

3 Numerical and dynamical analysis

In this section, we compare the described class of schemes with the well-known Newton’s procedure, that has second-order of convergence. We will compare it with some elements of the defined family of iterative methods (2). This comparison will be made, at a first stage, by using dynamical tools: we will use the software described in [2] in order to draw the dynamical planes associated to each one of the members of our proposed class of methods on some specific nonlinear functions. We will try to deduce from the observed behavior of the methods which elements of the family are more stable and reliable.

All these methods will be employed to solve some nonlinear systems, appearing in [7]:

• The first one is the Brown almost-linear function, which in case of dimension two is described by
\[
F_1(x) = (2x_1x_2 - 3, x_1x_2 - 1),
\]
whose real solutions are $\alpha_1 = (1, 1)^T$ and $\alpha_2 = (1/2, 2)^T$.

• and the Frendenstein-Roth problem, described by the nonlinear function
\[
F_2(x) = (-13 + x_1 + ((5 - x_2) - 2) x_2, -29 + x_1 + ((x_2 + 1)x_2 - 14) x_2),
\]
being the only real solution $\alpha = (5, 4)^T$. 

©CMMSE ISBN: 978-84-616-9216-3  Page 343 of 1485
In order to analyze the dynamical behavior of the proposed class of iterative methods, it is necessary to recall some basic dynamical concepts. For a more extensive review of these concepts, please see [1], [4] or [9].

Let us denote by \( G(x) \) the vectorial fixed-point function associated to the iterative method on polynomial \( p(x) \).

**Definition 1** Let \( G : \mathbb{R}^n \to \mathbb{R}^n \) be a vectorial rational function. The orbit of a point \( x^{(0)} \in \mathbb{R}^n \) is defined as the set of successive images of \( x^{(0)} \) by the vectorial rational function,

\[
\left\{ x^{(0)}, G(x^{(0)}), \ldots, G^m(x^{(0)}), \ldots \right\}.
\]

We recall a known result in Discrete Dynamics that gives the stability of fixed points for non-linear operators.

The dynamical behavior of the orbit of a point of \( \mathbb{R}^n \) can be classified depending on its asymptotic behavior. In this way, a point \( x^* \in \mathbb{R}^n \) is a fixed point of \( G \) if \( G(x^*) = x^* \).

**Theorem 2** ([9], page 558) Let \( G \) from \( \mathbb{R}^n \) to \( \mathbb{R}^n \) be \( C^2 \). Assume \( x^* \) is a fixed point. Let \( \lambda_1, \lambda_1, \ldots, \lambda_n \) be the eigenvalues of \( G'(x^*) \).

a) If all the eigenvalues \( \lambda_j \) have \( |\lambda_j| < 1 \), then \( x^* \) is attracting.

b) If one eigenvalue \( \lambda_{j_0} \) has \( |\lambda_{j_0}| > 1 \), then \( x^* \) is unstable, that is, repelling or saddle.

c) If all the eigenvalues \( \lambda_j \) have \( |\lambda_j| > 1 \), then \( x^* \) is repelling.

Then, if \( x^* \) is an attracting fixed point of the rational function \( G \), its basin of attraction \( A(x^*) \) is defined as the set of pre-images of any order such that

\[
A(x^*) = \left\{ x^{(0)} \in \mathbb{R}^n : G^m(x^{(0)}) \to x^*, m \to \infty \right\}.
\]

The set of points whose orbits tends to an attracting fixed point is defined as the Fatou set, \( \mathcal{F}(G) \). The complementary set in \( \mathbb{R}^n \), the Julia set \( \mathcal{J}(G) \), is the closure of the set consisting of its repelling fixed points, and establishes the boundaries between the basins of attraction.

The dynamical tools used are real planes that has been generated by slightly modifying the routines described in [2], designed for Matlab, with version R2013a is used along this paper. In them, a mesh of \( 400 \times 400 \) points has been used for any of the dynamical planes associated to each member of the parametric family. Then, if an starting point of this mesh converges to one of the roots of the polynomial (using \( 10^{-3} \) as the bound of the difference between the root and the iteration), it is painted in the color assigned to the root which has converged to. The color used is brighter when the number of iterations is lower. If it reaches the maximum number of iterations (40 in this case) without converging to any of the roots, it is painted in black.
In Figure 2 it can be observed that, for the nonlinear function $F_1(x)$, the dynamical behavior of Newton’ and Jarratt’s schemes is basically the same, with the only difference of the order of convergence, which is fourth-order in the last case.

If we observe the behavior of the different elements of the family showed in Figure 3, we notice that there exist different kinds of dynamical planes, depending of the value of the parameter. A very stable behavior is also found for values of the parameter close to zero, as in cases of Figures 3b and 3c, for $\beta = \mp \frac{1}{6}$, respectively. In these cases, the basins of attraction of the two roots are as clear and stable as in classical methods and the fourth-order of convergence is held. For $\beta = -\frac{3}{2}$, which reduces the error equation of the method, also a very stable behavior is found, but there are some black regions, which correspond with slower convergence (starting points of the method in these regions would take more than 40 iterations to converge to the root, at a distance lower than $10^{-3}$). Finally, when $\beta = 1$ is used (see Figure 3d), quite unstable behavior is found, with big black regions that, in this case, correspond to divergence points.

Respect to the second problem, it can be seen in Figures 4 and 5 that the convergence to the only solution of this function depends basically of the second component of the initial estimation. In all cases, there exist a band of stable behavior around the solution, where good properties of convergence can be found. Nevertheless, this region is wider in the case of classical methods (Figures 4a and 4b), but in case of the proposed methods, to be closer to the solution is needed in all cases, except those whose value of the parameter $\beta$ is near to zero. We also found stable behavior in these cases.

Now, we will undertake the numerical tests, with a more exigent error estimation and variable precision arithmetics of 2000 digits of mantissa. These test have been made by using the stopping
criterium $\| F(x^{(k+1)}) \| + \| x^{(k+1)} - x^{(k)} \| < 10^{-500}$. In this way we assure that, in case of convergence, we have done it to a solution. In Table 1, we show the numerical results obtained for functions $F_1(x)$ and $F_2(x)$, with different initial estimations. We show, for each method, the number of iterations and the root we have converged to and the approximated computational order of convergence $ACOC$ (see [3]),

$$p \approx ACOC = \frac{\ln(\| x^{(k+1)} - x^{(k)} \| / \| x^{(k)} - x^{(k-1)} \|)}{\ln(\| x^{(k)} - x^{(k-1)} \| / \| x^{(k-1)} - x^{(k-2)} \|)},$$

Figure 3: Dynamical planes of the proposed methods on $F_1(x)$
The value of $ACOC$ that appears in Table 1 is the last coordinate of vector (6) when the variation between its values is small. In some cases, the approximated order of convergence is not stable and it is not shown in the table. Observing Table 1, it can be noticed that the performance of the selected elements of the proposed family is at least equal and sometimes better than the classical Jarratt’s scheme. The value of the parameter $\beta$ associated to the best members of the class are
around 0 (like $\beta = \pm \frac{1}{6}$) and, in fact, the best results are obtained for $\beta = \frac{3}{8}$, that simplifies the error equation.

### Acknowledgements

This research was partially supported by el Ministerio de Ciencia y Tecnología MTM2011-28636-C02-01,02 and PAID SP20120474, from Universitat Politècnica de València.

### References


A class of bi-parametric families of iterative methods for nonlinear systems*

Alicia Cordero¹, Javier G. Maimó², Juan R. Torregrosa¹ and María P. Vassileva²

¹ Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Spain
² Area de Ciencias Básicas y Ambientales, Instituto Tecnológico de Santo Domingo (INTEC), Dominican Republic

emails: acordero@mat.upv.es, javier.garcia@intec.edu.do, jrtorre@mat.upv.es, maria.penkova@intec.edu.do

Abstract

In this paper we design, by using a combination of the Ostrowski’ and Chun’s methods, two bi-parametric families of predictor-corrector iterative schemes with order of convergence 4 for solving systems of nonlinear equations. The predictor of the first family is Newton’s method, and the predictor of the second one is Steffensen’s method. This makes the second family derivative free. Some numerical tests are performed to compare the proposed methods with existing ones and to confirm the theoretical results.

Key words: Nonlinear equation, system of nonlinear equations, divided differences, iterative schemes, optimal, methods, efficiency index
MSC 2000: 65H05, 65H10

1 Introduction

Solving nonlinear equations and systems is an important task in theory and practice, not only for Applied Mathematics, but also for many branches of Science and Engineering. A glance at the bibliography [16, 19, 17] show a high level of contemporary interest.

*This research was supported by FONDOCYT 2011-1-B1-33 República Dominicana and Ministerio de Ciencia y Tecnología, MTM2011-28636-C02-02, España.
Iterative methods for nonlinear systems

Let us consider the problem of finding a simple zero of the nonlinear function \( f : I \subset \mathbb{R} \to \mathbb{R} \), that is, a solution \( \xi \in I \) of the nonlinear equation \( f(x) = 0 \). The most used iterative techniques to determine these roots can be classified as: (a) methods that require only functional evaluations of \( f \), and (b) schemes whose formula require evaluations of the function and its derivatives. There are two simple and effective known methods that represent these classes: Steffensen’s scheme \cite{18}

\[
\begin{align*}
x_{k+1} &= x_k - \frac{f(x_k)}{f[\omega_k, x_k]},
\end{align*}
\]

where \( \omega_k = x_k + f(x_k) \) and \( f[\omega_k, x_k] = \frac{f(\omega_k) - f(x_k)}{\omega_k - x_k} \), and Newton’s procedure (see \cite{15})

\[
\begin{align*}
x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)},
\end{align*}
\]

where \( f'(x) \) is the first Fréchet derivative of the function \( f(x) \). The order of convergence of both methods is two.

Multipoint methods have been developed as a result of the search for iterative methods for solving nonlinear equations with fast convergence and small number of operations or functional assessments per step. The most important class of multistep schemes are the optimal methods in the sense of Kung-Traub conjecture \cite{12}.

The problem of solving a system of nonlinear equations is avoided as far as possible. Generally, the nonlinear system is approximated by a system of linear equations. When this is not satisfactory, the problem must be confronted directly. The direct way is to adapt the methods designed for the scalar case to several variables. A scalar variable is replaced by a vector incorporating all the variables. Hence arises the greatest difficulty to get new iterative methods for nonlinear systems, since not always the methods of nonlinear equations are extensible to systems directly.

The aim of this work is to design a new family of iterative methods for nonlinear equations using some of the known methods and subsequently extend it to systems of nonlinear equations. For this purpose we have used Ostrowski’ \cite{16} and Chun’s \cite{5} methods with iterative schemes

\[
\begin{align*}
x_{k+1} &= y_k - \frac{f(x_k)}{f(x_k) - 2f(y_k)} f'(x_k), \\
x_{k+1} &= y_k - \frac{f(x_k) + 2f(y_k) f'(x_k)}{f'(x_k)},
\end{align*}
\]

respectively, where \( y_k \) is the step of Newton’s method.

The paper is organized as follows: we start in Section 2 with the design of the families of iterative methods for nonlinear equations, with and without derivatives. Section 3 is devoted to the extension of the obtained families to systems of nonlinear equations by using divided difference operator. By means of standard test functions, in Section 4, we confirm the theoretical results. We finalize the paper with some concluding remarks in Section 5.
Design of the families for nonlinear equations

We propose a new family as a combination of Ostrowski’ and Chun’s methods in the form:

\[ y_k = x_k - \alpha \frac{f(x_k)}{f'(x_k)}, \]

\[ x_{k+1} = y_k - \left[ \frac{f(x_k)}{a_1 f(x_k) + a_2 f(y_k)} + \frac{b_1 f(x_k) + b_2 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f'(x_k)}, \]  

(5)

where \( \alpha, a_1, a_2, b_1 \) and \( b_2 \) are real parameters. In the following result we will show the values of the parameters that guarantee the order of convergence is at least 4.

Theorem 1 Let \( f : I \subseteq \mathbb{R} \to \mathbb{R} \) be a sufficiently differentiable function in the open interval \( I \), such that \( \xi \in I \) is a simple solution of the nonlinear equation \( f(x) = 0 \). Then, the sequence \( \{x_k\}_{k \geq 0} \) obtained by using expression (5) converges to \( \xi \) with local order of convergence at least four if \( \alpha = 1, a_2 = a_1^2(b_2 - 2), b_1 = 1 - \frac{1}{a_1} \) and for all \( a_1 y b_2 \in \mathbb{R} \) with \( a_1 \neq 0 \).

The iterative formula obtained from Ostrowski-Chun’s method (OC) is:

\[ x_{k+1} = y_k - \frac{1}{a_1} \left[ \frac{f(x_k)}{f(x_k) + a_1(b_2 - 2)f(y_k)} + \frac{(a_1 - 1)f(x_k) + a_1 b_2 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f'(x_k)}, \]  

(6)

that defines a two-parameter family of optimal methods with order of convergence four.

Particular cases:

1. When \( a_1 = 1 \), the iterative formula takes the form:

\[ x_{k+1} = y_k - \left[ \frac{f(x_k)}{f(x_k) + (b_2 - 2)f(y_k)} + \frac{b_2 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f'(x_k)}, \]

and we have a one parametric family that have the original methods as particular cases:

(a) if \( b_2 = 2 \), we have Chun’s method (4) and

(b) if \( b_2 = 0 \), we get Ostrowski’s scheme (3).

2. For any \( a_1 \neq 0 \) and \( b_2 = 0 \), the iterative formula is:

\[ x_{k+1} = y_k - \frac{f(x_k) - 2(a_1 - 1)f(y_k)}{f(x_k) - 2a_1 f(y_k)} \frac{f(y_k)}{f'(x_k)}, \]

If we express \( -2(a_1 - 1) = \beta \), then \( -2a_1 = \beta - 2 \) we get King’s family [13]

\[ x_{k+1} = y_k - \frac{f(x_k) + \beta f(y_k)}{f(x_k) + (\beta - 2)f(y_k)} \frac{f(y_k)}{f'(x_k)}. \]
3. For any $a_1 \neq 0$ and $b_2 = 1$, the iterative formula takes the form:

$$x_{k+1} = y_k - \frac{1}{a_1} \left[ \frac{f(x_k)}{f(x_k) - a_1 f(y_k)} + \frac{(a_1 - 1)f(x_k) + a_1 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f'(x_k)}.$$

At this point, can we get a similar family by approximating the derivatives by divided differences and preserving the order of convergence? The answer is given in the following result.

**Theorem 2** Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a sufficiently differentiable function in the open interval $I$, such that $\xi \in I$ is a simple solution of the nonlinear equation $f(x) = 0$. Then, the sequence $\{x_k\}_{k \geq 0}$ obtained by using the expression

$$y_k = x_k - \alpha \frac{f(x_k)}{f[z_k, x_k]},$$

$$x_{k+1} = y_k - \left[ \frac{f(x_k)}{a_1 f(x_k) + a_2 f(y_k)} + \frac{b_1 f(x_k) + b_2 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f[z_k, x_k]},$$

where $z_k = x_k + f^2(x_k)$ and $f[z_k, x_k] = \frac{f(z_k) - f(x_k)}{z_k - x_k}$, converges to $\xi$ with order of convergence four if $\alpha = 1$, $a_2 = a_1^2 (b_2 - 2)$, $b_1 = 1 - \frac{1}{a_1}$ and for all $a_1 \neq 0$ and $b_2 \in \mathbb{R}$.

**Particular cases:**

1. When $a_1 = 1$, the iterative formula takes the form:

$$x_{k+1} = y_k - \left[ \frac{f(x_k)}{f(x_k) + (b_2 - 2)f(y_k)} + \frac{b_2 f(y_k)}{f(x_k)} \right] \frac{f(y_k)}{f[z_k, x_k]}$$

and we have a one parametric family that includes the derivative-free versions of original schemes:

(a) if $b_2 = 2$, we have Chun’s derivative-free method (CM2), whose iterative expression is

$$x_{k+1} = y_k - \frac{f(x_k) + 2f(x_k)}{f(x_k)} \frac{f(y_k)}{f[z_k, x_k]},$$

(b) if $b_2 = 0$, we obtain Ostrowski’s derivative-free method (OM2), with the iterative expression

$$x_{k+1} = y_k - \frac{f(x_k)}{f(x_k) - 2f(x_k)} \frac{f(y_k)}{f[z_k, x_k]}.$$

©CMMSE ISBN: 978-84-616-9216-3 Page 353 of 1485
2. When $a_1 \neq 0$ and $b_2 = 0$, the iterative formula takes the form:

$$x_{k+1} = y_k - \frac{f(x_k) - 2(a_1 - 1)f(y_k) - f(y_k)}{f(x_k) - 2a_1 f(y_k)} f[y_k].$$

If we denote $-2(a_1 - 1) = \beta$, then $-2a_1 = \beta - 2$ and we get the derivative-free King’s family

$$x_{k+1} = y_k - \frac{f(x_k) + \beta f(y_k) - f(y_k)}{f(x_k) + (\beta - 2)f(y_k)} f[y_k].$$

3. When $a_1 \neq 0$ and $b_2 = 1$, the resulting iterative formula is:

$$x_{k+1} = y_k - \frac{1}{a_1} \left[ \frac{f(x_k)}{f(x_k) - a_1 f(y_k)} + \frac{(a_1 - 1)f(x_k) + a_1 f(y)}{f(x_k)} \right] f[y_k].$$

### 3 Extension of the families to systems of nonlinear equations

The objective of this section is to give a generalization to several variables of the families of iterative methods obtained in Section 2 preserving the local order of convergence. In order to get this aim, we are going to use the divided difference operator.

Let us consider a sufficiently differentiable function $F : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ in a convex set $\Omega \in \mathbb{R}^n$ and let $\xi$ be a solution of the nonlinear system $F(x) = 0$. The divided difference operator of $F$ on $\mathbb{R}^n$ is a mapping $[\cdot, \cdot; F] : \Omega \times \Omega \subset \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$ (see [15]), which is defined by $[x + h, x; F] = \int_0^1 F'(x + th) dt$, $\forall (x, h) \in \mathbb{R}^n \times \mathbb{R}^n$.

The extension to multivariate case of method (5) requires to rewrite the iterative expression in such a way that no functional evaluation of the nonlinear function remain at the denominator, as they will become vectors in the multivariate case. To get this aim, let us consider that, being the first step of the iterative process $y_k = x_k - \alpha \frac{f(x_k)}{f'(x_k)}$, $f(x_k)$ can be expressed as $f(x_k) = \frac{1}{\alpha}(x_k - y_k)f'(x_k)$. By using this, we can rewrite the quotient $\frac{f(y_k)}{f(x_k)}$ as

$$\frac{f(y_k)}{f(x_k)} = 1 - \alpha \frac{f[y_k]}{f'(x_k)}.$$

By using this transformation, the proposed family (5) is fully extensible to several variables,

$$\begin{align*}
y^{(k+1)} &= x^{(k)} - \alpha [F'(x^{(k)})]^{-1} F(x^{(k)}) \\
x^{(k+1)} &= y^{(k)} - G(x^{(k)}, y^{(k)}) [F'(x^{(k)})]^{-1} F(y^{(k)}),
\end{align*}$$

where

$$\begin{align*}
G(x^{(k)}, y^{(k)}) &= \left[ (a_1 + a_2) I - \alpha a_2 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right]^{-1} \\
&\phantom{=} + \left( b_1 + b_2 \right) I - \alpha b_2 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F],
\end{align*}$$

©CMMSE
Let $F : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a sufficiently differentiable function in a convex set $\Omega$ and $\xi \in \mathbb{R}^n$ be a solution of $F(x) = 0$. Let us suppose that $F'(x)$ is continuous and non-singular in $\xi$. Then, the sequence $\{x_k\}_{k \geq 0}$ obtained by using the iterative scheme (10), converges to $\xi$ with order of convergence at least four if $\alpha = 1$, $a_2 = a_1^2(b_2 - 2)$, $b_1 = 1 - \frac{1}{a_1}$ and for all $a_1$, $y$, $b_2 \in \mathbb{R}$ with $a_1 \neq 0$.

Under the assumptions made in the previous result, the iterative scheme of the b-parametric family (10) takes the form:

$$
\begin{align*}
    y^{(k)} &= x^{(k)} - [F'(x^{(k)})]^{-1}F(x^{(k)}), \\
    x^{(k+1)} &= y^{(k)} - G(x^{(k)}, y^{(k)})[F'(x^{(k)})]^{-1}F(y^{(k)}), \\
    G(x^{(k)}, y^{(k)}) &= \frac{1}{a_1} \left[ (1 + a_1 b_2 - 2a_1)I - a_1 (b_2 - 2) [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right]^{-1} \\
    &\quad + \frac{1}{a_1} \left[ (a_1 + a_1 b_2 - 1)I - b_2 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right].
\end{align*}
$$

**Particular cases:**

1. When $a_1 = 1$,

$$
G(x^{(k)}, y^{(k)}) = \left[ (b_2 - 1)I - (b_2 - 2) [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right]^{-1}
$$

and we have a one-parametric family. Some particular cases of this class are the following:

   (a) If $b_2 = 2$, we have Chun’s method transferred to systems (CM3)

$$
x^{(k+1)} = y^{(k)} - \left( I - 2 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right) [F'(x^{(k)})]^{-1} F(y^{(k)}). \tag{14}
$$

   (b) If $b_2 = 0$, we get Ostrowski’s method transferred to systems (OM3)

$$
x^{(k+1)} = y^{(k)} - \left( -I + 2 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F] \right)\left[ F'(x^{(k)}) \right]^{-1} F(y^{(k)}). \tag{15}
$$

2. For any $a_1 \neq 0$ and $b_2 = 0$,

$$
G(x^{(k)}, y^{(k)}) = \frac{(3 - 2a_1) I + 2(a_1 - 1) [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F]}{(1 - 2a_1) I + 2a_1 [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F]}. \tag{16}
$$

If we denote $-2(a_1 - 1) = \beta$ we get the King’s family transferred to systems

$$
x^{(k+1)} = y^{(k)} - \frac{(1 + \beta) I - \beta [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F]}{(\beta - 1) I - (\beta - 2) [F'(x^{(k)})]^{-1} [x^{(k)}, y^{(k)}; F]} [F'(x^{(k)})]^{-1} F(y^{(k)}). \tag{17}
$$
3. For any \( a_1 \neq 0 \) and \( b_2 = 1 \),
\[
x^{(k+1)} = y^{(k)} - \frac{1}{a_1} \left[ \left( (1 - a_1) I + a_1 [F'(x^{(k)})]^{-1} x^{(k)}, y^{(k)}; F \right)^{-1} \left[ F'(x^{(k)}) \right]^{-1} F(y^{(k)}) 
+ \frac{1}{a_1} \left( (2a_1 - 1) I - [F'(x^{(k)})]^{-1} x^{(k)}, y^{(k)}; F \right) \right] \left[ F'(x^{(k)}) \right]^{-1} F(y^{(k)}).
\]

\[
(18)
\]

4 Numerical results

In this section we show the numerical behavior of the proposed methods. In the tests made, variable precision arithmetics has been used, with 4000 digits of mantissa (in the numerical tests for nonlinear equations) and 1000 digits of mantissa (in the numerical tests for systems of nonlinear equations) in MATLAB R2013a.

To check the behavior of the proposed methods with derivatives for solving nonlinear equations, we use the following elements of the family of methods obtained:

1. \textbf{MA1}: \( a_1 = 2 \) and \( b_2 = 1 \)
\[
x_{k+1} = y_k - \frac{f(x_k)^2 - 2 f(y_k)^2}{f(x_k) [f(x_k) - 2 f(y_k)]} f(y_k).
\]

2. \textbf{MB1}: \( a_1 = -1 \) and \( b_2 = 1 \)
\[
x_{k+1} = y_k - \frac{f(x_k)^2 + 3 f(x_k) f(y_k) + 6 f(y_k)^2}{f(x_k) [f(x_k) + f(y_k)]} f(y_k).
\]

3. \textbf{MC1}: \( a_1 = -1 \) and \( b_2 = 4 \)
\[
x_{k+1} = y_k - \frac{f(x_k)^2 - 8 f(y_k)^2}{f(x_k) [f(x_k) - 2 f(y_k)]} f(y_k).
\]

where \( y_k \) is Newton’s step. In these numerical experiments, we compare MA1, MB1 and MC1 with Newton’s method (NM), Ostrowski’s method (OM) (3), Chun’s method (CM) (4) and Jarratt’s method (JM) [11]

\[
y_k = x_k - \frac{2 f(x_k)}{3 f'(x_k)},
\]

\[
x_{k+1} = x_k - \frac{1}{2} \left( \frac{3 f'(x_k)}{f(x_k) - f(y_k) f'(x_k)} \right).
\]

The results obtained are shown in Table 1. In it and in Table 2, \( x_0 \) and \( \xi \) are the initial guess used and the approximate solution obtained, respectively.

The elements of the family of derivative-free methods that we are going to use are:
1. **MA2**: \( a_1 = 2 \) and \( b_2 = 1 \)

\[
x_{k+1} = y_k - \frac{f(x_k)^2 - 2f(y_k)^2}{f(x_k) [f(x_k) - 2f(y_k)]} \frac{f(y_k)}{f(z_k, x_k)}
\]

2. **MB2**: \( a_1 = -1 \) and \( b_2 = 1 \)

\[
x_{k+1} = y_k - \frac{f(x_k)^2 + 3f(x_k)f(y_k) + 6f(y_k)^2}{f(x_k) [f(x_k) + f(y_k)]} \frac{f(y_k)}{f(z_k, x_k)}
\]

3. **MC2**: \( a_1 = -1 \) and \( b_2 = 4 \)

\[
x_{k+1} = y_k - \frac{f(x_k)^2 - 8f(y_k)^2}{f(x_k) [f(x_k) - 2f(y_k)]} \frac{f(y_k)}{f(z_k, x_k)}
\]

where \( y_k = x_k - \frac{f(x_k)}{f(z_k, x_k)} \), \( f[z_k, x_k] = \frac{f(z_k) - f(x_k)}{z_k - x_k} \) and \( z_k = x_k + f(x_k)^2 \). In this case, we compare our schemes with Steffensen’s method (SM) [18], LZM [14] and CT4 [3] (with \( \gamma = 1 \), \( a = 1 \), \( b = 1 \), \( c = 1 \) and \( d = 0 \))

\[
y_k = x_k - \frac{\gamma f(x_k)^2}{f(z_k) - f(x_k)}, \quad z_k = x_k + \gamma f(x_k),
\]

\[
x_{k+1} = y_k - \frac{f(x_k) - f(y_k)}{f[z_k, x_k]} \frac{f(y_k)}{f(z_k, x_k)}
\]

The results obtained are shown in Table 2.

In Table 3, we show the results obtained by using the following elements of the family (12), for the following values of \( a_1 \) and \( b_2 \):

1. **MA3**: \( a_1 = 2 \) and \( b_2 = 1 \)

2. **MB3**: \( a_1 = -1 \) and \( b_2 = 1 \)

3. **MC3**: \( a_1 = -1 \) and \( b_2 = 4 \)
In order to not preserve the local order of convergence we use in our computations the divided difference operator introduced in [9]. Tables 1, 2 and 3 show, for each initial estimation and every method, the absolute of the difference between the first three iterations and the approximated computational order of convergence \( \rho \) (usually called ACOC), defined by Cordero and Torregrosa in [7]

\[
\rho = \frac{\ln(\|x^{(k+1)} - x^{(k)}\|/\|x^{(k)} - x^{(k-1)}\|)}{\ln(\|x^{(k)} - x^{(k-1)}\|/\|x^{(k-1)} - x^{(k-2)}\|)}.
\]

5 Concluding remarks

We have presented two family of iterative methods for solution of nonlinear equations with and without derivatives, respectively. In addition, by using the first family we obtain a class of iterative methods for finding the solution of nonlinear systems.

The numerical results obtained in Section 4 reverified the theoretical results. Summarizing, we can conclude that the novel iterative methods have a good performance for solving nonlinear equations and systems.

References


Table 1: Test functions and numerical results for methods with derivatives

| Method | $\rho$ | $|x_0 - x_1|$ | $|x_2 - x_1|$ | $|x_3 - x_2|$ |
|--------|-------|--------------|--------------|--------------|
| NM     | 2.0000 | 5.764694e-01 | 1.536345e-01 | 1.311421e-02 |
| JM     | 4.0000 | 4.220146e-01 | 1.239055e-02 | 3.902397e-09 |
| OM     | 4.0000 | 4.220863e-01 | 1.246232e-02 | 4.033923e-09 |
| CM     | 4.0000 | 6.274630e-01 | 2.169164e-01 | 9.226259e-04 |
| MA1    | 4.0000 | 5.247747e-01 | 1.150881e-01 | 6.253507e-05 |
| MB1    | 4.0000 | 8.886550e-01 | 4.676852e-01 | 1.134584e-02 |
| MC1    | 4.0000 | 8.328398e-01 | 4.111556e-01 | 1.206016e-02 |

| $f_2(x) = x^2 - \exp(x) - 3x + 2$, $x_0 = 0.8$ and $\xi \approx 0.257530285439861$ |
|--------|-------|--------------|--------------|--------------|
| NM     | 2.0000 | 5.476537e-01 | 5.181467e-03 | 2.524890e-06 |
| JM     | 4.0000 | 5.421618e-01 | 3.079593e-04 | 5.488364e-17 |
| OM     | 4.0000 | 5.423528e-01 | 1.169339e-04 | 8.444066e-19 |
| CM     | 4.0000 | 5.423548e-01 | 1.148681e-04 | 2.181127e-19 |
| MA1    | 4.0000 | 5.423554e-01 | 1.143362e-04 | 7.465357e-20 |
| MB1    | 4.0000 | 5.423554e-01 | 1.143362e-04 | 7.465357e-20 |
| MC1    | 4.0000 | 5.423569e-01 | 1.128023e-04 | 3.255984e-19 |

| $f_3(x) = \cos x - x$, $x_0 = 1$ and $\xi \approx 0.739085133215161$ |
|--------|-------|--------------|--------------|--------------|
| NM     | 2.0000 | 1.250987e-01 | 1.250987e-02 | 2.775753e-05 |
| JM     | 4.0000 | 2.060841e-01 | 7.340976e-05 | 6.833501e-19 |
| OM     | 4.0000 | 2.060834e-01 | 8.075989e-05 | 1.087899e-18 |
| CM     | 4.0000 | 2.060782e-01 | 1.566593e-04 | 4.132425e-17 |
| MA1    | 4.0000 | 2.067962e-01 | 1.870969e-04 | 9.352209e-18 |
| MB1    | 4.0000 | 2.067415e-01 | 1.733835e-04 | 7.172162e-17 |
| MC1    | 4.0000 | 2.060823e-01 | 2.325586e-04 | 3.264915e-16 |

| $f_4(x) = \cos x - x \exp x + x^2$, $x_0 = 0.5$ and $\xi \approx 0.639154096332008$ |
|--------|-------|--------------|--------------|--------------|
| NM     | 2.0000 | 1.552987e-01 | 1.594170e-02 | 2.029048e-04 |
| JM     | 4.0000 | 1.392115e-01 | 5.742635e-05 | 1.721447e-18 |
| OM     | 4.0000 | 1.392052e-01 | 5.112554e-05 | 1.087899e-18 |
| CM     | 4.0000 | 1.403053e-01 | 1.151178e-03 | 3.631054e-12 |
| MA1    | 4.0000 | 1.397552e-01 | 6.011518e-04 | 1.445560e-13 |
| MB1    | 4.0000 | 1.407044e-01 | 1.550261e-03 | 1.469465e-11 |
| MC1    | 4.0000 | 1.414053e-01 | 2.251230e-03 | 1.022163e-10 |
Iterative methods for nonlinear systems

Table 2: Test functions and numerical results for methods without derivatives

\[ f_1(x) = \sin x - x^2 + 1, \ x_0 = 1 \text{ and } \xi \approx 1.409624004002596 \]

| Method | \( \rho \) | \( |x_0 - x_1| \) | \( |x_2 - x_1| \) | \( |x_3 - x_2| \) |
|--------|---------|----------------|----------------|----------------|
| ST     | 2.0000  | 3.120784e-01  | 8.967566e-02  | 7.812763e-03  |
| LJM    | 4.0000  | 4.092900e-01  | 3.339581e-04  | 2.213383e-15  |
| CT4    | 4.0000  | 4.153571e-01  | 5.733140e-03  | 5.174523e-10  |
| OM2    | 4.0000  | 4.696449e-01  | 6.005902e-02  | 3.811266e-05  |
| CM2    | 4.0000  | 4.438749e-01  | 3.425276e-02  | 1.871142e-06  |
| MA2    | 4.0000  | 4.567599e-01  | 4.714596e-02  | 1.005694e-05  |
| MB2    | 4.0000  | 4.409219e-01  | 3.129896e-02  | 1.050665e-06  |
| MC2    | 4.0000  | 4.181049e-01  | 8.480881e-03  | 2.837109e-09  |

\[ f_2(x) = x^2 - \exp(x) - 3x + 2, \ x_0 = 0.8 \text{ and } \xi \approx 0.257530285439861 \]

| Method | \( \rho \) | \( |x_0 - x_1| \) | \( |x_2 - x_1| \) | \( |x_3 - x_2| \) |
|--------|---------|----------------|----------------|----------------|
| ST     | 2.0000  | 4.561828e-01  | 8.424288e-02  | 2.042960e-03  |
| LJM    | 4.0000  | 5.398373e-01  | 2.632450e-03  | 1.783321e-12  |
| CT4    | 4.0000  | 5.401203e-01  | 2.349388e-03  | 1.061875e-12  |
| OM2    | 4.0000  | 1.474585e-02  | 9.742210e-03  | 7.402434e-03  |
| CM2    | 4.0000  | 2.553675e-01  | 2.887999e-01  | 1.697744e-03  |
| MA2    | 4.0000  | 2.503108e-01  | 1.363203e+00  | 3.397303e-24  |
| MB2    | 4.0000  | 2.291839e-01  | 3.176765e-01  | 4.390681e-03  |
| MC2    | 4.0000  | 5.254809e-01  | 1.698881e-02  | 1.270682e-08  |

\[ f_3(x) = \cos x - x, \ x_0 = 1 \text{ and } \xi \approx 0.739085133215161 \]

| Method | \( \rho \) | \( |x_0 - x_1| \) | \( |x_2 - x_1| \) | \( |x_3 - x_2| \) |
|--------|---------|----------------|----------------|----------------|
| ST     | 2.0000  | 2.719896e-01  | 1.105661e-02  | 1.816626e-05  |
| LJM    | 4.0000  | 2.608838e-01  | 3.108491e-05  | 4.065649e-21  |
| CT4    | 4.0000  | 2.608488e-01  | 6.605575e-05  | 2.209748e-19  |
| OM2    | 4.0000  | 2.613137e-01  | 3.988060e-04  | 2.806318e-15  |
| CM2    | 4.0000  | 2.609973e-01  | 8.244500e-05  | 3.138517e-18  |
| MA2    | 4.0000  | 2.611555e-01  | 2.406255e-04  | 2.998867e-16  |
| MB2    | 4.0000  | 2.609328e-01  | 1.796954e-05  | 5.959916e-21  |
| MC2    | 4.0000  | 2.606810e-01  | 2.339160e-04  | 7.421940e-17  |

\[ f_4(x) = \cos x - x \exp x + x^2, \ x_0 = 0.5 \text{ and } \xi \approx 0.639154096332008 \]

| Method | \( \rho \) | \( |x_0 - x_1| \) | \( |x_2 - x_1| \) | \( |x_3 - x_2| \) |
|--------|---------|----------------|----------------|----------------|
| ST     | 2.0000  | 1.214519e-01  | 1.736088e-02  | 3.411983e-04  |
| LJM    | 4.0000  | 1.390186e-01  | 1.355071e-04  | 1.339909e-16  |
| CT4    | 4.0000  | 1.392274e-01  | 7.333113e-05  | 8.391049e-18  |
| OM2    | 4.0000  | 1.387878e-01  | 3.663325e-04  | 6.237591e-14  |
| CM2    | 4.0000  | 1.388304e-01  | 3.236745e-04  | 1.693256e-14  |
| MA2    | 4.0000  | 1.388091e-01  | 3.450035e-04  | 3.546717e-14  |
| MB2    | 4.0000  | 1.388425e-01  | 3.115684e-04  | 1.000584e-14  |
| MC2    | 4.0000  | 1.388731e-01  | 2.810164e-04  | 2.372628e-15  |
Table 3: Test functions and results for nonlinear systems $F_1$ and $F_2$

$$F_1(x_1, x_2) = (\exp x_1 \exp x_2 + x_1 \cos x_2, x_1 + x_2 - 1)$$

$x^{(0)} = (3, -2)$

and $\xi_1 \approx 3.4675009642402$, $\xi_2 \approx -2.4675009642402$

| Method | $\rho$ | $||x^{(1)} - x^{(1)}||$ | $||x^{(2)} - x^{(1)}||$ | $||x^{(3)} - x^{(2)}||$ |
|--------|--------|--------------------------|--------------------------|--------------------------|
| NM     | 2.0000 | 6.611462e-01             | 4.423036e-03             | 3.446250e-06             |
| TM     | 4.0000 | 6.652901e-01             | 2.825451e-04             | 1.416637e-12             |
| JM     | 4.0000 | 6.652161e-01             | 3.565442e-04             | 4.535142e-16             |
| OM3    | 4.0000 | 6.653428e-01             | 2.299389e-04             | 7.883163e-17             |
| CM3    | 4.0000 | 6.653421e-01             | 2.305983e-04             | 1.426857e-16             |
| MA3    | 4.0000 | 6.653424e-01             | 2.302686e-04             | 1.05782e-16              |
| MB3    | 4.0000 | 6.653419e-01             | 2.307601e-04             | 1.588653e-16             |
| MC3    | 4.0000 | 6.653414e-01             | 2.312578e-04             | 2.079935e-16             |

$$F_2(x_1, x_2, x_3, x_4) = (x_1x_3 + x_4(x_2 + x_3), x_1x_3 + x_4(x_1 + x_3), x_1x_2 + x_4(x_1 + x_2), x_1x_2 + x_1x_3 + x_2x_3 - 1),$$

$x_0 = (1, 1, 1, -0.5)$

and $\xi_1 = \xi_2 = \xi_3 = 0.5773502691896257$, $\xi_4 = -0.2886751345948129$

| Method | $\rho$ | $||x^{(1)} - x^{(1)}||$ | $||x^{(2)} - x^{(1)}||$ | $||x^{(3)} - x^{(2)}||$ |
|--------|--------|--------------------------|--------------------------|--------------------------|
| NM     | 2.0000 | 6.00925e-01              | 1.502313e-01             | 1.073081e-02             |
| TM     | 4.0000 | 7.010794e-01             | 6.077163e-02             | 9.159487e-05             |
| JM     | 4.0000 | 7.511565e-01             | 1.078612e-02             | 1.469773e-09             |
| OM3    | 4.0000 | 7.511565e-01             | 1.078612e-02             | 1.469773e-09             |
| CM3    | 4.0000 | 7.344642e-01             | 2.747820e-02             | 2.877612e-07             |
| MA3    | 4.0000 | 7.428103e-01             | 1.913226e-02             | 4.194000e-08             |
| MB3    | 4.0000 | 7.320795e-01             | 2.986264e-02             | 4.737541e-07             |
| MC3    | 4.0000 | 7.177718e-01             | 4.416761e-02             | 3.246905e-06             |
Iterative methods for nonlinear systems

Table 4: Test functions and results for nonlinear systems $F_3$ and $F_4$

$F_3(x_1, x_2) = (x_1^2 - x_1 - x_2^2 - 1, \sin x_1 + x_2), \quad x_0 = (-0.15, -0.15) \quad \text{and} \quad \xi_1 \approx -0.8452567390376772, \; \xi_2 \approx -0.7481414932526368$

| Method | $\rho$ | $||x^{(1)} - x^{(1)}||$ | $||x^{(2)} - x^{(1)}||$ | $||x^{(3)} - x^{(2)}||$ |
|--------|--------|------------------------|------------------------|------------------------|
| NM     | 2.0000 | 1.190707e+00           | 2.832280e-01           | 9.495321e-03           |
| TM     | 4.0000 | 1.007175e+00           | 9.519725e-02           | 1.114063e-03           |
| JM     | 4.0000 | 8.813431e-01           | 3.580364e-02           | 7.498626e-08           |
| OM3    | 4.0000 | 8.827556e-01           | 3.438949e-02           | 6.051406e-08           |
| CM3    | 4.0000 | 6.896321e-01           | 2.324527e-01           | 2.701063e-04           |
| MA3    | 4.0000 | 7.858270e-01           | 1.332668e-01           | 2.713320e-05           |
| MB3    | 4.0000 | 6.772478e-01           | 2.492809e-01           | 2.096363e-04           |
| MC3    | 4.0000 | 5.011335e-01           | 4.326374e-01           | 2.746111e-03           |

$F_4(x_1, x_2, x_3) = (x_1^2 + x_2^2 + x_3^2 - 9, \; x_1 x_2 x_3 - 1, \; x_1 + x_2 - x_3^2), \quad x_0 = (2, -1.5, -0.5) \quad \text{and} \quad \xi_1 \approx 2.140258122005175, \; \xi_2 \approx -2.090294642255235, \; \xi_3 \approx -0.023525121071319$

| Method | $\rho$ | $||x^{(1)} - x^{(1)}||$ | $||x^{(2)} - x^{(1)}||$ | $||x^{(3)} - x^{(2)}||$ |
|--------|--------|------------------------|------------------------|------------------------|
| NM     | 2.0000 | 9.711265e-01           | 2.829497e-01           | 3.702405e-02           |
| TM     | 4.0000 | 3.201946e-01           | 3.551275e-01           | 9.073222e-02           |
| JM     | 4.0000 | 6.994381e-01           | 3.669324e-02           | 8.21706e-08            |
| OM3    | 4.0000 | 6.990541e-01           | 3.625677e-02           | 7.823164e-08           |
| CM3    | 4.0000 | 1.434861e+00           | 7.796443e-01           | 2.067793e-02           |
| MA2    | 4.0000 | 1.061901e+00           | 4.158193e-01           | 2.174787e-03           |
| MB2    | 4.0000 | 3.846125e+00           | 2.693861e+00           | 6.742142e-01           |
| MC2    | 4.0000 | 2.190413e+00           | 1.432442e+00           | 1.417313e-01           |
On generalization of the variants of Newton’s method for solving nonlinear equations

Alicia Cordero\textsuperscript{1} and Juan R. Torregrosa\textsuperscript{1}

\textsuperscript{1} Instituto de Matemáticas Multidisciplinar, Universitat Politècnica de València
Camino de Vera, s/n, 46022-Valencia, Spain
emails: acordero@mat.upv.es, jrtorre@mat.upv.es

Abstract

In this paper, an unified point of view that include the most of one-point Newton-type iterative methods for solving nonlinear equations is introduced. A simple idea to design iterative methods with quadratic or cubic convergence is also described. In addition, several numerical examples are given to illustrate and compare different known methods and some ones introduced by using this unifying idea.

Key words: Nonlinear equation, one-point method, Newton’s method, order of convergence, efficiency index, optimal order.

MSC 2000: 65H05

1 Introduction

Solving nonlinear equations is a classical problem which has interesting applications in several branches of sciences and engineering. Many optimization problems such as searching for a local minimizer of function [16], the potential equations in the transonic regime of dense gases in gasdynamics [20] and the boundary value problems encountered in kinetic theory of gases, elasticity [12] and problems in other applied areas can be reduced to nonlinear equations. In general, to compute their roots we must drawn on to iterative methods.

This paper is concerned with iterative methods to find a simple root $\alpha$ of a nonlinear equation $f(x) = 0$, where $f$ is a real function $f : I \subseteq \mathbb{R} \rightarrow \mathbb{R}$, defined in an open interval $I$. There are many iterative methods such as Newton’s method, Halley’ and super-Halley’s schemes, Chebyshev’s method, etc. and their variants (see [19] and the references therein).
Newton’s method is the best known and probably the most used algorithm for solving $f(x) = 0$. It is given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots$$

which converges quadratically in some neighborhood of $\alpha$, that is, there exists a positive constant $C$ such that

$$\lim_{k \to \infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = C.$$ 

More generally, for the sequence $\{x_k\}_{k=0}^\infty$ generated by an iterative method, if there exist positive constants $C$ and $p$ such that

$$\lim_{k \to \infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = C,$$

then the method is said to converge to $\alpha$ with the local order of convergence $p$.

Commonly, the efficiency of an iterative method is measured by the efficiency index, defined by Ostrowski [17] as $p^{1/d}$, where $p$ is the order of convergence and $d$ is the number of functional evaluations per iteration. Kung and Traub conjectured in [11] that the order of convergence of any iterative method, without memory, cannot exceed the bound $2^{d-1}$, called the optimal order. The efficiency index of Newton’s method is 1.414 and it is an optimal scheme.

The construction of numerical methods for solving nonlinear equations is an interesting task, which has attracted the attention of many authors for more than three centuries. These schemes can be classified in two big families: one-point and multipoint schemes, depending on if each iteration is formed by one step or by several ones, respectively. The classical methods mentioned before are one-point schemes. During the last years, numerous papers devoted to design one-point iterative methods for solving nonlinear equations, $f(x) = 0$, have appeared. These methods are developed from the classical algorithms by using Taylor interpolating polynomials, quadrature rules or some other techniques. However, due to the limitations and restrictions of the one-point iterative schemes, multipoint methods appeared in the literature. We can find a good survey of these multipoint schemes in [18].

In this work, we provide a systematic treatment of the one-point iterative methods by using the weight function procedure. We can include, under this unified point of view, all the one-point known methods, as far as we know, of order two and three.

The rest of the paper is organized as follows: in Section 2 we show many known one-point methods or classes of schemes of order two and three, describing their iterative expression as a modified Newton method with a weight function $H$, depending on a particular variable. In Section 3 we analyze the conditions that function $H$ must satisfy in order to obtain iterative methods of order two and three. Several numerical examples are given in Section 4 to illustrate and compare the efficiency of the methods considered in the paper.
2 Overview of classical and recent methods

In this section we are going to show some of the classical and more recent methods or families of schemes for finding a root of the equation $f(x) = 0$, with second and third order of convergence. Their iterative expressions will be described as a modified Newton method with a weight function $H$, depending on variable $u = \frac{f(x)}{f'(x)}$ or variable $w = \frac{f(x) f''(x)}{f'(x)^2}$, which is called degree of logarithmic convexity.

Kanwar and Tomar proposed in [9] the following parametric family of second order iterative methods

\begin{equation}
    x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) + \beta f(x_k)} = x_k - \frac{f'(x_k)}{f'(x_k) + \beta f(x_k)} \frac{f(x_k)}{f'(x_k)} = x_k - \frac{1}{1 + \beta u(x_k)} \frac{f(x_k)}{f'(x_k)},
\end{equation}

where $\beta$ is a parameter, derived by expanding a particular function in Taylor series. Let us observe that for $\beta = 0$ we obtain Newton’s scheme.

From this idea, Kou and Li in [10] described the bi-parametric family of methods of order two

\begin{equation}
    x_{k+1} = x_k - \left(1 + \frac{\lambda h(x_k)}{1 + \beta h(x_k)}\right) h(x_k),
\end{equation}

where $\lambda, \beta$ are parameters and $h(x_k) = \frac{f(x_k)}{f'(x_k) + \beta f(x_k)}$. We can transform this method in

\begin{equation}
    x_{k+1} = x_k - \left(1 + \frac{\lambda u(x_k)}{(1 + \beta u(x_k))(1 + 2\beta u(x_k))}\right) \frac{f(x_k)}{f'(x_k)}.
\end{equation}

In [14], Noor presented the following method with quadratic convergence:

\begin{equation}
    x_{k+1} = x_k - \frac{2 f(x_k)}{f'(x_k) \pm \sqrt{f'(x_k)^2 + 4\beta^3 f^3(x_k)}} = x_k - \frac{2}{1 \pm \sqrt{1 + 4\beta^3 f(x_k)^3 / f'(x_k)^2}} \frac{f(x_k)}{f'(x_k)},
\end{equation}

obtained by using Taylor polynomials along with an auxiliary equation.

All these methods are optimal in the sense of Kunt-Traub conjecture.

On the other hand, there are many third-order iterative methods requiring the evaluation of the second derivative of $f$. The most of the well-known one-point cubically convergent methods belong to the one-parameter class, called Chebyshev-Halley family

\begin{equation}
    x_{k+1} = x_k - \left(1 + \frac{1}{2} \frac{f(x_k) f''(x_k)}{f'(x_k)^2 - \beta f(x_k) f''(x_k)}\right) \frac{f(x_k)}{f'(x_k)} = x_k - \left(1 + \frac{1}{2} \frac{w(x_k)}{1 - \beta w(x_k)}\right) \frac{f(x_k)}{f'(x_k)}.
\end{equation}
This family includes Chebyshev’s method for $\beta = 0$, Halley’s scheme for $\beta = 1/2$, super-Halley’s method for $\beta = 1$ and Newton’s method when $\beta$ tends to $\pm \infty$.

Fang et al. [6] obtained the third order method

$$x_{k+1} = x_k - \frac{2f(x_k)}{f'(x_k) \pm \sqrt{f'(x_k)^2 - 2f(x_k)f''(x_k)}} = x_k - \frac{2}{1 \pm \sqrt{1 - 2w(x_k)f'(x_k)}} f(x_k),$$

(5)

by expanding $f$ in Taylor series about the origin, dropping the third and higher order terms, and solving the obtained quadratic equation.

Abbasbandy [1] and Chun [3] proposed and studied several one-step iterative methods by using the decomposition technique of Adomian. Specifically, Abbasbandy obtained with this technique the third order Chebyshev’s method, which was also obtained by Noor et al. [15] with another decomposition technique which does not involve the derivative of the Adomian polynomials. They designed the following predictor-corrector method with cubic convergence:

Predictor-step

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Corrector-step

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{(y_k - x_k)^2}{2f'(x_k)f''(x_k)} - \frac{(y_k + z_k - x_k)^2}{2f'(x_k)f''(x_k)},$$

where $z_k = -\frac{(y_k - x_k)^2}{2f'(x_k)} f''(x_k)$.

This method can be rewritten in the following form

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \left( 1 + \frac{1}{2} \frac{f(x_k)f''(x_k)}{f'(x_k)^2} \left( 1 + \frac{f(x_k)f''(x_k)}{2f'(x_k)^2} + \frac{f(x_k)^2f''(x_k)^2}{4f'(x_k)^4} \right) \right)$$

$$= x_k - \left( 1 + \frac{1}{2} w(x_k) + \frac{1}{2} w(x_k)^2 + \frac{1}{4} w(x_k)^3 \right) \frac{f(x_k)}{f'(x_k)},$$

(6)

In [8], Hansen and Patrick presented a parametric family of iterative methods, with order of convergence three, whose iterative expression is

$$x_{k+1} = x_k - \frac{(\lambda + 1)f(x_k)}{\lambda f'(x_k) \pm \sqrt{f'(x_k)^2 - (\lambda + 1)f(x_k)f''(x_k)}} = x_k - \frac{\lambda + 1}{\lambda \pm \sqrt{1 - (\lambda + 1)w(x_k)}} \frac{f(x_k)}{f'(x_k)},$$

(7)

In particular, if $\lambda = 0$, we obtain the iterative method
\[ x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \]

studied by Ostrowski in [17]. On the other hand, if \( \lambda = 1 \) we obtain the known Euler’s method.

Another parametric family of iterative methods with order of convergence three was developed by Gutiérrez and Hernández in [7]. Specifically, they described the family of methods

\[ x_{k+1} = x_k - \left( 1 + \frac{f(x_k)f''(x_k)}{2(f'(x_k)^2 - \lambda f(x_k)f''(x_k))} \right) \frac{f(x_k)}{f'(x_k)} = x_k - \left( 1 + \frac{w(x_k)}{2(1 - \lambda w(x_k))} \right) \frac{f(x_k)}{f'(x_k)}, \]

where \( \lambda \) is a real parameter. For \( \lambda = 1 \) super-Halley’s method is obtained.

More recently, Chun and Kim, by using a geometric approach based on the circle of curvature, constructed in [4] a new iterative method of order three, whose expression is

\[ x_{k+1} = x_k - \frac{f'(x_k)f(x_k)(f''(x_k)f(x_k) + 2 + 2f'(x_k)^2)}{2f'(x_k)^2(1 + f'(x_k)^2)f(x_k)f''(x_k)}. \]

Algebraic manipulations allow us to transform this iterative expression in

\[ x_{k+1} = x_k - \frac{w(x_k) + 2s(x_k)}{2s(x_k) - w(x_k)/f'(x_k)^2} \frac{f(x_k)}{f'(x_k)}, \]

where \( s(x_k) = 1 + \frac{1}{f'(x_k)^2} \).

Finally, we present the parametric family developed by Neta and Scott in [13], whose iterative formula is

\[ x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f(x_k)^2 f''(x_k)}{2f'(x_k)^3 - Af(x_k)f''(x_k)} = x_k - \left( 1 + \frac{w(x_k)}{2 - Aw(x_k)} \right) \frac{f(x_k)}{f'(x_k)}, \]

where \( A \) is a parameter. Upon choosing \( A = 1 \) we have Halley’s method. The choice \( A = 0 \) yields the well known Chebyshev’s method. This latter scheme is also a special case of Hansen and Patrick’s family (7) with \( \lambda = 1 \). The choice \( A = 2 \) gives the BSC method described by Basto et al. in [2].

As we have said, all these methods have order of convergence three and none of them is optimal in the sense of Kung-Traub conjecture.

### 3 Convergence analysis

In this section we present a unified and simple result for demonstrating the order of convergence of the methods described in Section 2. In fact, it is based on weight function
Variants of Newton’s method

procedure and allows us to establish the order of convergence of all one-point methods published, as far as we know, when their order is two or three. In addition, this result provides an easy procedure to design iterative methods with order of convergence two or three.

We can observe that the methods of order two presented in Section 2 have the general iterative expression

\[ x_{k+1} = x_k - H(u(x_k)) \frac{f(x_k)}{f'(x_k)}, \]  

(12)

where \( H(u) \) is a function of variable \( u(x) = \frac{f(x)}{f'(x)} \). For iterative methods described in the form (12) we can establish the following result.

**Theorem 1** Let \( f : I \subseteq \mathbb{R} \to \mathbb{R} \) be a real function with second derivative in \( I \). Let \( \alpha \in I \) a simple root of \( f(x) = 0 \). If we choose an initial guess close enough to \( \alpha \) and a sufficiently differentiable function \( H(u) \) such that \( H(0) = 1 \), then the methods described by (12) converge to \( \alpha \) with quadratic order of convergence, being their error equation

\[ e_{k+1} = (-H'(0) + c_2) e_k^2 + O(e_k^3), \]

where \( c_j = \frac{1}{j!} \frac{f^{(j)}(\alpha)}{f'(\alpha)}, j = 2, 3, \ldots \) and \( e_k = x_k - \alpha \).

In Table 1 we describe the function \( H(u) \) that guarantees the quadratic convergence of the methods of order two described in Section 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Function ( H(u) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton</td>
<td>( H(u) = 1 )</td>
</tr>
<tr>
<td>Kanwar-Tomar (1)</td>
<td>( H(u) = \frac{1}{1 + \beta u}, \beta \text{ parameter} )</td>
</tr>
<tr>
<td>Kou-Li (2)</td>
<td>( H(u) = 1 + \frac{\lambda u}{(1 + \beta u)(1 + 2\beta u)}, \lambda, \beta \text{ parameters} )</td>
</tr>
<tr>
<td>Noor (3)</td>
<td>( H(x) = \frac{2}{1 \pm \sqrt{1 + 4\beta^2 f(x)/x^2}}, \beta \text{ parameter} )</td>
</tr>
</tbody>
</table>

Table 1: Weight functions \( H(u) \) that express different quadratic methods

On the other hand, the methods of order three presented in Section 2 have the general iterative expression

\[ x_{k+1} = x_k - G(w(x_k)) \frac{f(x_k)}{f'(x_k)^2}, \]  

(13)

where \( G(w) \) is a function of variable \( w(x) = f(x)f''(x)/f'(x)^2 \). For iterative methods described in the form (13) we can establish the following result.
Theorem 2 Let us assume that $f(x)$ and $G(w)$ are sufficiently differentiable functions and $f(x)$ has a simple zero $\alpha \in I$. If the initial estimation $x_0$ is close enough to $\alpha$ and function $G(w)$ satisfies $G(0) = 1$ and $G'(0) = 1/2$, then the methods described by (13) converge to $\alpha$ with cubic order of convergence, being their error equation

$$e_{k+1} = -2((-1 + G''(0))c_2^2 + (1/2)c_3)e_k^3 + O(e_k^4),$$

where $c_j = \frac{1}{j!} \frac{f^{(j)}(\alpha)}{f'(\alpha)^j}$, $j = 2, 3, \ldots$ and $e_k = x_k - \alpha$.

In Table 2 we describe the functions $G(w)$ that guarantee the cubic convergence of the methods described in Section 2, since they satisfy the conditions of Theorem 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Function $G(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halley</td>
<td>$G(w) = \frac{2}{2 - w}$</td>
</tr>
<tr>
<td>super-Halley</td>
<td>$G(w) = \frac{w - 2}{2(w - 1)}$</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>$G(w) = 1 + \frac{1}{2}w$</td>
</tr>
<tr>
<td>Fang et al. (5)</td>
<td>$G(w) = \frac{2}{1 \pm \sqrt{1 - w}}$</td>
</tr>
<tr>
<td>Noor et al. (6)</td>
<td>$G(w) = 1 + \frac{1}{2}w + \frac{1}{4}w^2 + \frac{1}{4}w^3$</td>
</tr>
<tr>
<td>Hansen-Patrick (7)</td>
<td>$G(w) = \frac{\lambda + 1}{\lambda \pm \sqrt{1 - (\lambda + 1)w}}$, $\lambda$ parameter</td>
</tr>
<tr>
<td>Euler</td>
<td>$G(w) = \frac{2}{1 \pm \sqrt{1 - 2w}}$</td>
</tr>
<tr>
<td>Ostrowski (8)</td>
<td>$G(w) = \frac{2}{\sqrt{1 - w}}$</td>
</tr>
<tr>
<td>Gutiérrez-Hernández (9)</td>
<td>$G(w) = 1 + \frac{w}{2(1 - \lambda w)}$, $\lambda$ parameter</td>
</tr>
<tr>
<td>Chun-Kim (10)</td>
<td>$G(w) = \frac{w + 2s(x)}{2s(x) - w/f'(x)^2}$, $s(x) = 1 + \frac{1}{f'(x)^2}$</td>
</tr>
<tr>
<td>Neta-Scott (11)</td>
<td>$G(w) = 1 + \frac{2 - 4w}{2(1 - w)}$, $A$ parameter</td>
</tr>
<tr>
<td>Basto et al.</td>
<td>$G(w) = 1 + \frac{2 - 4w}{2(1 - w)}$</td>
</tr>
</tbody>
</table>

Table 2: Weight functions $G(w)$ that express different cubic methods
4 Numerical results

From Theorem 2, it is easy to design iterative methods that converge cubically. It is sufficient to choose functions $G(w)$ such that $G(0) = 1$ and $G'(0) = 1/2$. We can select functions $G(w)$ more simple than the ones in Table 2. For example, by taking $G(w) = e^{w/2}$ we obtain the iterative method

$$x_{k+1} = x_k - e^{\frac{f(x_k)f''(x_k)}{f'(x_k)^2}} f(x_k)$$

(14)

and with $G(w) = w^2 + w/2 + 1$ we obtain

$$x_{k+1} = x_k - \left( \frac{f(x_k)^2 f''(x_k)}{f'(x_k)^4} + \frac{f(x_k) f''(x_k)}{2 f'(x_k)} + 1 \right) \frac{f(x_k)}{f'(x_k)}.$$

(15)

Both methods have order of convergence three.

Now, we give numerical examples for comparing the effectiveness of some methods introduced in Section 2. Specifically, we compare Newton’, Chebyshev’, Ostrowski’ (8) and Chun-Kim’s method (10) with the methods described by (14) and (15), which are denoted in Table 3 by M1 and M2, respectively.

We use the following test functions.

- $f_1(x) = \cos x - x, \quad \alpha \approx 0.739085$,
- $f_2(x) = \sin^2 x - x^2 + 1, \quad \alpha \approx 1.404492$,
- $f_3(x) = xe^{x^2} - \sin^2 x + 3 \cos x + 5, \quad \alpha \approx -1.207648$,
- $f_4(x) = \sin x + x \cos x, \quad \alpha = 0$,
- $f_5(x) = x^2 e^{x^2} - \sin^2 x + x, \quad \alpha = 0$,
- $f_6(x) = (x - 1)^3 - 1, \quad \alpha = 2$,
- $f_7(x) = \frac{x^2 - 1}{x^2 + 1} + 1, \quad \alpha = 0$.

Numerical computations have been carried out using variable precision arithmetic, with 1000 digits, in MATLAB 7.13. The stopping criterion used is $|x_{k+1} - x_k| + |f(x_{k+1})| < 10^{-100}$. Therefore, we check that the sequence $\{x_k\}$ converges and that its limit is a solution of the nonlinear equation $f(x) = 0$. For every method and test function, we calculate the number of iterations, the value of $\text{incr} = |f(x_{k+1})|$ for the last iteration and the computational order of convergence ACOC, approximated by (see [5])

$$p \approx \text{ACOC} = \frac{\ln(|x_{k+1} - x_k| / |x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}| / |x_{k-1} - x_{k-2}|)}.$$  

(16)
The value of $ACOC$ that appears in Table 3 is the last coordinate of vector (16) when the variation between its values is small. In some cases, the approximated order of convergence is not stable and it is not shown in the table.

Table 3 summarizes several results obtained by using the mentioned methods (Newton, Chebyshev, Ostrowski, Chun-Kim, M1 and M2) in order to estimate a root of nonlinear functions from $f_1$ to $f_7$. For every function we specify the initial estimation $x_0$, the number of iterations, the value of function $f$ in the last iteration and the value of $ACOC$.

Table 3: Numerical results for different methods and several test functions

We can observe that, in general, the number of iterations and the value of $|f(x_k+1)|$ of all the methods is in a similar range, taking into account that Newton's method has order two while the rest of the schemes have order three. The computational order of convergence confirms the theoretical results, but we need to introduce some comments. For function $f_4(x)$ Newton’s method has order of convergence three, like the other methods, since $f'_4(\alpha) = 0$. In addition, as $f''_5(\alpha) = f''_5(\alpha) = 0$, all convergent methods have order four. For $f_7(x)$ all methods converge linearly since the zero of this function is not simple, it has multiplicity 2.
Acknowledgements

This work has been partially supported by Ministerio de Ciencia y Tecnología MTM2011-28636-C02-02.

References


Adjoint triples versus extended-order algebras

M Eugenia Cornejo\textsuperscript{1}, Jesús Medina\textsuperscript{2} and Eloisa Ramírez-Poussa\textsuperscript{1}

\textsuperscript{1} Department of Statistic and O.R., University of Cádiz. Spain
\textsuperscript{2} Department of Mathematics, University of Cádiz. Spain
emails: mariaeugenia.cornejo@uca.es, jesus.medina@uca.es, eloisa.ramirez@uca.es

Abstract

Adjoint triples and extended-order algebras provide general structures which have been developed to increase the flexibility in the framework where they are used. This paper relates adjoint triples to the operators considered in extended-order algebras. Moreover, the comparison between adjoint negations and the negations introduced in extended-order algebras is given.

Key words: Adjoint triples, extended-order algebras, negation operators.

1 Introduction

Adjoint triples are general operators which have been developed to increase the flexibility of the framework in which they are used, such as Logic Programming \cite{14}, Fuzzy Formal Concept Analysis \cite{13}, Fuzzy Relation Equations \cite{7} and Rough Set Theory \cite{5}. The main advantage of adjoint triples is the fact that their conjunctors are neither required to be commutative nor associative. These triples have widely been studied in several papers \cite{2, 4}.

In addition, from adjoint triples it can be defined two negation operators. Negation operators are widely studied in \cite{6, 9, 16, 17} and are very useful in fuzzy logic and logic programming, since these operators should be considered in order to simulate better the behavior of human brain. This paper considers adjoint negations obtained from an adjoint triple which generalize the pair of weak negations introduced by Georgescu and Popescu in \cite{11}.

The main goal of this paper is the comparison of w-eo algebras with the operators mentioned previously. An initial comparison was shown in \cite{3}, now we present a deeper
study including a comparative survey with adjoint negations and adjoint pairs. Extended-order algebras were introduced by C. Guido and P. Toto. In [6], an interesting study of extended-order algebras was developed and new operators in these algebras corresponding to the negation connectives were introduced. The comparison of the former with the adjoint negations is also presented.

The structure of this paper is the following: Section 2 establishes basic notions of adjoint triples and adjoint negations. The comparison between adjoint triples and extended-order algebras is made in Section 3 and the corresponding comparison among negation operators is given in Section 4. Finally, the paper finishes with some conclusions and prospects for future work.

2 Adjoint triples

This section recalls the definition of adjoint triple as well as some interesting properties derived from these operators. Moreover, from an adjoint triple a generalization of residuated negation [1, 10, 15] is defined, adjoint negations.

Definition 1 Let \((P_1, \leq_1), (P_2, \leq_2), (P_3, \leq_3)\) be posets and \&: \(P_1 \times P_2 \rightarrow P_3\), \(\vee': P_3 \times P_2 \rightarrow P_1\), \(\cap\): \(P_3 \times P_1 \rightarrow P_2\) be mappings, then \((\&, \vee', \cap)\) is an adjoint triple with respect to \(P_1, P_2, P_3\) if \&, \(\vee', \cap\) satisfy the adjoint property:

\[ x \leq_1 z \vee y \iff x \& y \leq_3 z \iff y \leq_2 z \cap x \]

where \(x \in P_1, y \in P_2\) and \(z \in P_3\).

The following monotony properties are obtained straightforwardly from adjoint property.

Proposition 1 If \((\&, \vee', \cap)\) is an adjoint triple w.r.t. the posets \((P_1, \leq_1), (P_2, \leq_2)\) and \((P_3, \leq_3)\), then

1. \& is order-preserving on both arguments.

2. \(\vee', \cap\) are order-preserving on the first argument and order-reversing on the second argument.

The boundary conditions satisfied by an adjoint triple with respect to three bounded posets are shown in next result.

Proposition 2 Given an adjoint triple \((\&, \vee', \cap)\) with respect to the bounded partially ordered sets \((P_1, \leq_1, \bot_1, \top_1), (P_2, \leq_2, \bot_2, \top_2)\) and \((P_3, \leq_3, \bot_3, \top_3)\), the following boundary conditions hold.
1. \( \perp_1 \land y = \perp_3 \) and \( x \land \perp_2 = \perp_3 \), for all \( x \in P_1, y \in P_2 \).

2. \( z \land \perp_1 = \top_2 \) and \( z \land \perp_2 = \top_1 \), for all \( z \in P_3 \).

The result below states that given the conjunctor of an adjoint triple, its residuated implications are unique.

**Proposition 3** Given a conjunctor \( \land \), if there exist its residuated implications \( \lor \) and \( \land \), they are unique.

Regarding an adjoint triple with respect to a lower bounded poset \( (P, \leq, \perp) \), adjoint negations are defined as follows.

**Definition 2** Given an adjoint triple \( (\land, \lor, \land) \) with respect to a lower bounded poset \( (P, \leq, \perp) \), the mappings \( n_s, n_n : P \to P \) defined, for all \( x, y \in P \), as
\[
    n_s(y) = \perp_1 \lor y \quad n_n(x) = \perp_2 \land x
\]
are called adjoint negations on \( P \).

The operators \( n_s \) and \( n_n \) satisfying that \( x = n_n(n_s(x)) = n_s(n_n(x)) \), for all \( x \in P \), are called strong adjoint negations on \( P \).

## 3 Extended-order algebras

Following the same motivation of adjoint triples, in order to reduce the mathematical requirements of the basic operators used for computation in the considered framework, other operators have been used in several settings. One of these operators is extended-order algebras [12]. An initial study where the relation between adjoint triples and extended-order algebras was given [3]. Now, a deeper comparison is made and it is proven that extended-order algebras are more restrictive.

### 3.1 Extended-order algebras with an operator

In [6, 12] general implications were considered in order to introduce different algebraic structures. This section only recalls several algebraic structures given in these papers and compares the considered operators with adjoint triples.

The first notion introduced in [12] is the definition of w-eo algebra, from which the rest of structures are presented.

**Definition 3** ([12]) Let \( P \) be a non-empty set, \( \to : P \times P \to P \) a binary operation and \( \top \) a fixed element of \( P \). The triple \( (P, \to, \top) \) is a w-eo algebra, if for all \( a, b, c \in P \) the following conditions are satisfied\(^1\)

\(^1\)Note that the names of the properties are those in [12].
(o₁) $a \to \top = \top$ (upper bound condition)

(o₂) $a \to a = \top$ (reflexivity condition)

(o₃) $a \to b = \top$ and $b \to a = \top$ then $a = b$ (antisymmetry condition)

(o₄) $a \to b = \top$ and $b \to c = \top$ then $a \to c = \top$ (weak transitivity condition)

From a w-eo algebra $(P, \to, \top)$ an ordering can be defined on the set $P$, which provides $P$ with a poset structure. This relation $\leq$ is defined as follows:

$$a \leq b \text{ if and only if } a \to b = \top, \text{ for all } a, b \in P \quad (1)$$

Straightforwardly, $\leq$ is an order relation in $P$, which was called the natural ordering in $P$ [12]. Note that the poset $(P, \leq)$ has a greatest element which coincides with the fixed element $\top$ of $P$.

Other interesting structures introduced in [6, 12] arise when the poset $(P, \leq)$ associated with the w-eo algebra $(P, \to, \top)$ is a complete lattice. In this case, we say that $(P, \to, \top)$ is a complete w-eo algebra $(P, \to, \top)$, in short, a w-eco algebra. In this case we will write $L$ and $\preceq$ instead of $P$ and $\leq$, respectively.

In this context, the following definition presents the right-distributive w-eco algebra.

**Definition 4 ([6])** Let $L$ be a non-empty set, $\to: L \times L \to L$ a binary operation and $\top$ a fixed element of $L$. The triple $(L, \to, \top)$ is a right-distributive w-eco algebra, if it is a w-eco algebra and satisfies the following condition

$$(d'_r) \text{ for any } a \in L, B \subseteq L: a \to \bigwedge_{b \in B} b = \bigwedge_{b \in B} (a \to b)$$

Once we have reminded the notions of w-eo algebra and right-distributive w-eco algebra, we will present several results which relate these notions to adjoint triples (Definition 1).

In order to present next results, extra properties are needed, for instance, that the implications in the adjoint triples will directly satisfy Equivalence (1) or that the conjunctor verifies a boundary condition with the top element. Moreover, the definition of the following mappings is required: $\vee^{\text{op}}: P_2 \times P_3 \to P_1, \wedge^{\text{op}}: P_1 \times P_3 \to P_2$, as $y \vee^{\text{op}} z = z \vee y$ and $x \wedge^{\text{op}} z = z \wedge x$, for all $x \in P_1, y \in P_2$ and $z \in P_3$.

**Proposition 4** Given a poset $(P, \leq)$, with a maximum element $\top$, and an adjoint triple $(\&_{\vee}, \vee, \wedge)$ with respect to $P$. The conjunctor satisfies $\top \&_{\vee} y = y$, for all $y \in P$, if and only if $(P, \vee^{\text{op}}, \top)$ is w-eo algebra and the natural ordering in $P$ is $\leq$.

Analogously, the following proposition is obtained.
Proposition 5 Given a poset \((P, \leq)\), with a maximum element \(\top\), and an adjoint triple \((\&, \\vee, \\wedge)\) with respect to \(P\). The conjunctor satisfies \(x \& \top = x\), for all \(x \in P\), if and only if \((P, \\wedge_{\text{op}}, \top)\) is w-ceo algebra and the natural ordering in \(P\) is \(\leq\).

As a consequence, adjoint triples provide more flexible structures and every property of them can be considered in the particular application frameworks in which these algebras are used.

3.2 Extended-order algebras with two operators

In [6, 12], from right-distributive w-ceo algebra \((L, \to, \top)\) was introduced a residuated operator \(\otimes: L \times L \to L\) which was defined as

\[
a \otimes x = \bigwedge \{t \in L \mid x \leq a \to t\}
\]

(2)

The structures presented in [6] were enriched with an additional binary operation and so, a triple is considered. This section compares this triple with adjoint triples.

Given a right-distributive w-ceo algebra \((L, \to, \top)\), the operator introduced in [6] is denoted as \(\rightsquigarrow: L \times L \to L\) and only needs to satisfy the equivalence

\[
a \preceq b \rightsquigarrow c \iff a \otimes b \preceq c \iff b \preceq a \to c
\]

(3)

for all \(a, b, c \in L\).

With respect to the adjoint triples, the flexibility of these last triples provides that \((\rightsquigarrow, \otimes, \to)\) straightforward is an adjoint triple.

Proposition 6 Given a complete lattice \((L, \leq)\) and the mappings \(\rightsquigarrow, \otimes\) and \(\to\) defined above, the triple \((\rightsquigarrow, \otimes, \to)\) is an adjoint triple with respect to \(L\).

The following structure introduced in [6] is the symmetrical w-ceo algebra, which considers both operators \(\rightsquigarrow\) and \(\to\).

Definition 5 ([6]) A w-ceo algebra \((L, \to, \top)\) is called symmetrical if there exists a binary operation \(\rightsquigarrow: L \times L \to L\) such that \((L, \rightsquigarrow, \top)\) is a w-ceo algebra, \(\to\) and \(\rightsquigarrow\) induce the same ordering and

\[
y \leq x \rightsquigarrow b \text{ if and only if } x \leq y \to b
\]

holds, for all \(b, x, y \in L\).

The w-ceo algebras \((L, \to, \top)\), \((L, \rightsquigarrow, \top)\) and their implications \(\to, \rightsquigarrow\) are said to be dual to each other.

Due to symmetrical character of this notion, \((L, \rightsquigarrow, \top)\) is symmetrical if and only if \((L, \to, \top)\) is symmetrical [6].

Every symmetrical right-distributive w-ceo algebra always provides an adjoint triple. However the counterpart is not true.
Proposition 7 Let \((L, \to, \top)\) be a symmetrical right-distributive w-ceo algebra and the operator \(\otimes: L \times L \to L\) defined by Equation (2), then \((\to, \otimes, \sim)\) is an adjoint triple on \(L\).

The next proposition shows the properties that an adjoint triple must satisfy in order to obtain a symmetrical right-distributive w-ceo algebra.

Proposition 8 Given a poset \((P, \leq)\), with a maximum element \(\top\), and an adjoint triple \((\&\,\!\!\,\", \\,\!\!\,\!\n\\r
\) with respect to \(P\). The conjunctor satisfies \(x \& \top = x\) and \(\top \& y = y\), for all \(x, y \in P\), if and only if \((P, \lor^{\text{op}}, \top)\) is a symmetrical right-distributive w-ceo algebra and the natural ordering in \(P\) is \(\leq\).

An analogous result is obtained with respect to \((P, \\,\!\!\,\!, \\,\!\!\,\!)\).

In a similar way that it has been related right-distributive structures, it could have done with the left-distributive structures which appear in [6] obtaining more flexibility of adjoint triples.

4 Adjoint negations and extended-order algebras with negations

The operations corresponding to the logic connective of negation were introduced in symmetrical algebras in [6]. In this section we will show a comparison between the negation operators presented in [6] and adjoint negations mentioned in Section 2.

In [6], it is defined the negation operators as follows.

Definition 6 Let \((L, \to, \top)\) a w-ceo algebra. We define the following unary operation

\[ [\cdot]^- : L \to L, \quad x \mapsto x^- = x \to \bot \]

If \((L, \to, \top)\) is a symmetrical w-ceo algebra, then we can define a further unary operator

\[ [\cdot]^\sim : L \to L, \quad x \mapsto x^\sim = x \preceq \bot \]

Both these operations are called negation and they are said to be dual to each other.

The negation \([\cdot]^- ([\cdot]^\sim\), respectively\) is involutive if \(x^- = x\) (\(x^\sim = x\), respectively), for every \(x \in L\).

The negations \([\cdot]^-\) and \([\cdot]^\sim\), and the symmetrical w-ceo algebra as well, are said to be cross-involutive if \(x^\sim = x^- = x\), for every \(x \in L\).

From definition above, basic properties of the negations were stated in [6].

Straightforwardly, these operators are particular cases of adjoint negations \(n_s\) and \(n_n\) defined in Section 2.
Proposition 9 Given a symmetrical \( w \)-ceo algebra \((L, \rightarrow, \top)\), the unary operation defined above, \([\cdot]^-\), \([\cdot]^-\) are adjoint negations.

As a consequence, the properties of the adjoint negations can be applied to \([\cdot]^-\), \([\cdot]^-\) on symmetrical \( w \)-ceo algebras \((L, \rightarrow, \top)\).

Almost all these properties given in [6] are satisfied by adjoint negations \(n_s\) and \(n_n\) given in Definition 2, although they satisfy less conditions.

For example, the negation operators \([\cdot]^-\) and \([\cdot]^-\) defined in [6] always satisfy the conditions \(\top^- = \bot\) and \(\top^- = \bot\). However, adjoint negations do not satisfy these properties, i.e. \(n_s(\top) \neq \bot\) and \(n_n(\top) \neq \bot\), in general.

Several of the properties that adjoint negations satisfy are given in the following propositions. These results generalize different properties stated in Proposition 6.2, 6.3 and 6.4 of [6], assuming less hypotheses.

Proposition 10 Let \((P, \leq, \bot, \top)\) be a bounded poset and \(n_s, n_n\) adjoint negations on \(P\). The following statements hold:

1. \(n_s(\bot) = n_n(\bot) = \top\);
2. If \(x \leq y\) then \(n_s(y) \leq n_s(x)\);
3. \(n_s n_n n_s = n_s\) and \(n_n n_s n_n = n_n\);
4. \(n_s n_n\) and \(n_n n_s\) are closure operators;
5. \(x \leq n_s(y)\) iff \(y \leq n_n(x)\), for all \(x, y \in P\);
6. When the supremum and the infimum exist, for any \(X, Y \subseteq P\),
   \[
   (a) \quad n_s(\bigvee_{y \in Y} y) = \bigwedge_{y \in Y} n_s(y),
   \]
   \[
   (b) \quad n_n(\bigvee_{x \in X} x) = \bigwedge_{x \in X} n_n(x).
   \]

The second one is associated with Proposition 6.2 of [6].

Proposition 11 Let \((\cup, \leq, \bot, \top)\) be an adjoint triple with respect to the bounded poset \((P, \leq, \bot, \top)\). The adjoint negation \(n_s\), obtained from the adjoint triple, satisfies the following properties, for all \(x, y \in P\) and \(\{x_i\}_{i \in I} \subseteq L\).
1. $n_s(x) \leq y \vee x$

2. $\bigvee_{i \in I} n_s(x_i) \leq n_s(\bigwedge_{i \in I} x_i)$

3. If $n_s$ is a strong adjoint negation on $P$, then $\bigvee_{i \in I} n_s(x_i) = n_s(\bigwedge_{i \in I} x_i)$

An analogous result is obtained considering $n_n$.

Finally, the following proposition presents interesting properties in a general framework, generalizing the ones given in Proposition 6.3 of [6].

**Proposition 12** Given an adjoint triple $(\& \vee \wedge)$ with respect to the bounded poset $(P, \leq, \bot, \top)$. The adjoint negation $n_n$, obtained from the residuated implication $\vee$, satisfies the following properties, for all $x, y \in P$

1. $x \& n_n(x) = \bot$

2. $x \leq n_n(y)$ if and only if $y \& x = \bot$

The operator $n_s$ satisfies a similar result taking into account that the adjoint product of $(L, \rightsquigarrow, \top)$ is the opposite of $\otimes$, i.e. $x \otimes^{op} y = y \otimes x$.

Therefore, the definition of adjoint negations is more general than the negation operators given in [6], since $n_s$ and $n_n$ satisfy the same properties requiring less conditions, in general.

## 5 Conclusions and future work

This paper continues studying more properties of adjoint triples.

For that purpose, we have presented adjoint negations which are a generalization of the definition of logic connective of negations.

This paper have recalled extended-order algebras which have been studied following the same motivation of adjoint triples in order to reduce the mathematical requirements for computation. The main contribution given by this paper is the comparison between these general operators and adjoint triples. Besides, it has been proven that when these operators have residuated implications they are a particular case of the adjoint triples. In addition, the properties introduced in different papers can be considered to offer extra properties in these frameworks.

Moreover, the comparison of general negation operators introduced in the multi-adjoint framework with the unary decreasing operators defined in symmetrical w-ceo algebras have been given. As a result, we show that the adjoint negations are more general than the former unary operations. Furthermore, the hypotheses of several properties have been weakened.

As future work, more properties of adjoint triples and adjoint negations will be studied.
M.E. Cornejo, J. Medina, E. Ramírez-Poussa

Acknowledgements

Partially supported by the Spanish Science Ministry project TIN2012-39353-C04-04.

References


An study for the Microwave Heating of a Half-Space through Lie symmetries and conservation laws

R. de la Rosa\textsuperscript{1}, M. L. Gandarias\textsuperscript{1} and M. S. Bruzón\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, University of Cádiz
emails: rafael.delarosa@uca.es, marialuz.gandarias@uca.es, m.bruzon@uca.es

Abstract

In this paper, we consider an equation describing microwave heating and we find the subclasses of equations which are nonlinearly self-adjoint. From a general theorem on conservation laws proved by Ibragimov we obtain conservation laws for this equation.

Key words: Classical Symmetries, Partial Differential Equation, Conservation laws

1 Introduction

Lie’s infinitesimal transformation method is an effective and systematic technique in handling partial differential equations (PDEs). Many phenomena in classical mechanics, fluid dynamics, elasticity, and many other applied areas as in the engineering are described by PDEs and the symmetry group techniques provide one method for obtaining solutions of this equations.

The classical method for finding symmetry reductions of partial differential equations is the Lie group method [21, 23, 24]. The fundamental basis of this method is that, when a differential equation is invariant under a Lie group of transformations, a reduction transformation exists. For partial differential equations (PDEs) with two independent variables a single group reduction transforms the PDE into an ordinary differential equations (ODEs), which in general are easier to solve.

The application of Lie groups methods to concrete physical systems involves tedious, mechanical computations. Programmable computer algebra systems such as MATHEMATICA, MACSYMA, MAXIMA, MAPLE and REDUCE, among other, are very powerful in such computations. An excellent survey of the different packages presently available and a discussion of their strengths and applications is given by Hereman [18].

©CMMSE
The use of microwave radiation for heating is common in many industrial situations such as cooking, sterilising, melting, smelting, sintering and drying. In [19] Hill and Pincombe considered the nonlinear heat equation and Maxwell’s equations for a linear conducting medium of density become

\[ \rho c(T) \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( K(T) \frac{\partial T}{\partial x} \right) + q(T)|E|^2, \]
\[ \frac{\partial H}{\partial x} + \frac{\partial}{\partial t} \left[ \epsilon(T) E \right] + \sigma(T) E = 0, \]
\[ \frac{\partial E}{\partial x} + \frac{\partial}{\partial t} \left[ \mu(T) H \right] = 0, \]
where \( T = T(x, y) \) is the temperature, \( c(T) \), \( k(T) \) and \( q(T) \) denote the temperature-dependent specific heat, thermal conductivity and body heating coefficient respectively, while \( \mu(T) \), \( \epsilon(T) \) and \( \sigma(T) \) denote the temperature-dependent magnetic permeability, electric permittivity and electrical conductivity of the medium and \( |E|^2 \) denotes the square of the modulus of the complex electric intensity. The authors attempted to improve the simple model utilised in both Hill and Smyth [20] and Coleman [15] by incorporating a spatial dependence on the heat source. The authors proved the model is consistent with that proposed by Coleman [14] in examining the Stefan problem for microwave heating.

The aim of this paper is to make an analysis of the heat equation

\[ u_t = [A(u)u_x]_x + E(x, t)C(u) \quad (1) \]

by using Lie classical symmetries. We determine, for the equation (1) the subclasses of equations which are nonlinearly self-adjoint. For these classes of nonlinearly self-adjoint equations we obtain conservation laws.

\[ \frac{\partial H}{\partial x} + \frac{\partial}{\partial t} \left[ \epsilon(T) E \right] + \sigma(T) E = 0, \]
\[ \frac{\partial E}{\partial x} + \frac{\partial}{\partial t} \left[ \mu(T) H \right] = 0, \]

\[ \rho c(T) \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( K(T) \frac{\partial T}{\partial x} \right) + q(T)|E|^2, \]
\[ \frac{\partial H}{\partial x} + \frac{\partial}{\partial t} \left[ \epsilon(T) E \right] + \sigma(T) E = 0, \]
\[ \frac{\partial E}{\partial x} + \frac{\partial}{\partial t} \left[ \mu(T) H \right] = 0, \]

where \( \epsilon \) is the group parameter. We require that this transformation leaves invariant the set of solutions of (1). This yields to an overdetermined, linear system of equations for the
infinitesimals \( \xi(x, t, u) \), \( \tau(x, t, u) \) and \( \eta(x, t, u) \). The associated Lie algebra of infinitesimal symmetries is the set of vector fields of the form

\[ v = \xi(x, t, u)\partial_x + \tau(x, t, u)\partial_t + \eta(x, t, u)\partial_u. \] (2)

Invariance of Eq. (1) under a Lie group of point transformations with infinitesimal generator (2) leads to a set of four determining equations. Solving this system we obtain \( \xi = \xi(x, t) \), \( \tau = \tau(t) \) and \( \eta = \eta(x, t, u) \) where \( \xi, \tau \) and \( \alpha \) are related by the following conditions:

\[ \eta A u + \tau A t - 2 \xi A x = 0, \]

\[ 2\eta A x + 2\eta A x - \xi A x + \xi t = 0, \]

\[ \eta AA u u - \eta (A u)^2 + \eta A A u + \eta u A^2 = 0, \]

\[ -\xi A C E A x - \tau A C E A t - \eta A C A E + \eta A u A E + \eta u A C E - 2\xi A C A E - \eta A A^2 + \eta t A = 0. \]

The solutions of this system depend on the functions \( A, C \) and \( E \) of Eq. (1). If \( A \) and \( C \) are arbitrary functions, the infinitesimals are:

\[ \xi = k_1 x + k_3, \] (3)

\[ \tau = 2k_1 t + k_2, \] (4)

\[ \eta = 0. \] (5)

Then the symmetries admitted by (1) are defined by the infinitesimal generators

\[ v_1 = \partial_x, \quad v_2 = \partial_t, \quad v_3 = x\partial_x + 2t\partial_t. \]

and \( E \) must satisfies the equation

\[(k_1 x + k_3) E_x + (2k_1 t + k_2) E_t + 2k_1 E = 0. \] (6)

- If \( A = u^k \) and \( C = u^n \), beside \( v_1, v_2 \) and \( v_3 \) the equation admits new symmetries,
  - If \( E = \gamma(t)x^{2(1-n)/k} \), \( v_4 = x\partial_x + \frac{2}{k}u\partial_u \).
  - If \( E = \theta(x)t^{\frac{n-k-1}{k}} \), \( v_5 = t\partial_t - \frac{1}{k}u\partial_u \).
• If $A = e^{ku}$ and $C = e^{nu}$, beside $v_1$, $v_2$ and $v_3$ the equation admits new symmetries,

- If $E = \gamma(t)x^{\frac{2n}{k}}$, $v_6 = x\partial_x + \frac{2}{k}\partial_u$.
- If $E = \theta(x)t^{\frac{n+k}{k}}$, $v_7 = t\partial_t - \frac{1}{k}\partial_u$.

3 Symbolic manipulation program

In this section we show how the free software MAXIMA program symmgrp2009.max derived by W. Heremann can be used to calculate the determining equations for the potential symmetries of the GBBMB equation (1). To use symmgrp2009.max, we have to convert (1) into the appropriate MAXIMA syntax: $x[1]$ and $x[2]$ represent the independent variables $x$ and $t$, respectively, $u[1]$ and $u[2]$ represent the dependent variables $u$ and $v$, respectively, $u[1, [1, 0]]$ represents $u_x$, $u[1, [1, 1]]$ represents $u_{xt}$, $u[2, [1, 0]]$ represents $v_x$, and $u[2, [0, 1]]$ represents $v_t$. Hence (1) is rewritten as

$$u[1, [0, 1]] - \text{diff}(C, u[1]) \ast u[1, [1, 0]]^2 - A \ast u[1, [2, 0]] - E \ast C = 0$$

The infinitesimals $\xi$, $\tau$ and $\eta$ are represented by $\text{eta1}$, $\text{eta2}$ and $\text{phi1}$, respectively. The program symmgrp2009.max automatically computes the determining equations for the infinitesimals. The batchfile batch containing the MAXIMA commands to implement the program symmgrp2009.max, which we have called clasico.mac is

```maxima
kill(all);
batchload("C:\CLA\symmgrp2009.max");
/*classical symmetries Maxwell equation*/
batch("C:\clasico.dat");
symmetry(1,0,0);
printeqn(lode);
save("lodegnlh.lsp",lode);
for j thru q do (x[j]:=concat(x,j));
for j thru q do (u[j]:=concat(u,j));
ev(lode)$
gnlhode:ev(%,x1=x,x2=t,u1=u);
grind: true$
stringout("gnlhode",gnlhode);
derivabbrev: true;
```

The first lines of this file are standard to symmgrp.max and explained in [13]. The last lines are in order to create an output that suitable for solving the determining equations. This changes $x[1]$, $x[2]$ and $u[1]$ to $x,t$ and $u$, respectively. The file clasica.mac in turn batches the file clasica.dat which contains the requisite data about (1).
4 Determination of self-adjointness, nonlinear self-adjointness equations

**Definition 1.** Consider an $s$th-order partial differential equation

$$F(x, u, u_{(1)}, \ldots, u_{(s)}) = 0 \quad (7)$$

with independent variables $x = (x^1, \ldots, x^n)$ and a dependent variable $u$, where $u_{(1)} = \{u_i\}$, $u_{(2)} = \{u_{ij}\}$, $\ldots$ denote the sets of the partial derivatives of the first, second, etc. orders, $u_i = \partial u / \partial x^i$, $u_{ij} = \partial^2 u / \partial x^i \partial x^j$. The formal Lagrangian is defined as

$$\mathcal{L} = v F \left( x, u, u_{(1)}, \ldots, u_{(s)} \right), \quad (8)$$

where $v = v(x, t)$ is a new dependent variable. The adjoint equation to (7) is

$$F^* \left( x, u, v, u_{(1)}, v_{(1)}, \ldots, u_{(s)}, v_{(s)} \right) = 0, \quad (9)$$

with

$$F^* \left( x, u, v, u_{(1)}, v_{(1)}, \ldots, u_{(s)}, v_{(s)} \right) = \frac{\delta (v F)}{\delta u},$$

where

$$\frac{\delta}{\delta u} = \frac{\partial}{\partial u} + \sum_{s=1}^{\infty} (-1)^s D_{i_1} \cdots D_{i_s} \frac{\partial}{\partial u_{i_1 \cdots i_s}}.$$
denotes the variational derivatives (the Euler-Lagrange operator). Here
\[ D_i = \frac{\partial}{\partial x^i} + u_i \frac{\partial}{\partial u} + u_{ij} \frac{\partial}{\partial u_j} + \cdots \]
are the total differentiations.

In [22] Ibragimov introduced a theorem. The theorem is valid for any system of differential equations where the number of equations is equal to the number of dependent variables. The new theorem does not require existence of a Lagrangian and this theorem is based on a concept of an adjoint equation for nonlinear equations.

Given
\[ F = u_t - [A(u)u_x]_x - E(x,t)C(u), \]
we obtain
\[ F^* \equiv -v C_u E - v_{xx} A - v_t. \] (10)
Setting \( v = u \),
\[ F^* \equiv -u C_u E - u_{xx} A - u_t. \]

Comparing \( F^* \) with \( F \) we obtain the following result:

**Proposition 1.** Equation \( F \equiv u_t - [A(u)u_x]_x - E(x,t)C(u) = 0 \) is not self-adjoint.

Many equations having remarkable symmetry properties and physical significance are not self-adjoint. Therefore one cannot eliminate the nonlocal variables from conservation laws of these equations by setting \( v = u \). In [22] the concept of self-adjoint equation has been generalized by introducing the definition of nonlinearly self-adjoint equations.

Equation (1) is said to be nonlinearly selfadjoint if the equation obtained from the adjoint equation (9) by the substitution
\[ v = \varphi(x, t, u), \] (11)
such that \( \varphi(x, t, u) \neq 0 \),
\[ F^*(x, u, u, u(1), u(1), \ldots, u(s), u(s)) = 0 \]
is identical with the original equation (7), i.e.
\[ F^* \big|_{v=\varphi} = \lambda F. \] (12)

Consequently, we get the following result:

**Proposition 2.** Equation \( F \equiv u_t - [A(u)u_x]_x - E(x, t)C(u) = 0 \) is nonlinearly self-adjoint if \( C = c \), with \( c \) constant, i.e. when it has the following form
\[ u_t - [A(u)u_x]_x - c E(x, t) = 0. \] (13)
5 Conservation laws

Given a PDE a conservation law is of the form

\[ D_t \rho + D_x J = 0, \]

where \( \rho \) is the conserved density, \( J \) is the associated flux, \( D_x J = \frac{\partial J}{\partial x} + \sum_{k=0}^{N} \frac{\partial J}{\partial u_{kx}} u_{(k+1)x}, \) \( N \) is the order of \( J \), and \( D_t \rho = \frac{\partial \rho}{\partial t} + \sum_{k=0}^{M} \frac{\partial \rho}{\partial u_{kx}} D_x u_t, \) with \( M \) the order of \( \rho \).

In order to construct conservation laws we use the following theorem on conservation laws proved in [22].

**Theorem 1** Any Lie point, Lie-Bäcklund or non-local symmetry

\[ X = \xi^i(x, u, u^{(1)}, \ldots) \frac{\partial}{\partial x^i} + \eta(x, u, u^{(1)}, \ldots) \frac{\partial}{\partial u} \]

of equation (7) provides a conservation law \( D_i(C^i) = 0 \) for the simultaneous system (7), (9). The conserved vector is given by

\[ C^i = \xi^i L + W \left[ \frac{\partial L}{\partial u_i} - D_j \left( \frac{\partial L}{\partial u_{ij}} \right) D_k \left( \frac{\partial L}{\partial u_{ijk}} \right) - \cdots \right] \\
+ D_j(W) \left[ \frac{\partial L}{\partial u_{ij}} - D_k \left( \frac{\partial L}{\partial u_{ijk}} \right) + \cdots \right] + D_j D_k(W) \left[ \frac{\partial L}{\partial u_{ijk}} - \cdots \right] + \cdots, \]

where \( W \) and \( L \) are defined as follows:

\[ W = \eta - \xi^i u_j, \quad L = v F(x, u, u^{(1)}, \ldots, u^{(s)}). \]

Let us apply Theorem 1 to the quasi self-adjoint equation (13). We will write generators of point transformation group admitted by Eq. (13) in the form

\[ X = \xi^1 \frac{\partial}{\partial t} + \xi^2 \frac{\partial}{\partial x} + \eta \frac{\partial}{\partial u} \]

by setting \( t = x^1 \), \( x = x^2 \). The conservation law will be written

\[ D_t(C^1) + D_x(C^2) = 0. \]

For \( C \) constant function, equation (1) is nonlinearly self-adjoint when \( h = k \). This equation admits the generator \( v_3 = x \partial_x + 2t \partial_t. \)

- From generator \( v_3 = x \partial_x + 2t \partial_t \), the normal form for this group is
An study for the Microwave Heating of a Half-Space through Lie symmetries

\[ W = -u_x x - 2 tu_t. \]

The vector components are

\[ C^1 = -2 t v C E - 2 t (u_x)^2 v A_u - 2 t u_{xx} v A \]
\[ -u_x v x, \]

\[ C^2 = -v x C E + 2 t u_t u_x v A_u - u_x v_x x A \]
\[ -2 t u_t v_x A + u_x v A + 2 t u_{tx} v A \]
\[ +u_t v x. \]

Setting \( v = k \) in (18)

\[ C^1 = -2 k t C E - 2 k t (u_x)^2 A_u \]
\[ -2 k t u_{xx} A - k u_x x, \]

\[ C^2 = -k x C E + 2 k t u_t u_x A_u + k u_x A \]
\[ +2 k t u_{tx} A + k u_t x. \]

Transfering the terms \( D_x (\ldots) \) from \( C^1 \) to \( C^2 \) and simplifying, we obtain the conserved vector \( C = (C^1, C^2) \) with components given by

\[ C^1 = k u - 2 k t C E, \]

\[ C^2 = -k x C E - k u_x A. \]

For generators \( v_1 \) and \( v_2 \) we obtain trivial conservation laws.

6 Conclusions

In this work we have considered an equation describing microwave heating. By using free software Maxima, we have derived the Lie classical symmetries. We have determined the subclasses of this equations which are self-adjoint and nonlinearly self-adjoint. We found a class of nonlinearly self-adjoint of this equation which are not self-adjoint. By using a general theorem on conservation laws proved by Nail Ibragimov we derived conservation laws for some of these partial differential equations without classical Lagrangians.
Acknowledgements

The authors acknowledge the financial support from Junta de Andalucía group FQM–201.

References


Error analysis in the reconstruction of a convolution kernel in a semilinear parabolic problem

R. H. De Staelen\textsuperscript{1} and M. Slodicka\textsuperscript{1}

\textsuperscript{1} Department of Mathematical Analysis, Faculty of Engineering and Architecture, Ghent University, B-9000 Ghent, Belgium
emails: rob.destaelen@ugent.be, marian.slodicka@ugent.be

Abstract

A semilinear parabolic problem of second order with an unknown convolution kernel is considered. We proved, based on a given global measurement, the existence of a unique weak solution. In this paper we perform an error analysis of our numerical algorithm based on Rothe’s method.

\textit{Key words}: parabolic IBVP, convolution kernel, reconstruction, error analysis

1 Introduction

We are interested in determining the solution \( u \) and reconstruction of the convolution kernel \( K \) of the following nonlinear problem

\[
\begin{aligned}
\partial_t u - \Delta u + K(t)h + (K \ast u)(t) &= f(u, \nabla u), \quad \text{in } \Omega \times \Theta, \\
-\nabla u \cdot n &= g, \quad \text{on } \Gamma \times \Theta, \\
u(x, 0) &= u_0(x),
\end{aligned}
\]

where \( \Omega \) is a Lipschitz domain in \( \mathbb{R}^N \), \( N \geq 3 \), with \( \partial \Omega = \Gamma \) and \( \Theta = [0, T] \), \( T > 0 \), the time frame, when a global measurement

\[
\int_\Omega u(x, t) \, dx = m(t)
\]

is known.
Such type of problems arise for example in the theory of reactive contaminant transport. In [1] one considers the following differential equation

\[ \partial_t C + \nabla \cdot (VC) - \Delta C = -\frac{\rho_b}{n} \partial_t S \]

for the aqueous concentration \( C \) and sorbed concentration per unit mass of solid \( S \) with mass transformation rate in first order form of

\[ \partial_t S = K_r (K_d C - S) \]

with desorption rate \( K_r \) and equilibrium distribution coefficient \( K_d \).

We proved the following:

**Theorem** (see [2]). Suppose \( f \) is everywhere differentiable in \( \mathbb{R}^{1+N} \) and Lipschitz continuous, \( g \in C^2(\Theta, L^2(\Gamma)) \), \( h \in C^0(\Theta, H^1(\Omega)) \) and \( (h,1) \neq 0 \), \( m \in C^2(\Theta, \mathbb{R}) \) and \( u_0 \in H^2(\Omega) \). Then there exists a unique couple weak solutions \((u, K)\) to (1), where \( u \in C(\Theta, L^2(\Omega)) \cap L^\infty(\Theta, H^1(\Omega)) \) which is Lipschitz continuous in time and \( K \in C(\Theta) \cap H^1(\Theta) \).

### 2 Time discretization

We apply the Rothe method [3]. Consider an equidistant time-partitioning of the time frame \( \Theta \) with a step \( \tau = T/n \), for any \( n \in \mathbb{N} \). We use the notation \( t_i = i \tau \) and for any function \( z \) we write

\[ z_i = z(t_i), \quad \delta z_i = \frac{z_i - z_{i-1}}{\tau}. \]

Now, let us introduce the following piecewise linear function in time

\[ u_n : \Theta \to L^2(\Omega) : t \mapsto \begin{cases} u_0 & t = 0 \\ u_{i-1} + (t - t_{i-1})\delta u_i & t \in (t_{i-1}, t_i], \end{cases}, \quad 0 \leq i \leq n, \]

and a step function

\[ \bar{u}_n : \Theta \to L^2(\Omega) : t \mapsto \begin{cases} u_0 & t = 0 \\ u_i & t \in (t_{i-1}, t_i], \end{cases}, \quad 0 \leq i \leq n. \]

Similarly we define \( \bar{K}_n, \bar{h}_n, \bar{g}_n, \bar{m}_n \) and \( \bar{m}'_n \). At time \( t_i \) we infer from the variational formulation of (1) the backward Euler scheme

\[ (\delta u_i, \phi) + (\Delta u_i, \phi) + K_i(h_i, \phi) + \sum_{k=1}^{i} (K_k u_{i-k} \tau, \phi) = (f_{i-1}, \phi). \]
where $f_i = f(u_i, \nabla u_i)$. We obtain from (2) that

$$(\delta u_i, \phi) + (\nabla u_i, \nabla \phi) + (g_i, \phi) + K_i(h_i, \phi) + \sum_{k=1}^{i} (K_k u_{i-k}, \phi) = (f_{i-1}, \phi) \quad \text{(DP)}$$

and alike for the measurement operator

$$m_i' + (g_i, 1) + K_i(h_i, 1) + \sum_{k=1}^{i} K_k m_{i-k} = (f_{i-1}, 1). \quad \text{(DMP)}$$

Using Rothe’s functions we can write (DP) and (DMP) on the whole time frame as

$$(\partial_t u_n, \phi) + (\nabla \bar{u}_n, \nabla \phi) + (\bar{g}_n, \phi) + \sum_{k=1}^{\lfloor t \rfloor} \bar{K}_n(t_k) \bar{u}_n(t - t_k), \phi) = (f(\bar{u}_n(t - \tau), \nabla \bar{u}_n(t - \tau)), \phi). \quad \text{(DP)}$$

and

$$\bar{m}_n' + (\bar{g}_n, 1) + \bar{K}_n(\bar{h}_n, 1) + \sum_{k=1}^{\lfloor t \rfloor} \bar{K}_n(t_k) \bar{m}_n(t - t_k) = (f(\bar{u}_n(t - \tau), \nabla \bar{u}_n(t - \tau)), 1). \quad \text{(DMP)}$$

In this paper we investigate the error

$$\int_0^T \|u_n(t) - u(t)\|_{L^2(\Omega)}^2 \, dt \quad \text{and} \quad \int_0^T \|K_n(t) - K(t)\| \, dt$$

in terms of the time step $\tau = T/n$.

References


A hybrid algorithm for the split generalized equilibrium and the system of variational inequality problems

Jitsupa Deepho\textsuperscript{1} and Wiyada Kumam\textsuperscript{2}

\textsuperscript{1} Department of Mathematics, Faculty of Science, King Mongkut’s University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thrung Kru, Bangkok 10140, Thailand

\textsuperscript{2} Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT) 39, Moo 1, Rangsit-Nakhonnayok Rd., Klong 6, Thanyaburi, Pathumthani 12110, Thailand

emails: jitsupa.deepho@mail.kmutt.ac.th, wiyada.kum@mail.rmutt.ac.th

Abstract

In this paper, we give a hybrid extragradient iterative method for finding the approximate element of the common set of solutions of the split generalized equilibrium problem, system of variational inequality problems, variational inequality problem and fixed point problem for a strictly pseudocontractive mapping in a real Hilbert space. We prove a strong convergence theorem based on this method under certain mild conditions.

Key words: Split generalized equilibrium problem, Variational inequality problem, System of variational inequality problems, Strictly pseudocontractive mappings, Fixed-point problem

MSC 2000: 47H10, 47J05, 47J20, 47J25

1 Introduction

Throughout the paper, unless otherwise stated, let $H_1$ and $H_2$ be real Hilbert spaces with inner product $\langle \cdot , \cdot \rangle$ and norm $\| \cdot \|$. Let $C$ and $Q$ be nonempty closed convex subsets of $H_1$ and $H_2$, respectively. Let $\{x_n\}$ be a sequence in $H_1$, then $x_n \to x$ (respectively, $x_n \rightharpoonup x$) will denote strong (respectively, weak) convergence of the sequence $\{x_n\}$.

Let $T : C \to H$ be a nonlinear mapping. Then $T$ is called
(i) monotone, if
\[ \langle Tx - Ty, x - y \rangle \geq 0, \quad \forall x, y \in H_1; \]

(ii) \(\alpha\)-strongly monotone, if
\[ \langle Tx - Ty, x - y \rangle \geq \alpha \|x - y\|^2, \quad \forall x, y \in H_1; \]

(iii) \(\alpha\)-inverse strongly monotone (or, \(\alpha\)-ism), if there exists a constant \(\alpha > 0\) such that
\[ \langle Tx - Ty, x - y \rangle \geq \alpha \|Tx - Ty\|^2, \quad \forall x, y \in H_1; \]

(iv) \(k\)-Lipschitz continuous, if there exists a constant \(k > 0\) such that
\[ \|Tx - Ty\| \leq k\|x - y\|, \quad \forall x, y \in H_1. \]

It is easy to observe that every \(\alpha\)-strongly monotone mapping \(T\) is monotone and Lipschitz continuous. A mapping \(S : C \to C\) is said to be \(k\)-strict pseudocontractive, if there exists a constant \(0 \leq k < 1\) such that
\[ \|Sx - Sy\|^2 \leq \|x - y\|^2 + k\|(I - S)x - (I - S)y\|^2, \quad \forall x, y \in C. \]

The fixed point problem (in short, FPP) for the mapping \(S\) is to find \(x \in C\) such that
\[ Sx = x \quad (1.1) \]

We denote \(F(S)\), the set of solutions of FPP (1.1).

Given a nonlinear mapping \(A : C \to H\). Then the variational inequality problem (in short, VIP) is to find \(x \in C\) such that
\[ \langle Ax, y - x \rangle \leq 0, \quad \forall y \in C. \quad (1.2) \]

The solution of VIP (1.2) is denoted by \(\Theta\). It is well known that if \(A\) is strongly monotone and Lipschitz continuous mapping on \(C\) then VIP (1.2) has a unique solution. There are several different approaches towards solving this problem in finite dimensional and infinite dimensional spaces see [1–4, 6–9, 24] and the research in this direction is intensively continued.

In 2003, Takahashi and Toyoda [10] introduced an iterative method for finding an element of the set \(F(S) \cap \Gamma\). In 1976, Korpelevich [11] introduced an iterative method for finding a solution of nonconstrained variational inequality problem in finite dimensional Euclidean space \(\mathbb{R}^n\) so-called extragradient method. On the other hand, Xu [12] considered an iterative method know as viscosity approximation method. Later on, the idea of Korpelevich’s [11] was generalized and extended by many authors, see, e.g. [1, 2, 13, 14] for finding the element of the set \(F(S) \cap \Theta\).
Next, we consider the following system of variational inequality problems (in short, SVIP): find \((x^*, y^*) \in C \times C\) such that

\[
\begin{cases}
\langle \rho_1 B_1 y^* + x^* - y^*, x - x^* \rangle \geq 0, & \forall x \in C, \\
\langle \rho_2 B_2 x^* + y^* - x^*, x - y^* \rangle \geq 0, & \forall x \in C,
\end{cases}
\] (1.3)

where \(B_i : C \to C\) is a nonlinear mapping and \(\rho_i > 0\) for each \(i = 1, 2\). The set of solutions of SVIP (1.3) is denoted by \(\Omega\).

Some special case of SVIP (1.3):

(i) If \(B_1 = B_2 = B\) then SVIP (1.3) is reduced to the system of problems of finding \((x^*, y^*) \in C \times C\) such that

\[
\begin{cases}
\langle \rho_1 B y^* + x^* - y^*, x - x^* \rangle \geq 0, & \forall x \in C, \\
\langle \rho_2 B x^* + y^* - x^*, x - y^* \rangle \geq 0, & \forall x \in C,
\end{cases}
\] (1.4)

which is considered and studied by Verma [15].

(ii) If \(x^* = y^*\) in problem (1.4) then problem (1.4) is reduced to the following classical variational inequality problem of finding \(x^* \in C\) such that

\[
\langle B x^*, x - x^* \rangle \geq 0, \quad \forall x \in C.
\] (1.5)

In order to find the solution of SVIP (1.3), Ceng et al. [3] introduced an approximation method known as relaxed extragradient method. Some related works, we refer to see [16–18].

The mixed equilibrium problem (in short, MEP) is to find \(x \in C\) such that

\[
F(x, y) + \langle Dx, y - x \rangle \geq 0, \quad \forall y \in C,
\] (1.6)

where \(F : C \times C \to \mathbb{R}\) is bifunction and \(D : C \to H\) is an nonlinear mappings. This problem was introduced and studied by Moudafi and Théra [19] and Moudafi [20]. The set of solutions of MEP (1.6) is denoted by:

\[
MEP(F) := \{x \in C : F(x, y) + \langle Dx, y - x \rangle \geq 0, \forall y \in C\}.
\] (1.7)

If \(D = 0\) then \(MEP (1.6)\) is reduced to the equilibrium problem (in short, EP) is to find \(x \in C\) such that

\[
F(x, y) \geq 0, \quad \forall y \in C.
\] (1.8)

The solution set of EP (1.8) is denoted by \(EP(F)\).

Various methods have been proposed to solve the equilibrium problems; see, e.g., [21–25]. In 1997, Combettes and Hirstoaga [26] introduced an iterative method of finding the best approximation to the initial data when the set \(EP(F)\) is nonempty. Recently, Plubtieng
and Punpaeng [27] introduced an iterative method for finding the common element of the set $F(S) \cap \Theta \cap EP(F)$.

Recently, Moudafi [28] introduced the following split equilibrium problem (in short, SEP):

Let $F_1 : C \times C \to \mathbb{R}$ and $F_2 : Q \times Q \to \mathbb{R}$ be nonlinear bifunctions and $A : H_1 \to H_2$ be a bounded linear operator, then the split equilibrium problem (SEP) is to find $x^* \in C$ such that

$$F_1(x^*, x) \geq 0, \ \forall x \in C,$$

and such that

$$y^* = Ax^* \in Q \text{ solves } F_2(y^*, y) \geq 0, \ \forall y \in Q. \quad (1.9)$$

When looked separately, (1.9) is the classical equilibrium problem EP and we denoted its solution set by $EP(F_1)$. The SEP (1.9) and (1.10) constitutes a pair of equilibrium problems which have to be solved so that the image $y^* = Ax^*$ under a given bounded linear operator $A$, of the solution $x^*$ of the EP (1.9) in $H_1$ is the solution of another EP (1.10) by $EP(F_2)$.

The solution set SEP (1.9) and (1.10) is denoted by $\Xi = \{ p \in EP(F_1) : Ap \in EP(F_2) \}$.

In 2013, Kazmi and Rivi [29] consider a split generalized equilibrium problem (in short, SGEP): find $x^* \in C$ such that

$$F_1(x^*, x) + h_1(x^*, x) \geq 0, \ \forall x \in C,$$

and such that

$$y^* = Ax^* \in Q \text{ solves } F_2(y^*, y) + h_2(y^*, y) \geq 0, \ \forall y \in Q, \quad (1.11)$$

where $F_1, h_1 : C \times C \to \mathbb{R}$ and $F_2, h_2 : Q \times Q \to \mathbb{R}$ are nonlinear bifunctions and $A : H_1 \to H_2$ is a bounded linear operator.

They denoted the solution set of generalized equilibrium problem (GEP) $(1.11)$ and GEP $(1.12)$ by $GEP(F_1, h_1)$ and $GEP(F_2, h_2)$, respectively. The solution set of SGEP $(1.11)$-$(1.12)$ is denoted by $\Gamma = \{ p \in GEP(F_1, h_1) : Ap \in GEP(F_2, h_2) \}$.

If $h_1 = 0$ and $h_2 = 0$, then SGEP $(1.11)$-$(1.12)$ reduces to SEP $(1.9)$-$(1.10)$. If $h_2 = 0$ and $F_2 = 0$, then SGEP $(1.11)$-$(1.12)$ reduces to the equilibrium problem considered by Cianciaruso et al. [30].

In this paper motivated by the work of Kazmi and Rivi [29, 33], we give a hybrid extragradient method for finding the approximation set of the common set of solutions of $FPP$ $(1.1)$, $VIP$ $(1.2)$, $SVIP$ $(1.3)$ and SGEP $(1.11)$-$(1.12)$ for a strictly pseudocontractive mapping in real Hilbert space. We establish a strong convergence theorem based on this method.
2 Preliminaries

We recall some concepts and results which are needed in sequel. A mapping $P_C$ is said to be metric projection of $H_1$ onto $C$ if for every point $x \in H_1$, there exists a unique nearest point in $C$ denoted by $P_Cx$ such that

$$\|x - P_Cx\| \leq \|x - y\|, \quad \forall y \in C.$$  

It is well known that $P_C$ is a nonexpansive mapping and is characterized by the following property:

$$\langle x - y, P_Cx - P_Cy \rangle \leq \|P_Cx - P_Cy\|^2, \quad \forall x, y \in H_1. \quad (2.1)$$

Moreover, $P_Cx$ is characterized by the following properties:

$$\langle x - P_Cx, y - P_Cx \rangle \leq 0, \quad (2.2)$$

and

$$\|x - y\|^2 \geq \|x - P_Cx\|^2 + \|y - P_Cx\|^2, \quad \forall x \in H_1, y \in C. \quad (2.3)$$

It is known that every nonexpansive operator $T : H_1 \to H_1$ satisfies, for all $(x, y) \in H_1 \times H_1$, the inequality

$$\langle (x - T(x)) - (y - T(y)), T(y) - T(x) \rangle \leq \frac{1}{2} \|(T(x) - x) - (T(y) - y)\|^2 \quad (2.4)$$

and therefore, we get, for all $(x, y) \in H_1 \times \text{Fix}(T),$

$$\langle x - T(x), y - T(x) \rangle \leq \frac{1}{2} \|T(x) - x\|^2, \quad (2.5)$$

(see, e.g., Theorem 3 in [31] and Theorem 1 in [32]).

Let $A$ be a monotone mapping of $C$ into $H_1$, it is easy to observe that,

$$u \in \Gamma \iff u = P_C(u - \lambda Au), \quad \forall \lambda > 0. \quad (2.6)$$

It is also known that $H_1$ satisfies that Opial condition [34], i.e., for any sequence $\{x_n\} \subset H_1$ with $x_n \rightharpoonup x$, the inequality

$$\liminf_{n \to \infty} \|x_n - x\| < \liminf_{n \to \infty} \|x_n - y\|$$

holds for every $y \in H_1$ with $x \neq y$.

A set valued mapping $B : H_1 \to 2^{H_1}$ is called monotone if for all $x, y \in H_1, u \in Bx$ and $v \in By$ imply $\langle x - y, u - v \rangle \geq 0$. A monotone mapping $B : H_1 \to 2^{H_1}$ is maximal if the graph $G(B)$ of $B$ is not properly contained in the graph of any other monotone mapping.
It is known that a monotone mapping $B$ is maximal if and only if for $(x, u) \in H_1 \times H_1, \langle x - y, u - v \rangle \geq 0$, for every $(y, v) \in G(B)$ implies $u \in Bx$. Let $D : C \to H_1$ be an inverse strongly monotone mapping and let $N_Cx$ be the normal cone to $C$ at $x \in C$, i.e., $N_Cx := \{z \in H_1 : \langle y - x, z \rangle \geq 0, \forall y \in C\}$. Define $Bx = \begin{cases} 
abla x + N_Cx, & \text{if } x \in C, \\ \emptyset, & \text{if } x \notin C. \end{cases}$

Then, $B$ is maximal monotone and $0 \in Bx$ if and only if $x \in VI(C, B)$ (see [35] for more details).

Lemma 2.1 [36]. For any $(x^*, y^*) \in C \times C, (x^*, y^*)$ is solution of $SVIP (1.3)$ if and only if $x^*$ is a point of the mapping $M : C \to C$ defined by $M(x) = PC[P_C(x - \mu_2B_2x) - \mu_1B_1(x - \mu_2B_2x)], \forall x \in C,$ (2.7)

where $y^* = PC(x^* - \mu_2B_2x^*), \mu_i \in (0, 2\beta_i)$ and $B_i : C \to H_1$ is a nonlinear mapping for each $i = 1, 2$.

Lemma 2.2 [37] Let $F : C \times C \to \mathbb{R}$ be bifunction satisfying the following assumptions:

(i) $F(x, x) \geq 0$ for all $x \in C$;

(ii) $F$ is monotone, i.e., $F(x, y) + F(y, x) \leq 0$ for all $x \in C;$

(iii) $F$ is upper hemicontinuous, i.e., for each $x, y, z \in C$, $\limsup_{t \to 0} F(tx + (1 - t)x, y) \leq F(x, y);$ 

(iv) For each $x \in C$ fixed, the function $y \mapsto F(x, y)$ is convex and lower semicontinuous; let $h : C \times C \to \mathbb{R}$ such that

(i) $h(x, y) \geq 0$ for all $x \in C$,

(ii) For each $y \in C$ fixed, the function $x \mapsto h(x, y)$ is upper semicontinuous,

(iii) For each $x \in C$ fixed, the function $y \mapsto h(x, y)$ is convex and lower semicontinuous, and assume that for fixed $r > 0$ and $z \in C$, there exists a nonempty compact convex subset $K$ of $H_1$ and $x \in C \cap K$ such that $F(y, x) + h(y, x) + \frac{1}{r}\langle y - x, x - z \rangle < 0, \forall y \in C \backslash K$.

The proof of the following lemma is similar to the proof of Lemma 2.13 in [37] and hence omitted.
Lemma 2.3 Assume that $F_1, h_1 : C \times C \to \mathbb{R}$ satisfying Assumption 1. Let $r > 0$ and $x \in H_1$. Then, there exists $z \in C$ such that

$$F_1(z, y) + h_1(z, y) + \frac{1}{r} \langle y - z, z - x \rangle \geq 0, \forall y \in C.$$ 

Lemma 2.4 \[38\] Assume that the bifunctions $F_1, h_1 : C \times C \to \mathbb{R}$ satisfying Assumption 1 and $h_1$ is monotone. For $r > 0$ and for all $x \in H_1$, define a mapping $T_{r}^{(F_1, h_1)} : H_1 \to C$ as follows:

$$T_{r}^{(F_1, h_1)}(x) = \left\{ z \in C : F_1(z, y) + h_1(z, y) + \frac{1}{r} \langle y - z, z - x \rangle \geq 0, \forall y \in C \right\}.$$ 

Then, the following hold:

1. $T_{r}^{(F_1, h_1)}$ is single-valued.
2. $T_{r}^{(F_1, h_1)}$ is firmly nonexpansive, i.e.,

$$\|T_{r}^{(F_1, h_1)}x - T_{r}^{(F_1, h_1)}y\|^2 \leq \langle T_{r}^{(F_1, h_1)}x - T_{r}^{(F_1, h_1)}y, x - y \rangle, \forall x, y \in H_1.$$ 

3. $Fix(T_{r}^{(F_1, h_1)}) = GEP(F_1, h_1)$.
4. $GEP(F_1, h_1)$ is compact and convex.

Further, assume that $F_2, h_2 : Q \times Q \to \mathbb{R}$ satisfying Assumption 1. For $s > 0$ and for all $w \in H_2$, define a mapping $T_{s}^{(F_2, h_2)} : H_2 \to Q$ as follows:

$$T_{s}^{(F_2, h_2)}(w) = \left\{ d \in Q : F_2(d, e) + h_2(d, e) + \frac{1}{s} \langle e - d, d - w \rangle \geq 0, \forall e \in Q \right\}.$$ 

Then, we easily observe that $T_{s}^{(F_2, h_2)}$ is single-valued and firmly nonexpansive, $GEP(F_2, h_2, Q)$ is compact and convex, and $Fix(T_{s}^{(F_2, h_2)}) = GEP(F_2, h_2, Q)$, where $GEP(F_2, h_2, Q)$ is the solution set of the following generalized equilibrium problem:

Find $y^* \in Q$ such that $F_2(y^*, y) + h_2(y^*, y) \geq 0, \forall y \in Q$.

We observe that $GEP(F_2, h_2) \subseteq GEP(F_2, h_2, Q)$. Further, it is easy to prove that $\Gamma$ is a closed and convex set.

Remark 2.5 Lemmas 2.3 and 2.4 are slight generalizations of Lemma 3.5 in [30] where the equilibrium condition $F_1(x, x) = h_1(x, x) = 0$ has been relaxed to $F_1(x, x) \geq 0$ and $h_1(x, x) \geq 0$ for all $x \in C$. Further, the monotonicity of $h_1$ in Lemma 2.3 is not required.
Lemma 2.6 [39]. Let $S : C \to C$ be a $k$-strictly pseudocontractive mapping. Let $\gamma$ and $\delta$ be two real numbers. Assume that $(\gamma + \delta)k \leq \gamma$. Then
\[
\|\gamma(x - y) + \delta(Sx - Sy)\| \leq (\gamma + \delta)\|x - y\|. \tag{2.8}
\]

Lemma 2.7 [40]. Let $S : C \to C$ be a $k$-strict pseudocontractive mapping. Then:
(i) $S$ satisfies the Lipschitz condition
\[
\|Sx - Sy\| \leq \frac{1 + k}{1 - k}\|x - y\|, \quad \forall x, y \in C; \tag{2.9}
\]
(ii) The mapping $I - S$ is demiclosed at $0$, i.e., if $\{x_n\}$ is a sequence in $C$ such that $x_n \to \bar{x}$ and $(I - S)x_n \to 0$. Then $(I - S)\bar{x} = 0$;
(iii) The set $F(S)$ of $S$ is closed and convex so that the projection $P_{F(S)}$ is well defined.

Lemma 2.8 [41]. Let $(E, \langle \cdot, \cdot \rangle)$ be an inner product space, then for all $x, y \in E$ and $\alpha, \beta, \gamma \in [0, 1]$, we have
\[
\|\alpha x + \beta y + \gamma z\|^2 = \alpha\|x\|^2 + \beta\|y\|^2 + \gamma\|z\|^2 - \alpha\beta\|x - y\|^2 - \alpha\gamma\|x - z\|^2 - \beta\gamma\|y - z\|^2.
\]

Lemma 2.9 The following inequality hold in real Hilbert space $H_1$:
\[
\|x + y\|^2 \leq \|x\|^2 + \langle y, x + y \rangle, \quad \forall x, y \in H_1; \tag{2.10}
\]

Lemma 2.10 [42]. Let $\{a_n\}$ be a sequence of nonnegative real numbers such that
\[
a_{n+1} \leq (1 - \alpha_n)a_n + \delta_n, \quad n \geq 0,
\]
where $\{\alpha_n\}$ is a sequence in $[0, 1)$ and $\delta_n$ is a sequence in $\mathbb{R}$ such that
(i) $\sum_{n=1}^{\infty} \alpha_n = \infty$;
(ii) $\limsup_{n \to \infty} \frac{\delta_n}{\alpha_n} \leq 0$ or $\sum_{n=1}^{\infty} |\alpha_n| < \infty$.
Then $\lim_{n \to \infty} a_n = 0$.

Lemma 2.11 [39] Let $F_1 : C \times C \to \mathbb{R}$ be a bifunction satisfying Assumption 1 hold and let $T_{F_1}^r$ be defined as in Lemma 2.4 for $r > 0$. Let $x, y \in H_1$ and $r_1, r_2 > 0$. Then
\[
\|T_{F_1}^{r_2}y - T_{F_1}^{r_1}x\| \leq \|y - x\| + \left|\frac{r_2 - r_1}{r_2}\right|\|T_{F_1}^{r_1}y - y\|.
\]

Lemma 2.12 [45]. Let $\{x_n\}$ and $\{y_n\}$ be bounded sequences in a Banach space $X$ and $\{\beta_n\}$ be a sequence in $[0, 1]$ with $0 < \liminf_{n \to \infty} \beta_n \leq \limsup_{n \to \infty} \beta_n < 1$. Suppose $x_{n+1} = (1 - \beta_n)y_n + \beta_n x_n$ for all $n \geq 0$ and $\limsup_{n \to \infty} (\|y_{n+1} - y_n\| - \|x_{n+1} - x_n\|) \leq 0$. Then $\lim_{n \to \infty} \|y_n - x_n\| = 0$.

Notation. Let $\{x_n\}$ be a sequence in $H_1$, then $x_n \to x$ (respectively, $x_n \to x$) denotes strong (respectively, weak) convergence of the sequence $\{x_n\}$ to a point $x \in H_1$. 

3 Strong convergence theorem

Theorem 3.1 Let $H_1$ and $H_2$ be two real Hilbert spaces and $C \subset H_1$ and $Q \subset H_2$ be nonempty closed convex subsets of real Hilbert spaces $H_1$ and $H_2$, respectively. For each $i = 1, 2$, let $D, B_i : C \to H_1$ be $\alpha, \beta_i$-inverse strongly monotone mappings, respectively. Let $A : H_1 \to H_2$ be a bounded linear operator. Let $F_1, h_1 : C \times C \to \mathbb{R}$ and $F_2, h_2 : Q \times Q \to \mathbb{R}$ satisfying Lemma 2.2; $h_1, h_2$ are monotone and $F_2$ is upper semicontinuous and $S : C \to C$ be a $k$-strict pseudocontractive mapping such that $Y := F(S) \cap \Gamma \cap \Omega \cap \Theta \neq \emptyset$. Let $f$ be a $\rho$-contraction mapping with $\rho \in [0, \frac{1}{2})$. For a given $x_0 \in C$, let the iterative sequences $\{u_n\}, \{x_n\}, \{y_n\}$ and $\{z_n\}$ be generated by

\[
\begin{cases}
    u_n = T_{r_n}^{(F_1, h_1)}(x_n + \xi A^*(T_{r_n}^{(T_2, h_2)} - I)Ax_n); \\
    z_n = P_C(u_n - \lambda_n Du_n); \\
    y_n = \alpha_n f(x_n) + (1 - \alpha_n)P_C[z_n - \mu_2 B_2 z_n] - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)]; \\
    x_{n+1} = \beta_n x_n + \gamma_n y_n + \delta_n S y_n,
\end{cases}
\]

where $\mu_i \in (0, 2\beta_i)$, for each $i = 1, 2$, $r_n \in (0, \infty)$, $\lambda_n \in (0, 2\alpha)$ and $\xi \in (0, \frac{1}{2})$, $L$ is the spectral radius of the operator $A^*A$ and $A^*$ is the adjoint of $A$ and $\{\alpha_n\}, \{\beta_n\}, \{\gamma_n\}$ and $\{\delta_n\}$ are the sequences in $(0, 1)$ satisfying the following conditions:

(i) $\beta_n + \gamma_n + \delta_n = 1$ and $(\gamma_n + \delta_n)k \leq \gamma_n$, for all $n \geq 0$;

(ii) $\lim_{n \to \infty} \alpha_n = 0$ and $\sum_{n=0}^{\infty} \alpha_n = \infty$;

(iii) $0 < \liminf_{n \to \infty} \beta_n \leq \limsup_{n \to \infty} \beta_n < 1$ and $\liminf_{n \to \infty} \delta_n > 0$;

(iv) $\liminf_{n \to \infty} r_n > 0$, $\sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;

(v) $\lim n \to \infty \left(\frac{n+1}{1-\beta_{n+1}} - \frac{n}{1-\beta_n}\right) = 0$;

(vi) $0 < \liminf_{n \to \infty} \lambda_n \leq \limsup_{n \to \infty} \lambda_n < 2\alpha$ and $\lim_{n \to \infty} |\lambda_{n+1} - \lambda_n| = 0$.

Then the sequence $\{x_n\}$ converges strongly to $z \in P_Y$ where $z = P_Y f(z)$.

Proof 1 Step 1. We will show that the sequence $\{x_n\}$ is bounded.

First, we show that the mapping $(I - r_n D)$ is nonexpansive. For any $x, y \in C$,

\[
\begin{align*}
    &||(I - r_n D)x - (I - r_n D)y||^2 \\
    &\leq ||(x - y) - r_n (Dx - Dy)||^2 = ||x - y||^2 - 2r_n (x - y, Dx - Dy) + r_n^2 ||Dx - Dy||^2 \\
    &\leq ||x - y||^2 - r_n (2\alpha - r_n) ||Dx - Dy||^2 \leq ||x - y||^2.
\end{align*}
\]

Similarly we can show that the mapping $(I - \lambda_n D)$ and $(I - \mu_i B_i)$ are nonexpansive for each $i = 1, 2$. 

Let \( x^* \in \Upsilon \), i.e., \( x^* \in \Gamma \), and we have \( x^* = T_{r_n}^{(F_1, h_1)} x^* \) and \( Ax^* = T_{r_n}^{(F_2, h_2)} Ax^* \).

We estimate

\[
\| u_n - x^* \|^2 = \| T_{r_n}^{(F_1, h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n) - x^* \|^2
\]

\[
= \| T_{r_n}^{(F_1, h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n) - T_{r_n}^{(F_1, h_1)}x^* \|^2
\]

\[
\leq \| x_n + \xi A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n - x^* \|^2
\]

\[
\leq \| x_n - x^* \|^2 + \xi^2 \| A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2 + 2\xi \langle x_n - x^*, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle.
\]

Thus, we have

\[
\| u_n - x^* \|^2 \leq \| x_n - x^* \|^2 + \xi^2 \langle (T_{r_n}^{(F_2, h_2)} - I)Ax_n, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle
\]

\[
+ 2\xi \langle x_n - x^*, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle.
\]

Now, we have

\[
\xi^2 \langle (T_{r_n}^{(F_2, h_2)} - I)Ax_n, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle \leq L \xi^2 \langle (T_{r_n}^{(F_2, h_2)} - I)Ax_n, (T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle
\]

\[
= L \xi^2 \| (T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2.
\]

Denoting \( \Lambda := 2\xi \langle x_n - x^*, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle \) and using (2.5), we have

\[
\Lambda = 2\xi \langle x_n - x^*, A^*(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle
\]

\[
= 2\xi \langle A(x_n - x^*), (T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle
\]

\[
= 2\xi \left\{ \langle T_{r_n}^{(F_2, h_2)}Ax_n - Ap_h(T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle - \langle (T_{r_n}^{(F_2, h_2)} - I)Ax_n \rangle \right\}
\]

\[
\leq 2\xi \left\{ \frac{1}{2} \| (T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2 - \| (T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2 \right\}
\]

\[
\leq -\xi \| (T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2.
\]

Using (3.4), (3.5) and (3.6), we obtain

\[
\| u_n - x^* \|^2 \leq \| x_n - x^* \|^2 + \xi (L\delta - 1) \| (T_{r_n}^{(F_2, h_2)} - I)Ax_n \|^2.
\]

Since \( \xi \in (0, \frac{1}{L}) \), we obtain

\[
\| u_n - x^* \|^2 \leq \| x_n - x^* \|^2.
\]

Since \( x^* \in \Upsilon \), we have

\[
x^* = P_C[P_C(x^* - \mu_2B_2x^*) - \mu_1B_1P_C(x^* - \mu_2B_2x^*)].
\]
Putting $y^* = P_C(x^* - \mu_2 B_2 x^*)$ we see that $x^* = P_C(y^* - \mu_1 B_1 y^*)$.
Since the mapping $D : C \to H_1$ is $\alpha$-inverse strongly monotone, we have

$$
\|z_n - x^*\|^2 = \|P_C(u_n - \lambda_n Du_n) - P_C(x^* - \lambda_n Dx^*)\|^2 \leq \|(u_n - \lambda_n Du_n) - (x^* - \lambda_n Dx^*)\|^2 \\
\leq \|(u_n - x^*) - \lambda_n(Du_n - Dx^*)\|^2 \leq \|u_n - x^\|^2 - \lambda_n(2\alpha - \lambda_n)\|Du_n - Dx^*\|^2 \\
\leq \|u_n - x^\|^2 \leq \|x_n - x^\|^2.
$$

(3.9)

Setting $k_n := P_C[P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)]$ and $v_n := P_C(z_n - \mu_2 B_2 z_n)$.
It follows that

$$
\|v_n - y^*\|^2 = \|P_C(z_n - \mu_2 B_2 z_n) - P_C(x^* - \mu_2 B_2 x^*)\|^2 \leq \|(z_n - \mu_2 B_2 z_n) - (x^* - \mu_2 B_2 x^*)\|^2 \\
\leq \|z_n - x^\|^2 - \mu_2(2\beta_2 - \mu_2)\|B_2 z_n - B_2 x^*\|^2 \leq \|z_n - x^\|^2 \leq \|x_n - x^\|^2.
$$

(3.10)

Furthermore, we have

$$
\|k_n - x^\|^2 \\
= \|P_C[P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)] - P_C[P_C(x^* - \mu_2 B_2 x^*) - \mu_1 B_1 P_C(x^* - \mu_2 B_2 x^*)]\|^2 \\
\leq \|P_C[z_n - \mu_2 B_2 z_n] - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)] - [P_C(x^* - \mu_2 B_2 x^*) - \mu_1 B_1 P_C(x^* - \mu_2 B_2 x^*)]\|^2 \\
\leq \|P_C[z_n - \mu_2 B_2 z_n] - P_C(x^* - \mu_2 B_2 x^*) - \mu_1[B_1 P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(x^* - \mu_2 B_2 x^*)]\|^2 \\
\leq \|P_C[z_n - \mu_2 B_2 z_n] - P_C(x^* - \mu_2 B_2 x^*)\|^2 - \mu_1(2\beta_1 - \mu_1)\|B_1 v_n - B_1 y^*\|^2 \\
\leq \|z_n - x^\|^2 - \mu_2(2\beta_2 - \mu_2)\|B_2 z_n - B_2 x^*\|^2 - \mu_1(2\beta_1 - \mu_1)\|B_1 v_n - B_1 y^*\|^2 \\
\leq \|z_n - x^\|^2 \\
\leq \|x_n - x^\|^2.
$$

(3.11)

Next, we estimate

$$
\|y_n - x^\| = \|\alpha_n(f(x_n) - x^*) + (1 - \alpha_n)(k_n - x^*)\| \leq \alpha_n\|f(x_n) - x^*\| + (1 - \alpha_n)\|k_n - x^*\| \\
\leq \alpha_n(\|f(x_n) - f(x^*)\| + \|f(x^*) - x^*\|) + (1 - \alpha_n)\|k_n - x^*\| \\
\leq \alpha_n(\rho\|x_n - x^*\| + \|f(x^*) - x^*\|) + (1 - \alpha_n)\|x_n - x^*\| \\
= [1 - (1 - \rho)\alpha_n]\|x_n - x^*\| + (1 - \alpha_n)\|f(x^*) - x^*\|/\rho \\
\leq \max \left\{\|x_n - x^*\|, \frac{\|f(x^*) - x^*\|}{1 - \rho}\right\}.
$$

(3.12)

By mathematic induction on $n$, we obtain $\|x_n - x^\| \leq \max \left\{\|x_n - x^*\|, \frac{\|f(x^*) - x^*\|}{1 - \rho}\right\}$, for every $n \geq 0$ and $x_0 \in C$. 

Hence \( \{x_n\} \) is bounded and consequently, we deduce that \( \{u_n\}, \{z_n\}, \{v_n\} \) and \( \{k_n\} \) are bounded.

**Step 2.** We will show that \( \|x_{n+1} - x_n\| \to 0 \).

On the other hand, from the nonexpansivity of the mapping \((I - \lambda_n D)\), we have

\[
\|z_{n+1} - z_n\| = \|P_C(u_{n+1} - \lambda_n D u_{n+1}) - P_C(u_n - \lambda_n D u_n)\| \\
\leq \|(u_{n+1} - \lambda_n D u_{n+1}) - (u_n - \lambda_n D u_n)\| \\
= \|(u_{n+1} - u_n) - \lambda_n(D u_{n+1} - D u_n) + (\lambda_n - \lambda_n) D u_n\| \\
\leq \|(u_{n+1} - u_n) - \lambda_n(D u_{n+1} - D u_n)\| + \|\lambda_n - \lambda_n\|\|D u_n\| \\
\leq \|u_{n+1} - u_n\| + |\lambda_n - \lambda_n|\|D u_n\|. \tag{3.13}
\]

Next, we estimate

\[
\|k_{n+1} - k_n\|^2 = \|P_C[P_C(z_{n+1} - \mu_2 B_2 z_{n+1}) - \mu_1 B_1 P_C(z_{n+1} - \mu_2 B_2 z_{n+1})] \\
- P_C[P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)]\|^2 \\
\leq \|P_C(z_{n+1} - \mu_2 B_2 z_{n+1}) - \mu_1 B_1 P_C(z_{n+1} - \mu_2 B_2 z_{n+1})\|^2 \\
- \|P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)\|^2 \\
\leq \|P_C(z_{n+1} - \mu_2 B_2 z_{n+1}) - P_C(z_n - \mu_2 B_2 z_n)\|^2 \\
- \mu_1(2\beta_1 - \mu_1)\|B_1 P_C(z_{n+1} - \mu_2 B_2 z_{n+1}) - B_1 P_C(z_n - \mu_2 B_2 z_n)\|^2 \\
\leq \|P_C(z_{n+1} - \mu_2 B_2 z_{n+1}) - P_C(z_n - \mu_2 B_2 z_n)\|^2 \\
\leq \|(z_{n+1} - z_n) - \mu_2(B_2 z_{n+1} - B_2 z_n)\|^2 \\
\leq \|z_{n+1} - z_n\|^2 - \mu_2(2\beta_2 - \mu_2)\|B_2 z_{n+1} - B_2 z_n\|^2\|k_{n+1} - k_n\|^2 \\
\leq \|z_{n+1} - z_n\|^2. \tag{3.14}
\]

From (3.13) and (3.14), we have

\[
\|k_{n+1} - k_n\| \leq \|u_{n+1} - u_n\| + |\lambda_{n+1} - \lambda_n|\|D u_n\|. \tag{3.15}
\]

We observe that

\[
\|y_{n+1} - y_n\|^2 = \|k_{n+1} - \alpha_{n+1}[f(x_{n+1}) - k_{n+1} - k_n - \alpha_n[f(x_n) - k_n]]\| \\
\leq \|k_{n+1} - k_n\| + \alpha_{n+1}\|f(x_{n+1}) - k_{n+1}\| + \alpha_n\|f(x_n) - k_n\| \\
\leq \|u_{n+1} - u_n\| + |\lambda_{n+1} - \lambda_n|\|D u_n\| + \alpha_n\|f(x_{n+1}) - k_{n+1}\| \\
+ \alpha_n\|f(x_n) - k_n\|. \tag{3.16}
\]

Since \( T_{r_{n+1}}^{(F_1,h_1)} \) and \( T_{r_{n+1}}^{(F_2,h_2)} \) both are firmly nonexpansive, for \( \xi \in (0, \frac{1}{2}) \), the mapping \( T_{r_{n+1}}^{(F_1,h_1)}(I + \xi A^* (T_{r_{n+1}}^{(F_2,h_2)} - I) A) \) is nonexpansive, see [43, 44]. Further, since \( u_n =
\[ T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n) \text{ and } u_{n+1} = T_{r_{n+1}}^{(F_1,h_1)}(x_{n+1} + \xi A^*(T_{r_{n+1}}^{(F_2,h_2)} - I)Ax_{n+1}) , \]

it follows from Lemma 2.11 that

\[
\|u_{n+1} - u_n\| \leq \|T_{r_{n+1}}^{(F_1,h_1)}(x_{n+1} + \xi A^*(T_{r_{n+1}}^{(F_2,h_2)} - I)Ax_{n+1}) - T_{r_{n+1}}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_{n+1}}^{(F_2,h_2)} - I)Ax_n)\| \\
+ \|T_{r_{n+1}}^{(F_1,h_1)}(x_{n+1} + \xi A^*(T_{r_{n+1}}^{(F_2,h_2)} - I)Ax_{n+1}) - T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n)\| \\
\leq \|x_{n+1} - x_n\| + \|x_n + \xi A^*(T_{r_{n+1}}^{(F_2,h_2)} - I)Ax_{n+1}) - (x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n)\| \\
+ \left| 1 - \frac{r_n}{r_{n+1}}\right| \|T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n) - (x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n)\| \\
\leq \|x_{n+1} - x_n\| + \xi \|A\| \|\|T_{r_{n+1}}^{(F_2,h_2)}Ax_{n+1} - T_{r_n}^{(F_2,h_2)}Ax_n\| + \delta_n \\
\leq \|x_{n+1} - x_n\| + \xi \|A\| \left| 1 - \frac{r_n}{r_{n+1}}\right| \|T_{r_n}^{(F_2,h_2)}Ax_n - Ax_n\| + \delta_n \\
= \|x_{n+1} - x_n\| + \xi \|A\| \sigma_n + \varphi_n \quad (3.17)
\]

where

\[ \sigma_n := \left| 1 - \frac{r_n}{r_{n+1}}\right| \|T_{r_n}^{(F_2,h_2)}Ax_n - Ax_n\| \]

and

\[ \varphi_n := \left| 1 - \frac{r_n}{r_{n+1}}\right| \|T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n) - (x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n)\|. \]

Using (3.16) and (3.17), we have

\[
\|y_{n+1} - y_n\| \leq \|x_{n+1} - x_n\| + \xi \|A\| \sigma_n + \varphi_n + |\lambda_{n+1} - \lambda_n| \|Du_n\| + \alpha_{n+1} \|f(x_{n+1}) - k_{n+1}\| + \alpha_n \|f(x_n) - k_n\| \quad (3.18)
\]

Setting \( x_{n+1} = \beta_n x_n + (1 - \beta_n)e_n \), which implies from (3.1) that

\[ \gamma_n = \frac{x_{n+1} - \beta_n x_n}{1 - \beta_n} = \gamma_n y_n + \delta_n S y_n \]

Further, it follows that

\[
e_{n+1} - e_n = \gamma_n y_n + \delta_n S y_n \]

where

\[ e_n = \frac{x_{n+1} - \beta_n x_n}{1 - \beta_n} = \gamma_n y_n + \delta_n S y_n \]

\[ \frac{\gamma_{n+1} y_{n+1} + \delta_{n+1} S y_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_n y_n + \delta_n S y_n}{1 - \beta_n} \]

\[ = \frac{\gamma_{n+1} y_{n+1} + \delta_{n+1} S y_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_{n+1} y_{n+1} + \delta_{n+1} S y_{n+1}}{1 - \beta_{n+1}} \]

\[ + \frac{\gamma_{n+1} y_{n+1} + \delta_{n+1} S y_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_{n+1} y_{n+1} + \delta_{n+1} S y_{n+1}}{1 - \beta_n} \]

\[ = \gamma_{n+1} (y_{n+1} - y_n) + \delta_{n+1} (S y_{n+1} - S y_n) + \left( \frac{\gamma_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_n}{1 - \beta_n} \right) y_n \]

\[ + \left( \frac{\delta_{n+1}}{1 - \beta_{n+1}} - \frac{\delta_n}{1 - \beta_n} \right) S y_n \]
Since \((\gamma_n + \delta_n)k \leq \gamma_n\), for all \(n \geq 0\), from Lemma 2.6, we get
\[
\|\gamma_{n+1}(y_{n+1} - y_n) + \delta_{n+1}(Sy_{n+1} - Sy_n)\| \leq (\gamma_{n+1} + \delta_{n+1})\|y_{n+1} - y_n\| \tag{3.19}
\]

Hence we obtain
\[
\|e_{n+1} - e_n\| = \left\| \frac{\gamma_{n+1}(y_{n+1} - y_n) + \delta_{n+1}(Sy_{n+1} - Sy_n)}{1 - \beta_{n+1}} + \frac{\gamma_{n+1} - \gamma_n}{1 - \beta_n} \|y_n\| + \frac{\delta_{n+1}}{1 - \beta_{n+1}} \|Sy_n\| \right\|
\leq \frac{\gamma_{n+1} + \delta_{n+1}}{1 - \beta_{n+1}} \|y_{n+1} - y_n\| + \left\| \frac{\gamma_{n+1} - \gamma_n}{1 - \beta_{n+1}} \|y_n\| + \|Sy_n\| \right\|
\]
\[
= \|y_{n+1} - y_n\| + \left\| \frac{\gamma_{n+1} - \gamma_n}{1 - \beta_{n+1}} \|y_n\| + \|Sy_n\| \right\|.
\]

From (3.18), we obtain
\[
\|e_{n+1} - e_n\| \leq \|x_{n+1} - x_n\| + \xi\|A\|\|\sigma_n + \varrho_n + |\lambda_{n+1} - \lambda_n||Du_n| + \alpha_{n+1}\|f(x_{n+1}) - k_{n+1}\| + \alpha_n\|f(x_n) - k_n\| + \left\| \frac{\gamma_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_n}{1 - \beta_n} \right\| \|y_n\| + \|Sy_n\|, \tag{3.20}
\]
which implies that
\[
\|e_{n+1} - e_n\| - \|x_{n+1} - x_n\| \leq \xi\|A\|\|\sigma_n + \varrho_n + |\lambda_{n+1} - \lambda_n||Du_n| + \alpha_{n+1}\|f(x_{n+1}) - k_{n+1}\| + \alpha_n\|f(x_n) - k_n\| + \left\| \frac{\gamma_{n+1}}{1 - \beta_{n+1}} - \frac{\gamma_n}{1 - \beta_n} \right\| \|y_n\| + \|Sy_n\|.
\]

Hence it follows by conditions (i)-(vi) that
\[
\limsup_{n \to \infty} [\|e_{n+1} - e_n\| - \|x_{n+1} - x_n\|] \leq 0. \tag{3.21}
\]

Hence Lemma 2.12 and (3.21), we get \(\lim_{n \to \infty} \|e_n - x_n\| = 0\) and
\[
\lim_{n \to \infty} \|x_{n+1} - x_n\| = \lim_{n \to \infty} (1 - \beta_n)\|e_n - x_n\| = 0. \tag{3.22}
\]

**Step 3.** We will show that \(\|x_n - u_n\| \to 0\) as \(n \to \infty\).
Since $x^* \in \mathcal{Y}$, by using Lemma 2.6 and (3.11), we obtain
\[
\|x_{n+1} - x^*\|^2 \\
= \|\beta_n(x_n - x^*) + \gamma_n(y_n - x^*) + \delta_n(Sy_n - y^*)\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\left\| \frac{1}{\gamma_n + \delta_n}(\gamma_n(u_n - x^*)) + \delta_n(Sy_n - x^*) \right\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\|y_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)[\alpha_n\|f(x_n) - x^*\|^2 + (1 - \alpha_n)\|k_n - x^*\|] \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)\|k_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)\|z_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)\|u_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)\|z_n - x^*\|^2 + \xi(\|x_n - x^*\|^2 + \xi(L\delta - 1)\|T_{r_n}^{(F_2,h_2)} - I\|Ax_n\|^2) \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (1 - \beta_n)\|\|x_n - x^*\|^2 + \xi(\|x_n - x^*\|^2 + \xi(L\delta - 1)\|T_{r_n}^{(F_2,h_2)} - I\|Ax_n\|^2) \\
= \|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 - (1 - \beta_n)\xi(1 - L\delta)\|T_{r_n}^{(F_2,h_2)} - I\|Ax_n\|^2. \quad (3.23)
\]
Therefore
\[
(1 - \beta_n)\xi(1 - L\delta)\|T_{r_n}^{(F_2,h_2)} - I\|Ax_n\|^2 \leq \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 \\
\leq (\|x_n - x^*\|^2 + \|x_{n+1} - x^*\|)\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2.
\]
Since $\alpha_n \to 0, 0 < \limsup_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1, \|x_{n+1} - x_n\| \to 0$, we obtain
\[
\lim_{n\to\infty} \|T_{r_n}^{(F_2,h_2)} - I\|Ax_n\| = 0. \quad (3.24)
\]
Observe that
\[
\|u_n - x^*\|^2 = \|T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n) - x^*\|^2 \\
= \|T_{r_n}^{(F_1,h_1)}(x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n) - T_{r_n}^{(F_1,h_1)}x^*\|^2 \\
\leq \langle u_n - x^*, x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n - x^* \rangle \\
= \frac{1}{2}\left\{ \|u_n - x^*\|^2 + \|x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n - x^*\|^2 \\
- \|(u_n - x^*) - [x_n + \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n - x^*]\|^2 \right\} \\
= \frac{1}{2}\left\{ \|u_n - x^*\|^2 + \|x_n - x^*\|^2 - \|u_n - x_n - \xi A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|^2 \right\} \\
= \frac{1}{2}\left\{ \|u_n - x^*\|^2 + \|x_n - x^*\|^2 - \|u_n - x_n\|^2 + \xi^2\|A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|^2 \\
-2\xi\langle u_n - x_n, A^*(T_{r_n}^{(F_2,h_2)} - I)Ax_n \rangle \right\}.
\]
Hence, we obtain

$$
\|u_n - x^*\|^2 \leq \|x_n - x^\|2 - \|u_n - x_n\|^2 + 2\xi\|A(u_n - x_n)\|\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|. \quad (3.25)
$$

We consider

$$
\|x_{n+1} - x^\|^2 = \beta_n\|x_n - x^\|^2 + \gamma_n\|y_n - x^\|^2 + \delta_n\|Sy_n - x^\|^2 \\
\leq \beta_n\|x_n - x^\|^2 + (\gamma_n + \delta_n)\left(1 - \frac{1}{\gamma_n + \delta_n}\gamma_n\|y_n - x^\|^2 + \delta_n\|Sy_n - x^\|^2\right) \\
\leq \beta_n\|x_n - x^\|^2 + \gamma_n\|f(x_n) - x^\|^2 + \gamma_n\|\gamma_n\|\|k_n - x^\|^2 \\
\leq \beta_n\|x_n - x^\|^2 + \gamma_n\|f(x_n) - x^\|^2 + \gamma_n\|\gamma_n\|\|u_n - x^\|^2 \\
\leq \beta_n\|x_n - x^\|^2 + \gamma_n\|f(x_n) - x^\|^2 + \gamma_n\|\gamma_n\| \\
\times [(\|x_n - x^\|^2 - \|u_n - x_n\|^2 + 2\xi\|A(u_n - x_n)\|\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|] \\
= \|x_n - x^\|^2 + \gamma_n\|f(x_n) - x^\|^2 - (1 - \beta_n)\|u_n - x_n\|^2 \\
+ (1 - \beta_n)2\xi\|A(u_n - x_n)\|\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|. \quad (3.26)
$$

Therefore

$$
(1 - \beta_n)\|u_n - x_n\|^2 \leq (\|x_n - x^\|^2 - \|x_{n+1} - x^\|^2 + \alpha_n\|f(x_n) - x^\|^2 \\
+ (1 - \beta_n)2\xi\|A(u_n - x_n)\|\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\| \\
\leq (\|x_n - x^\|^2 + \|x_{n+1} - x^\|^2)\|x_n - x_{n+1}\| + \alpha_n\|f(x_n) - x^\|^2 \\
+ (1 - \beta_n)2\xi\|A(u_n - x_n)\|\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\|.
$$

Since $\alpha_n \to \infty$, $0 < \lim inf_{n \to \infty} \beta_n \leq \lim sup_{n \to \infty} \beta_n < 1$, $\|x_{n+1} - x_n\| \to 0$, $\|(T_{r_n}^{(F_2,h_2)} - I)Ax_n\| \to 0$ as $n \to \infty$, we obtain

$$
\lim_{n \to \infty} \|u_n - x_n\| = 0. \quad (3.27)
$$

**Step 4.** We will show that $\lim_{n \to \infty} ||Du_n - Dx^*|| = \lim_{n \to \infty} ||B_1v_n - B_1y^*|| = \lim_{n \to \infty} ||B_2z_n - B_2x^*|| = 0.$
From (3.1), (3.9) and (3.11), we have

\[
\|x_{n+1} - x^*\|^2
= \|\beta_n(x_n - x^*) + \gamma_n(y_n - x^*) + \delta_n(Sy_n - x^*)\|^2
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\frac{1}{\gamma_n + \delta_n}(\gamma_n(y_n - x^*) + \delta_n(Sy_n - x^*))^2
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\|y_n - x^*\|^2
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)[\alpha_n\|f(x_n) - x^*\|^2 + (1 - \alpha_n)\|k_n - x^*\|^2]
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)[\|z_n - x^*\|^2 - \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2
\]

\[
- \mu_1(2\beta_1 - \mu_1)\|B_1v_n - B_1y^*\|^2
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)[\|z_n - x^*\|^2 - \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2
\]

\[
- \mu_1(2\beta_1 - \mu_1)\|B_1v_n - B_1y^*\|^2
\]

\[
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)[\|z_n - x^*\|^2 - \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2
\]

\[
- \mu_1(2\beta_1 - \mu_1)\|B_1v_n - B_1y^*\|^2
\]

\[
\leq \|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (\gamma_n + \delta_n)[\|z_n - x^*\|^2 - \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2
\]

\[
+ \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2 + \mu_1(2\beta_1 - \mu_1)\|B_1v_n - B_1y^*\|^2
\].

Hence from (3.28), we have

\[
(\gamma_n + \delta_n)(\lambda_n(2\alpha - \lambda_n)\|Du_n - Dx^*\|^2 + \mu_2(\beta_2 - \mu_2)\|B_2z_n - B_2x^*\|^2 + \mu_1(2\beta_1 - \mu_1)\|B_1v_n - B_1y^*\|^2)
\]

\[
\leq \|x_n - x^*\|^2 + \|x_{n+1} - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2
\]

\[
\leq (\|x_n - x^*\| + \|x_{n+1} - x^*\|)\|x_n - x_{n+1}\| + \alpha_n\|f(x_n) - x^*\|^2.
\]

Since \(0 < \liminf_{n \to \infty} \lambda_n \leq \limsup_{n \to \infty} \lambda_n < 2\alpha, \|x_{n+1} - x_n\| \to 0, \alpha_n \to 0\) and \(\liminf_{n \to \infty} (\gamma_n + \delta_n) > 0\), we obtain

\[
\lim_{n \to \infty} \|Du_n - Dx^*\| = \lim_{n \to \infty} \|B_1v_n - B_1y^*\| = \lim_{n \to \infty} \|B_2z_n - B_2x^*\| = 0. \tag{3.29}
\]

**Step 5.** We will show that \(\lim_{n \to \infty} \|u_n - z_n\| = \lim_{n \to \infty} \|z_n - k_n\| = \lim_{n \to \infty} \|k_n - u_n\| = \lim_{n \to \infty} \|y_n - z_n\| = \lim_{n \to \infty} \|y_n - x_n\| = \lim_{n \to \infty} \|z_n - v_n - (x^* - y^*)\| = \lim_{n \to \infty} \|v_n - k_n + (x^* - y^*)\| = 0\).
Further, we observe that

\[ \|z_n - x^*\|^2 \]
\[ = \|P_C(u_n - \lambda_n Du_n) - P_C(x^* - \lambda_n Dx^*)\|^2 \]
\[ \leq \langle (u_n - \lambda_n Du_n) - (x^* - \lambda_n Dx^*), z_n - x^* \rangle \]
\[ \leq \frac{1}{2}\{\|u_n - x^* - \lambda_n(Du_n - Dx^*)\|^2 + \|z_n - x^*\|^2 - \|u_n - x^* - \lambda_n(Du_n - Dx^*) - (z_n - x^*)\|^2\} \]
\[ \leq \frac{1}{2}\{\|u_n - x^*\|^2 + \|z_n - x^*\|^2 - \|u_n - z_n - \lambda_n(Du_n - Dx^*)\|^2\} \]
\[ \leq \frac{1}{2}\{\|x_n - x^*\|^2 + \|z_n - x^*\|^2 - \|u_n - z_n\|^2 + 2\lambda_n\langle u_n - z_n, Du_n - z_n \rangle - \lambda_n^2\|Du_n - Dx^*\|^2\} \]
\[ \leq \frac{1}{2}\{\|x_n - x^*\|^2 + \|z_n - x^*\|^2 - \|u_n - z_n\|^2 + 2\lambda_n\|u_n - z_n\|\|Du_n - Dx^*\|^2\}. \]

Hence,

\[ \|z_n - x^*\|^2 \leq \|x_n - x^*\|^2 - \|u_n - z_n\|^2 + 2\lambda_n\|u_n - z_n\|\|Du_n - Dx^*\|^2. \quad (3.30) \]

Next, we estimate

\[ \|v_n - y^*\|^2 \]
\[ = \|P_C(z_n - \mu_2 B_2 z_n) - P_C(x^* - \mu_2 B_2 x^*)\|^2 \]
\[ \leq \langle (z_n - \mu_2 B_2 z_n) - (x^* - \mu_2 B_2 x^*), v_n - y^* \rangle \]
\[ \leq \frac{1}{2}\{\|z_n - x^* - \mu_2(B_2 z_n - B_2 x^*)\|^2 + \|v_n - y^*\|^2 - \|z_n - x^* - \mu_2(B_2 z_n - B_2 x^*) - (v_n - y^*)\|^2\} \]
\[ \leq \frac{1}{2}\{\|z_n - x^*\|^2 + \|v_n - y^*\|^2 - \|z_n - v_n - \mu_2(B_2 z_n - B_2 x^*) - (x^* - y^*)\|^2\} \]
\[ \leq \frac{1}{2}\{\|z_n - x^*\|^2 + \|v_n - y^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 + 2\mu_2\langle z_n - v_n - (x^* - y^*), B_2 z_n - B_2 x^* \rangle \]
\[ - \mu_2^2\|B_2 z_n - B_2 x^*\|^2 \}
\[ \leq \frac{1}{2}\{\|z_n - x^*\|^2 + \|v_n - y^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 + 2\mu_2\|z_n - v_n - (x^* - y^*)\|\|B_2 z_n - B_2 x^*\|^2\}. \]

Hence,

\[ \|v_n - y^*\|^2 \leq \|z_n - x^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 + 2\mu_2\|z_n - v_n - (x^* - y^*)\|\|B_2 z_n - B_2 x^*\|^2 \]
\[ \leq \|x_n - x^*\|^2 - \|u_n - z_n\|^2 + 2\lambda_n\|u_n - z_n\|\|Du_n - Dx^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 \]
\[ + 2\mu_2\|z_n - v_n - (x^* - y^*)\|\|B_2 z_n - B_2 x^*\|^2. \quad (3.31) \]
Similarly, we also estimate

\[
\|k_n - x^*\|^2 \\
= \|PC(v_n - \mu_1B_1v_n) - PC(y^* - \mu_1B_1y^*)\|^2 \\
\leq \langle (v_n - \mu_1B_1v_n) - (y^* - \mu_1B_1y^*), k_n - x^* \rangle \\
\leq \frac{1}{2} \{ \|v_n - y^*\|^2 + \|k_n - x^*\|^2 - \|v_n - y^* - \mu_1(B_1v_n - B_1y^*) - (k_n - x^*)\|^2 \} \\
\leq \frac{1}{2} \{ \|v_n - y^*\|^2 + \|k_n - x^*\|^2 - \|v_n - k_n - \mu_1(B_1v_n - B_1y^*) + (x^* - y^*)\|^2 \} \\
\leq \frac{1}{2} \{ \|v_n - y^*\|^2 + \|k_n - x^*\|^2 - \|v_n - k_n + (x^* - y^*)\|^2 + 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\| \} \\
- \mu_2^2\|B_1v_n - B_1y^*\|^2 \\
\leq \frac{1}{2} \{ \|v_n - y^*\|^2 + \|k_n - x^*\|^2 - \|v_n - k_n + (x^* - y^*)\|^2 + 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\| \}. \\
\]

Hence, from (3.31),

\[
\|k_n - x^*\|^2 \leq \|v_n - y^*\|^2 - \|v_n - k_n + (x^* - y^*)\|^2 + 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\| \\
\leq \|x_n - x^*\|^2 - \|u_n - z_n\|^2 + 2\lambda_n\|u_n - z_n\| \|Du_n - Dx^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 \\
+ 2\mu_2\|z_n - v_n - (x^* - y^*)\| \|B_2z_n - B_2x^*\| - \|v_n - k_n + (x^* - y^*)\|^2 \\
+ 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\|. \\
\]

(3.32)

From Lemma 2.6, (3.1) and (3.32), we have

\[
\|x_{n+1} - x^*\|^2 = \|\beta_n(x_n - x^*) + \gamma_n(y_n - x^*) + \delta_n(Sy_n - x^*)\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\left\|\frac{1}{\gamma_n + \delta_n}(\gamma_n(y_n - x^*) + \delta_n(Sy_n - x^*))\right\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + \|\gamma_n + \delta_n\|y_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + (\gamma_n + \delta_n)\|\alpha_n\|f(x_n) - x^*\|^2 + (1 - \alpha_n)\|k_n - x^*\|^2 \\
\leq \beta_n\|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (1 - \beta_n)\times\|x_n - x^*\|^2 - \|u_n - z_n\|^2 \\
+ 2\lambda_n\|u_n - z_n\|\|Du_n - Dx^*\|^2 - \|z_n - v_n - (x^* - y^*)\|^2 \\
+ 2\mu_2\|z_n - v_n - (x^* - y^*)\| \|B_2z_n - B_2x^*\| + 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\| \\
= \|x_n - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 + (1 - \beta_n)\times\|x_n - x^*\|^2 - \|u_n - z_n\|^2 \\
+ 2\lambda_n\|u_n - z_n\|\|Du_n - Dx^*\|^2 \\
+ 2\mu_2\|z_n - v_n - (x^* - y^*)\| \|B_2z_n - B_2x^*\| + 2\mu_1\|v_n - k_n + (x^* - y^*)\| \|B_1v_n - B_1y^*\| \\
-(1 - \beta_n)\|u_n - z_n\|^2 + \|z_n - v_n - x^* - (x^* - y^*)\|^2 + \|v_n - k_n + (x^* - y^*)\|^2 \\
\]

©CMMSE
which yield

\[
(1 - \beta_n)\|u_n - z_n\|^2 + \|z_n - v_n - x^* - (x^* - y^*)\|^2 + \|v_n - k_n + (x^* - y^*)\|^2 \\
\leq \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 + \alpha_n\|f(x_n) - x^*\|^2 \\
+ (1 - \beta_n) \times [2\lambda_n\|u_n - z_n\|\|Du_n - D\nabla x^*\|^2 + 2\mu_2\|z_n - v_n - (x^* - y^*)\|\|B_2z_n - B_2x^*\| \\
+ 2\mu_1\|v_n - k_n + (x^* - y^*)\|\|B_1v_n - B_1y^*\|]
\]

\[
\leq \left(\|x_n - x^*\| + \|x_{n+1} - x^*\|\right)\|x_n - x_{n+1}\| + \alpha_n\|f(x_n) - x^*\|^2 \\
+ (1 - \beta_n) \times [2\lambda_n\|u_n - z_n\|\|Du_n - D\nabla x^*\|^2 + 2\mu_2\|z_n - v_n - (x^* - y^*)\|\|B_2z_n - B_2x^*\| \\
+ 2\mu_1\|v_n - k_n + (x^* - y^*)\|\|B_1v_n - B_1y^*\|]. \\
(3.33)
\]

Since \(0 < \limsup_{n \to \infty} \beta_n \leq 1, 0 < \lambda_n \leq 2\alpha_n, \|x_{n+1} - x_n\| \to 0, \alpha_n \to 0\) and \(\lim_{n \to \infty} \|Du_n - D\nabla x^*\| = 0, \lim_{n \to \infty} \|B_2z_n - B_2x^*\| = 0\) and \(\lim_{n \to \infty} \|B_1v_n - B_1y^*\| = 0\), it follows from the boundedness of \(\{x_n\}, \{u_n\}, \{k_n\}, \{z_n\}\) and \(\{v_n\}\) that

\[
\lim_{n \to \infty} \|u_n - z_n\| = \lim_{n \to \infty} \|z_n - v_n - (x^* - y^*)\| = \lim_{n \to \infty} \|v_n - k_n + (x^* - y^*)\| = 0. \\
(3.34)
\]

Consequently from (3.34), it follows that

\[
\|z_n - k_n\| \leq \|z_n - v_n - (x^* - y^*)\| + \|v_n - k_n + (x^* - y^*)\| \to 0 \quad \text{as} \quad n \to \infty. \\
(3.35)
\]

Also from (3.34) and (3.35), we have

\[
\|u_n - k_n\| \leq \|u_n - z_n\| + \|z_n - k_n\| \to 0 \quad \text{as} \quad n \to \infty, \\
(3.36)
\]

and

\[
\|y_n - k_n\| \leq \alpha_n\|f(x_n) - k_n\| \to 0 \quad \text{as} \quad n \to \infty, \\
(3.37)
\]

and from (3.36) and (3.37), we have

\[
\|u_n - y_n\| \leq \|u_n - k_n\| + \|k_n - y_n\| \to 0 \quad \text{as} \quad n \to \infty, \\
(3.38)
\]

and from (3.27) and (3.38), we have

\[
\|y_n - x_n\| \leq \|y_n - u_n\| + \|u_n - x_n\| \to 0 \quad \text{as} \quad n \to \infty, \\
(3.39)
\]

**Step 6.** We will show that \(w \in \mathcal{Y}\).

**Step 6.1** We will show that \(w \in \text{Fix}(S)\). Since from (3.1), we get

\[
\delta_n\|(Sy_n - x_n)\| \leq \|x_{n+1} - x_n\| + \gamma_n\|y_n - x_n\|,
\]

it follows from (3.22) and (3.39), we get

\[
\lim_{n \to \infty} \|Sy_n - x_n\| = 0, \\
(3.40)
\]
and it follows from (3.39) and (3.40), we get

\[ \|Sy_n - y_n\| \leq \|Sy_n - x_n\| + \|x_n - y_n\|. \]

Also since \( H \) is reflexive and \( \{y_n\} \) is bounded, without loss of generality we can assume that \( y_n \to w \) for some \( w \in C \). It follows from Lemma 2.7 (ii) that \( w \in \text{Fix}(S) \).

**Step 6.2** We will show that \( w \in \Gamma \).

First, we will show \( w \in \text{GEP}(F_1, h_1) \).

Since \( u_n = J_{\Gamma}^{(F_1, h_1)} x_n \), we have

\[ F_1(u_n, y) + h_1(u_n, y) + \frac{1}{r_n} \langle y - u_n, u_n - x_n \rangle \geq 0, \quad \forall y \in C. \]

It follows from the monotonicity of \( F_1 \) that

\[ h_1(u_n, y) + \frac{1}{r_n} \langle y - u_n, u_n - x_n \rangle \geq F_1(y, u_n), \]

and hence replacing \( n \) by \( n_i \), we get

\[ h_1(u_{n_i}, y) + \frac{1}{r_{n_i}} \langle y - u_{n_i}, u_{n_i} - x_{n_i} \rangle \geq F_1(y, u_{n_i}). \]

Since \( \|u_n - x_n\| \to 0, \|y_n - x_n\| \to 0, \) and \( y_n \to w \), we get \( u_{n_i} \to w \) and \( \frac{u_{n_i} - x_{n_i}}{r_{n_i}} \to 0. \)

It follows by Lemma 2.2 (iv) that \( 0 \geq F_1(y, w), \forall w \in C \). For any \( g \) with \( 0 < g \leq 1 \) and \( y \in C \), let \( y_g = gy + (1 - g)w, \forall x \in C \). Since \( y \in C, w \in C \), we have \( y_g \in C \), and hence, \( F_1(y_g, w) \leq 0 \). So, from Lemma 2.2 (i) and (iv), we have

\[
0 = F_1(y_g, y) + h_1(y_g, y) \\
\leq g[F_1(y_g, y) + h_1(y_g, y)] + (1 - g)[F_1(y_g, w) + h_1(y_g, w)] \\
\leq g[F_1(y_g, y) + h_1(y_g, y)] + (1 - g)[F_1(w, y_g) + h_1(w, y_g)] \\
\leq [F_1(y_g, y) + h_1(y_g, y)].
\]

Therefore, \( 0 \leq F_1(y_g, y) + h_1(y_g, y) \). From Lemma 2.2 (iii), we have \( 0 \leq F_1(w, y) + h_1(w, y) \). This implies that \( w \in \text{GEP}(F_1, h_1) \).

Next, we show that \( Aw \in \text{GEP}(F_2, h_2) \). Since \( \|u_n - x_n\| \to 0, u_n \to w \) as \( n \to \infty \) and \( \{x_n\} \) is bounded, there exists a subsequence \( \{x_{n_k}\} \) of \( \{x_n\} \) such that \( x_{n_k} \to w \), and since \( A \) is bounded linear operator, so \( Ax_{n_k} \to Aw \).

Now, setting \( v_{n_k} = Ax_{n_k} - T_{r_{n_k}} F_2 Ax_{n_k} \). It follows from (3.24) that \( \lim_{i \to \infty} v_{n_k} = 0 \) and \( Ax_{n_k} - v_{n_k} = T_{r_{n_k}} F_2 Ax_{n_k} \).

Therefore, from Lemma 2.4, we have

\[
F_2(Ax_{n_k} - v_{n_k}, z) + h_2(Ax_{n_k} - v_{n_k}, z) + \frac{1}{r_{n_k}}(z - (Ax_{n_k} - v_{n_k}), (Ax_{n_k} - v_{n_k}) - Ax_{n_k}) \geq 0, \quad \forall z \in Q.
\]
Since \( F_2 \) and \( h_2 \) are upper semicontinuous in the first argument, taking \( \limsup \) to above inequality as \( k \to \infty \) and using condition (iv), we obtain

\[
F_2(Aw, z) + h_2(Aw, z) \geq 0, \quad \forall z \in Q,
\]

which means that \( Aw \in \text{GEP}(F_2, h_2) \) and hence \( w \in \Gamma \).

**Step 6.3** We will show that \( w \in \Omega \).

Take any \( x, y \in C \). Using (2.7), we estimate

\[
\|M(x) - M(y)\|^2
\]

\[
= \|P_C[P_C(x - \mu_2 B_2 x) - \mu_1 B_1 P_C(x - \mu_2 B_2 x)] - P_C[P_C(y - \mu_2 B_2 y) - \mu_1 B_1 P_C(y - \mu_2 B_2 y)]\|^2
\]

\[
\leq \|P_C(x - \mu_2 B_2 x) - P_C(y - \mu_2 B_2 y)\|^2 - \mu_1 B_1 P_C((x - \mu_2 B_2 x) - B_1 P_C(y - \mu_2 B_2 y))\|^2
\]

\[
\leq \|x - \mu_2 B_2 x\| - \|y - \mu_2 B_2 y\|^2
\]

\[
\leq \|x - y\|^2.
\]

This implies that \( M : C \to C \) is nonexpansive. Next, we have

\[
\|y_n - M(y_n)\|
\]

\[
= \alpha_n \|f(x_n) - M(y_n)\| + (1 - \alpha_n) \|P_C[P_C(z_n - \mu_2 B_2 z_n) - \mu_1 B_1 P_C(z_n - \mu_2 B_2 z_n)] - M(y_n)\|
\]

\[
= \alpha_n \|f(x_n) - M(y_n)\| + (1 - \alpha_n) \|M(z_n) - M(k_n)\|
\]

\[
\leq \alpha_n \|f(x_n) - M(y_n)\| + (1 - \alpha_n) \|z_n - k_n\|. \quad \text{(3.41)}
\]

Since \( \alpha_n \to 0 \) and \( \|z_n - k_n\| \to 0 \) as \( n \to \infty \), (3.41) implies \( \liminf_{n \to \infty} \|y_n - M(y_n)\| = 0 \) and hence by Lemma 2.7 (ii), it follows that \( w \in M(w) \). Further, it follows from Lemma 2.1 that \( w \in \Omega \).

**Step 6.4** We will show that \( w \in \Theta \).

Indeed, let

\[
L v = \begin{cases} 
Av + N_{Cv} & \text{if } v \in C, \\
\emptyset & \text{if } v \not\in C,
\end{cases}
\]

where \( N_{Cv} := \{w \in H_1 : \langle v - u, w \rangle \geq 0, \forall u \in C\} \) is the normal cone to \( C \) at \( v \in C \). Then \( L \) is maximal monotone and \( 0 \in L v \) if and only if \( v \in \Theta \); (see [46]) for more details. Let \( G(L) \) denote the graph of \( U \) and let \( (v, u) \in G(L) \). Then we have

\[
u \in L v = Av + N_{Cv},
\]

and hence

\[
u - Av \in N_{Cv}.
\]
Next, we have
\[ \langle v - t, u - Av \rangle \geq 0, \forall t \in C. \]

Since \( z_n \in C, \forall n \), so we have
\[ \langle v - z_n, u - Av \rangle \geq 0. \tag{3.42} \]

On the other hand, it follows from \( z_n = P_C(u_n - \lambda_n Du_n) \) and \( v \in C \) that
\[ \langle v - z_n, z_n - (I - \lambda_n D)u_n \rangle \geq 0 \]
and hence
\[ \langle v - z_n, z_n - u_n + Du_n \rangle \geq 0. \]

Further, from (3.42) and inverse strongly monotonicity of \( D \), we have
\[
\langle v - z_{n_i}, u \rangle \geq \langle v - z_{n_i}, Av \rangle \\
\geq \langle v - z_{n_i}, Av \rangle - \left( \langle v - z_{n_i}, \frac{z_{n_i} - u_{n_i}}{\lambda_{n_i}} + Du_{n_i} \rangle \right) \\
= \langle v - z_{n_i}, Av - Dz_{n_i} \rangle + \langle v - z_{n_i}, Dz_{n_i} - Du_{n_i} \rangle - \left( \langle v - z_{n_i}, \frac{z_{n_i} - u_{n_i}}{\lambda_{n_i}} \rangle \right) \\
\geq \langle v - z_{n_i}, Dz_{n_i} - Du_{n_i} \rangle - \left( \langle v - z_{n_i}, \frac{z_{n_i} - u_{n_i}}{\lambda_{n_i}} \rangle \right). \tag{3.43}
\]

Since \( z_{n_i} \to w \) and \( \|u_{n_i} - z_{n_i}\| \to 0 \) as \( i \to \infty \), hence we obtain \( \langle v - w, u \rangle \geq 0. \) Since \( L \) is maximal monotone, we have \( w \in L^{-1}0 \) and hence \( w \in \Theta. \) Thus we have \( w \in \Theta. \)

**Step 7.** We claim that \( \limsup_{n \to \infty} \langle f(z) - z, x_n - z \rangle \leq 0, \) where \( z = P_\Theta f(z). \)

Since \( \{x_n\} \) is bounded, there exists a subsequence \( \{x_{n_i}\} \) of \( \{x_n\} \) such that
\[
\limsup_{n \to \infty} \langle f(z) - z, x_n - z \rangle = \limsup_{i \to \infty} \langle f(z) - z, x_{n_i} - z \rangle = \langle f(z) - z, w - z \rangle \leq 0. \tag{3.44}
\]

Further, from (2.7) and definition of \( k_n \), we have
\[
\|k_n - z\| = \|M(z_n) - M(z)\| \leq \|z_n - z\| \leq \|x_n - z\|.
\]

Next, we have
\[
\langle f(x_n) - z, y_n - z \rangle = \langle f(x_n) - z, x_n - z \rangle + \langle x_n - z, y_n - z \rangle \\
= \langle f(x_n) - f(z), x_n - z \rangle + \langle f(z) - z, x_n - z \rangle + \langle f(x_n) - z, y_n - x_n \rangle \\
\leq \rho \|x_n - z\|^2 + \|f(z) - z, x_n - z \rangle + \|f(x_n) - z\| \|y_n - x_n\|. \tag{3.45}
\]
Finally, we show that $x_n \to z$.

\[
\|x_{n+1} - z\|^2 = \|\beta_n(x_n - z) + \gamma_n(y_n - z) + \delta_n(Sy_n - z)\|^2 \\
\leq \beta_n\|x_n - z\|^2 + (\gamma_n + \delta_n)\left(\frac{1}{(\gamma_n + \delta_n)}(\gamma_n(y_n - z) + \delta_n(Sy_n - z))\right)^2 \\
\leq \beta_n\|x_n - z\|^2 + (\gamma_n + \delta_n)\|y_n - z\|^2 \\
\leq [1 - (\gamma_n + \delta_n)\alpha_n]\|x_n - z\|^2 + (\gamma_n + \delta_n)2\alpha_n\langle f(x_n) - z, y_n - z \rangle \\
\leq [1 - (1 - 2\rho)(\gamma_n + \delta_n)\alpha_n]\|x_n - z\|^2 + (\gamma_n + \delta_n)2\alpha_n\langle f(z) - z, x_n - z \rangle \\
+ \|f(x_n) - z\|\|y_n - x_n\| \\
\leq [1 - (1 - 2\rho)(\gamma_n + \delta_n)\alpha_n]\|x_n - z\|^2 + (1 - 2\rho)(\gamma_n + \delta_n)\alpha_n \\
\times 2\{(f(z) - z, x_n - z) + \|f(x_n) - z\|\|y_n - x_n\|\}\}. \\
\tag{3.46}
\]

Since $\liminf_{n \to \infty} (1 - 2\rho)(\gamma_n + \delta_n) > 0$. It follows that $\sum_{n=0}^{\infty} (1 - 2\rho)(\gamma_n + \delta_n)\alpha_n = \infty$ and hence we notice that

\[
\limsup_{n \to \infty} \frac{2\{(f(z) - z, x_n - z) + \|f(x_n) - z\|\|y_n - x_n\|\}}{1 - \rho} \leq 0, \\
\tag{3.47}
\]

since $\limsup_{n \to \infty} \langle f(z) - z, x_n - z \rangle \leq 0$ and $\lim_{n \to \infty} \|x_n - y_n\| = 0$. Thus all the conditions of Lemma 2.10 are satisfied. Hence we deduce that $x_n \to z$. This completes the proof.

**Corollary 3.2** Let $H_1$ and $H_2$ be two real Hilbert spaces and $C \subset H_1$ and $Q \subset H_2$ be nonempty closed convex subsets of real Hilbert spaces $H_1$ and $H_2$, respectively. For each $i = 1, 2$, let $D, B_i : C \to H_1$ be $\alpha_i, \beta_i$-inverse strongly monotone mappings, respectively. Let $F_1, h_1 : C \times C \to \mathbb{R}$ and $F_2, h_2 : Q \times Q \to \mathbb{R}$ satisfying Lemma 2.2; $h_1, h_2$ are monotone and $F_2$ is upper semicontinuous and $S : C \to C$ be a $k$-strict pseudocontractive mapping such that $\mathcal{Y} := F(S) \cap \Gamma \cap \Omega \neq \emptyset$. Let $f$ be a $\rho$-contraction mapping with $\rho \in [0, \frac{1}{2})$. For a given $x_0 \in C$, let the iterative sequences $\{u_n\}, \{x_n\}$ and $\{y_n\}$ be generated by

\[
\begin{cases}
  u_n = T_{r_n}^{(F_1, h_1)}(x_n + \xi A^* (T_{r_n}^{(F_2, h_2)} - I)Ax_n); \\
  y_n = \alpha_n f(x_n) + (1 - \alpha_n)P_C[P_C(x_n - \mu_2B_2u_n) - \mu_1B_1P_C(x_n - \mu_2B_2u_n)]; \\
  x_{n+1} = \beta_n x_n + \gamma_n y_n + \delta_n Sy_n,
\end{cases} \\
\tag{3.48}
\]

where $\mu_i \in (0, 2\beta_i)$, for each $i = 1, 2$, $r_n \subset (0, \infty), \xi \in (0, \frac{1}{\eta}), L$ is the spectral radius of the operator $A^*A$ and $A^*$ is the adjoint of $A$ and $\{\alpha_n\}, \{\beta_n\}, \{\gamma_n\}$ and $\{\delta_n\}$ are the sequences in $(0, 1)$ satisfying the following conditions:
Then the sequence \( \{x_n\} \) converges strongly to \( z \in P_{\Upsilon} \) where \( z = P_{\Upsilon} f(z) \).

**Proof 2** Putting \( D = 0 \) in Theorem 3.1. Then conclusion of Corollary 3.2 is obtained.

**Corollary 3.3** Let \( H_1 \) and \( H_2 \) be two real Hilbert spaces and \( C \subset H_1 \) and \( Q \subset H_2 \) be nonempty closed convex subsets of real Hilbert spaces \( H_1 \) and \( H_2 \), respectively. For each \( i = 1, 2 \), let \( B_i : C \to H_1 \) be \( \beta_i \)-inverse strongly monotone mappings. Let \( F_1, h_1 : C \times C \to \mathbb{R} \) and \( F_2, h_2 : Q \times Q \to \mathbb{R} \) satisfying Lemma 2.2; \( h_1, h_2 \) are monotone and \( F_2 \) is upper semicontinuous and \( S : C \to C \) be a nonexpansive mapping such that \( \Upsilon := F(S) \cap \Gamma \cap \Omega \neq \emptyset \). Let \( f \) be a \( \rho \)-contraction mapping with \( \rho \in (0, \frac{1}{2}) \). For a given \( x_0 \in C \), let the iterative sequences \( \{u_n\}, \{x_n\} \) and \( \{y_n\} \) be generated by

\[
\begin{align*}
\left\{ \begin{array}{l}
u_n = T_{r_n}^{(F_1, h_1)} (x_n + \xi A^* (T_{r_n}^{(F_2, h_2)} - I) Ax_n) ; \\
y_n = \alpha_n f(x_n) + (1 - \alpha_n) PC [P_C (u_n - \mu_2 B_2 u_n) - \mu_1 B_1 P_C (u_n - \mu_2 B_2 u_n)]; \\
x_{n+1} = \beta_n x_n + \gamma_n y_n + \delta_n S y_n ,
\end{array} \right. 
\end{align*}
\]

where \( \mu_i \in (0, 2\beta_i) \), for each \( i = 1, 2, r_n \subset (0, \infty), \xi \in (0, \frac{1}{\rho}), L \) is the spectral radius of the operator \( A^* A \) and \( A^* \) is the adjoint of \( A \) and \( \{\alpha_n\}, \{\beta_n\}, \{\gamma_n\} \) and \( \{\delta_n\} \) are the sequences in (0, 1) satisfying the following conditions:

(i) \( \beta_n + \gamma_n + \delta_n = 1 \);

(ii) \( \lim_{n \to \infty} \alpha_n = 0 \) and \( \sum_{n=0}^{\infty} \alpha_n = \infty \);

(iii) \( 0 < \lim \inf_{n \to \infty} \beta_n \leq \lim \sup_{n \to \infty} \beta_n < 1 \) and \( \lim \inf_{n \to \infty} \delta_n > 0 \);

(iv) \( \lim \inf_{n \to \infty} r_n > 0, \sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty \);

(v) \( \lim n \to \infty \left( \frac{\gamma_n + 1}{1 - \beta_{n+1}} - \frac{\gamma_n}{1 - \beta_n} \right) = 0 \).

Then the sequence \( \{x_n\} \) converges strongly to \( z \in P_{\Upsilon} \) where \( z = P_{\Upsilon} f(z) \).

**Proof 3** Putting \( D = 0 \) in Theorem 3.1 and taking \( S \) is nonexpansive. Then conclusion of Corollary 3.3 is obtained.
Corollary 3.4 Let $H_1$ and $H_2$ be two real Hilbert spaces and $C \subset H_1$ and $Q \subset H_2$ be nonempty closed convex subsets of real Hilbert spaces $H_1$ and $H_2$, respectively. Let $D : C \to H_1$ be $\alpha$-inverse strongly monotone mappings. Let $F_1, h_1 : C \times C \to \mathbb{R}$ and $F_2, h_2 : Q \times Q \to \mathbb{R}$ satisfying Lemma 2.2; $h_1, h_2$ are monotone and $F_2$ is upper semicontinuous and $S : C \to C$ be a nonexpansive mapping such that $\Upsilon := F(S) \cap \Gamma \cap \Omega \neq \emptyset$. Let $f$ be a $\rho$-contraction mapping with $\rho \in (0, \frac{1}{2})$. For a given $x_0 \in C$, let the iterative sequences $\{u_n\}, \{x_n\}$ and $\{y_n\}$ be generated by

$$
\begin{align*}
\begin{cases}
u_n &= T^{(F_1, h_1)}(x_n + \delta A^*(T^{(F_2, h_2)} - I)Ax_n); \\
z_n &= P_C(u_n - \lambda_n Du_n); \\
x_{n+1} &= \beta_n x_n + \gamma_n y_n + \delta_n Sy_n,
\end{cases}
\end{align*}
$$

where $r_n \subset (0, \infty), \lambda_n \in (0, 2\alpha), \xi \in (0, \frac{1}{2}), L$ is the spectral radius of the operator $A^*A$ and $A^*$ is the adjoint of $A$ and $\{\beta_n\}, \{\gamma_n\}$ and $\{\delta_n\}$ are the sequences in $(0, 1)$ satisfying the following conditions:

(i) $\beta_n + \gamma_n + \delta_n = 1$;

(ii) $0 < \lim \inf_{n \to \infty} \beta_n \leq \lim \sup_{n \to \infty} \beta_n < 1$ and $\lim \inf_{n \to \infty} \delta_n > 0$;

(iii) $\lim \inf_{n \to \infty} r_n > 0, \sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;

(iv) $\lim n \to \infty \left(\frac{\gamma_{n+1}}{1-\beta_{n+1}} - \frac{\gamma_n}{1-\beta_n}\right) = 0$;

(v) $0 < \lim \inf_{n \to \infty} \lambda_n \leq \lim \sup_{n \to \infty} \lambda_n < 2\alpha$ and $\lim_{n \to \infty} |\lambda_{n+1} - \lambda_n| = 0$.

Then the sequence $\{x_n\}$ converges strongly to $z \in P_\Upsilon$ where $z = P_\Upsilon f(z)$.

Proof 4 Putting $\alpha_n = 0, \forall n \geq 0$ and $B_1 = B_2 = 0$ in Theorem 3.1 and taking $S$ is nonexpansive. Then conclusion of Corollary 3.4 is obtained.

Acknowledgements

Financial support from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0033/2554) to student’s Jitsupa Deepho and advisor’s Poom Kumam and the King Mongkut’s University of Technology Thonburi (KMUTT) are acknowledged. Moreover, the second author was supported by the National Research Council of Thailand to the Faculty of Science and Technology, RMUTT/2557.
References


A mathematical model of a single population with habitat fragmentation in progress

Rodrigo Del-Valle\textsuperscript{1} and Fernando Córdova-Lepe\textsuperscript{2,1}

\textsuperscript{1} Facultad de Ciencias Básicas, Universidad Católica del Maule
\textsuperscript{2} Facultad de Ciencias Básicas, Universidad Metropolitana de Ciencias de la Educación
emails: rvalle@ucm.cl, fcordova@ucm.cl

Abstract

For a single species, we considered that the habitat of its population, during certain interval of time, is being fragmented by the emergence of an interior border, which divide it in two disjoint patches. Assuming logistic laws for times before and after the fragmentation interval, we connect past with future to study the abundance on the fragments while they appear. The above is got it by the postulating and analyzing of a natural differential system that rules the abundances of the transition.

Key words: Habitat fragmentation Population abundance Logistic growth

1 Introduction

Fragmentation of ecosystems are often evaluated as a progressive phenomenon directly linked to the transforming human action. Furthermore, it is considered having a very high impact as a factor of loss of biodiversity and abundance of populations, for an example in native forest fragmentation (see [1]). At a level of habitat, fragmentation is a phenomenon of transformation of the landscape, that is, a habitat that was a continuous medium after a time becomes two or more fragments.

In the literature, we recognize at least two meanings to the concept of fragmentation: First: The fragmentation can be understood as the process in which at least one of the patches which form the habitat has a division that is accompanied with reduction in the size of the sum of the emerging patches and / or a gradual isolation of these. The above generally occurs within what is called a hostile matrix (see [3]). The work, at the theoretical or practical level, with mentioned concept has an important drawback, since this concept does
not allow to separate and calibrate the different concurrent effects that can be distinguished: the division, the habitat loss and the isolation. Although in nature, the division of a patch (as a process of emergence of edge, *i.e.*, border) usually occurs in conjunction with some loss of habitat and also with continued isolation, some authors recognize that this mixture, as a concept unifying factors, is a weakness for the study of fragmentation (see [5]), it is also likely that these contributing factors have independent effects on habitat richness and abundance of populations (see [2]).

Second: The appearance of fragments as a simple division in an ideal sense, is other conceptual possibility. Then, exaggerating the reality, we can consider by fragmentation as the splitting process by the development of a border or edge of codimension one, which is not a priori a habitat loss. This means that in a habitat formed by one planar patch, emerges a border which is a curve or separatrix which divides the patch at least in other two. This option, which favorese the analysis, will be the approach assumed by the present article. We also consider that when this fragmentation has been set, there is no possibility of migration between patches, that is, it generates a total isolation.

It is recognized the need to further study the effects of fragmentation processes in the above second meaning. In [4], it is shown that experimental studies with this sense of fragmentation per se are quite few. From a total of 120 works found (period Jan. 2000 - Nov. 2003 in ISI Web of Knowledge) related to experimental studies of fragmentation, only five papers separate the effects of fragmentation of those effects of habitat loss. Restricting ourselves to study only the theoretical implications of fragmentation per se by mathematical modeling, this paper aims to contribute in this direction. Moreover, some authors suggest that modeling gives better compression expectations when it is taken at the landscape level, analyzing one or a small group of species, rather than the entire ecosystem. We limit ourselves to study the consequences on the abundance of a single population in its habitat that has a fragmentation in two patches and only during the interval of time that this process is completed.

The literature shows many works centered in try to understand the effect of the habitat fragmentation on abundance of the populations. A simple model that works with a single species, modeled by a logistic law of growth can be found in [6]. Their work resolve the following question “Does habitat fragmentation influence the abundance of organisms in the absence of habitat loss?” It also will be our main research question, but centered in the dynamic during the interval time that the process late.

The paper will be organized as follows: Firstly, in Section 2, we represent the fragmentation transition by considering an nonautonomous differential system. Secondly, in Section 3, we present the implications on the abundance of the total and partial population of the fragmentation process by means of an analytical study of the introduced model. We finish with Section 4, giving some numerical simulations to illustrate the behavior possibilities.
2 The mathematical model of fragmentation

Let us consider that the population is undergoing a process of fragmentation during a time interval \([t_i, t_f]\), \(t_i < t_f\). This population is located in a habitat represented in mathematical terms by a region (open, connected and non-empty) bounded \(\Omega\) of \(\mathbb{R}^2\). To represent the habitat division into two patches and the appearance of edge, we assume that \(\Omega\) is the union of three sets: two disjoint regions \(A\) and \(B\), and also a common border \(\Gamma = \partial A \cap \partial B\), which is an arcconnected set. Moreover, we assume that \(\Gamma\) is the graph of a regular and simple curve \(\alpha : (t_i, t_f) \to \Omega\) such that \(\alpha(t_i)\) and \(\alpha(t_f)\) exist.

Denoting \(\Gamma(t) = \{\alpha(s) : s \in (t_i, t_f)\}\) with \(t \in [t_i, t_f]\), we have that, at any instant \(t\) of the process, the habitat is represented by the set \(\Omega(t) = \Omega \setminus \Gamma(t)\). Notice that \(\Omega(t_i) = \Omega\) and \(\Omega(t_f) = A \cup B\), because \(\Gamma(t_i) = \emptyset\) and \(\Gamma(t_f) = \Gamma\) respectively. Then, at \(t = t_i\) the population is in a unique patch \(\Omega\), but at the final instant, \(t = t_f\), the population is distributed in two isolated patches \(A\) and \(B\). In set-theoretic terms, the loss of original habitat to a time \(t \in (t_i, t_f)\) is \(\Gamma(t)\). As this is a set of measure (area) null, we will say, by abuse of language, we will say that we have a process of fragmentation by arise of edge, but without habitat loss.

Our interest is in connecting population dynamics that existed before the onset of fragmentation \((t < t_i)\), with that which occurs after completion of the fragmentation, \((t > t_f)\). To simplify this scenario, we assume a before and after fragmentation of the population modeled by classical logistic equations.

To work with population densities, we will assume that the sets \(\Omega\), \(A\) and \(B\), and those considered hereafter, are quantified by a nonnegative measure (area) \(m(\cdot)\). Notice that, \(m(\Omega) = m(\Omega(t)) = m(A) + m(B)\), for each \(t \in \mathbb{R}\).

A unique patch \((t < t_i)\): If we represent by \(x_A(t)\) and \(x_B(t)\) the abundances of the population in the zones \(A\) and \(B\) respectively at time \(t\), and we assume that there are no individuals in areas of zero measure, then we will consider that the total population, \(x_A(t) + x_B(t)\), is given by the logistic equation:

\[
(x_A + x_B) = r(x_A + x_B)
\]

where \(r\) denotes the intrinsic rate of growth and \(K\) the carrying capacity of \(\Omega\).

Two patches \((t > t_f)\): The zones \(A\) and \(B\) are now physically divided and there is not migration between them (isolation). So they form two closed patches. However, there was not loss of habitat in the process. So we have two ordinary differential equation governing the growth one for any patch. This is,

\[
\begin{align*}
\dot{x}_A(t) &= r x_A(t) \left(1 - \frac{x_A(t)}{K_A}\right), \\
\dot{x}_B(t) &= r x_B(t) \left(1 - \frac{x_B(t)}{K_B}\right).
\end{align*}
\]
The transition ($t_i < t < t_f$):

We are counting the population size according to a border that is forming. It is expected that:

\[
\begin{align*}
    \dot{x}_A(t) &= x_A(t) f_t(x_A(t), x_B(t)), \\
    \dot{x}_B(t) &= x_B(t) g_t(x_A(t), x_B(t)),
\end{align*}
\]

(3)

where, $f_t(\cdot, \cdot)$ and $g_t(\cdot, \cdot)$ could be linear functions of the pair $(x_A, x_B)$, for any $t \in [t_i, t_f]$.

In order to postulate, for any $t \in [t_i, t_f]$, a form for the functions $f_t, g_t : [0, \infty]^2 \to \mathbb{R}$, we follow the classical deduction logistic model $x'(\cdot) = r(x(t))x(t)$. Here, in each patch the per capita rate of growth is decomposed into the difference $r(x) = b(x) - d(x)$, with the natality $b(x)$ as the mortality $d(x)$, are linearly affected by the density $D$. Then,

\[
b(x) = b_0 - b \cdot D \quad \text{and} \quad d(x) = d_0 + d \cdot D,
\]

(4)

where $b_0$ and $d_0$ are respectively the natality and mortality to very low densities. Moreover, $b$ and $d$ are the decreasing (resp. increasing) of the natality (resp. mortality) per unit density.

The density $D$ at the instant $t = t_i$, this is, before fragmentation, and that affects both the patch individuals $A$ and $B$, is $D = (x_A(t_i) + x_B(t_i))/(m(A) + m(B))$, where $m(\cdot)$ is some space measure, for instance, area. Substituting this value $D$ in (4), we obtain (1), with $r = b_0 - d_0$ and $K = K_A + K_B$, where

\[
K_C = m(C) \frac{b_0 - d_0}{b + d}, \quad C \in \{A, B\}.
\]

Now at $t = t_f$, the density $D$ in patch $A$ (resp. $B$) is $x_A/m(A)$ (resp. $x_B/m(B)$). Substituting this density in (4), we obtain the first equation of (2) (resp. the second).

These last two paragraphs lead us to wonder about the density at time $t$ intermediate in $[t_i, t_f]$. Note that it is natural to think that $A$, at each time $t$, while it is not completely isolated, is an area influenced (for purposes of limiting rates) by the density in a territory $B_\eta$, whose measure represents a fraction $\eta(t)$ of $m(B)$, i.e., we can work with a weighted average density:

\[
D_A(t) = p \cdot \text{(density in } A) + q \cdot \text{(density in } B_\eta),
\]

(5)

certain positives $p$ and $q$, such that $p + q = 1$.

The weights $p$ and $q$ can be assumed representing proportionally the influencing zones on the rates of $A$, this is, $A$ and $B_\eta$. As we have a measure and $m(A \cup B_\eta) = m(A) + \eta \cdot m(B)$, we suppose:

\[
p = \frac{m(A)}{m(A) + \eta \cdot m(B)} \quad \text{and} \quad q = \frac{\eta \cdot m(B)}{m(A) + \eta \cdot m(B)}.
\]

Substituting in (5), we obtain

\[
D_A(t) = \frac{x_A + \eta(t) \cdot x_B}{m(A) + \eta(t) \cdot m(B)}.
\]
If, at the same instant $t$, the focus is in zone $B$, then assuming a fraction $\xi$ of $A$ which is influencing $B$, and assuming the average perspective, we have that the density $D_B$ in area $B$ is given by:

$$D_B(t) = \frac{\xi(t) \cdot x_A + x_B}{\xi(t) \cdot m(A) + m(B)}.$$

Doing the calculations, by replacing these densities in (4) for each of the areas $A$ and $B$, we obtain the following differential system:

$$\begin{cases}
    x'_A = x_A f_t(x_A, x_B) = rx_A \left\{1 - \frac{x_A + \eta(t)x_B}{K_A + \eta(t)K_B}\right\}, \\
    x'_B = x_B g_t(x_A, x_B) = rx_B \left\{1 - \frac{\xi(t)x_A + x_B}{\xi(t)K_A + K_B}\right\},
\end{cases}$$

where $\eta, \xi : [t_i, t_f] \rightarrow [0, 1]$ are decreasing functions, such that $\eta(t_i) = \xi(t_i) = 1$ and $\eta(t_f) = \xi(t_f) = 0$. The functions $\eta(\cdot)$ and $\xi(\cdot)$ pretend to be a measure of the “connectivity” between patches $A$ and $B$ at time $t$, $t \in [t_i, t_f]$. Hereinafter, for simplicity, we assume that $\eta(\cdot)$ and $\xi(\cdot)$ are equal.

There is another route to postulate the system (6). It is the case when we are inspired in a simple competition model with variable currying capacity and competitors for any subpopulation.

Notice that equation (1) can be deduced from the addition of the equations of (6), because $\eta(t) = \xi(t) = 1$, for each $t \leq t_i$.

### 3 Main results

#### 3.1 Equal densities

**Theorem:** Let us consider that $(x_A(\cdot), x_B(\cdot))$ is a solution of (6) such that for $t = t_i$ we have $x_A(t_i)/K_A = x_B(t_i)/K_B$. Then for each $t \in [t_i, t_f]$, we have:

(a) $x_A(t)/K_A = x_B(t)/K_B$.

(b) $x_A(t) + x_B(t)$ satisfy (1).

**Remark:** The equality (a) indicates that when the regions $A$ and $B$ present at $t = t_i$ same densities, these persist over time as long as the fragmentation. In this case, with respect to the total abundance, the fragmentation makes no difference.

**Proof (a):** Indeed, considering the integral versions of the equations of the system (6), we have

$$|\alpha(t) - \beta(t)| \leq |\alpha(t_i) - \beta(t_i)| + r \int_{t_i}^{t} |\Delta(s)|ds, \quad t \in [t_i, t_f],$$

(7)
where \(\alpha(\cdot) = x_A(\cdot)/K_A, \beta(\cdot) = x_B(\cdot)/K_B\) and
\[
\Delta = \alpha \left\{ 1 - \frac{\alpha + \eta \beta(K_B/K_A)}{1 + \eta(K_B/K_A)} \right\} - \beta \left\{ 1 - \frac{\eta \alpha(K_A/K_B) + \beta}{\eta(K_A/K_B) + 1} \right\}
\]

Notice that,
\[
\Delta = (\alpha - \beta) \left\{ 1 - \frac{\alpha(1 + \lambda \eta) + \beta(1 + \lambda^{-1} \eta)}{(1 + \eta \lambda)(1 + \eta \lambda^{-1})} \right\},
\]
where \(\lambda = K_A/K_B\).

Let us define \(M = \max\{K_A + K_B, x_A(t_i), x_B(t_i)\}\). Then \(x_A(t), x_B(t) \leq M\) for each \(t \geq t_i\). Indeed, if there exists \(t_1 > t_i\) with \(x_A(t_1) > M \geq x_A(t_i)\), then there exists \(t_2 \in (t_i, t_1]\) such that \(x_A(t_2) > M\) and \(x_A'(t_2) > 0\). However, \(x_A(t_2) > K_A + K_B\) implies
\[
1 - \frac{x_A(t_2) + \eta x_B(t_2)}{K_A + \eta(t_2)K_B} < -\frac{(1 - \eta(t_2))K_B + \eta(t_2)x_B(t_2)}{K_A + \eta(t_2)K_B},
\]
this is, \(x_A'(t_2) < 0\), a contradiction. Similarly, it is possible to prove that \(x_B(t) \leq M\).

Then \(\alpha(\cdot)\) and \(\beta(\cdot)\) are bounded functions. Therefore, there exists \(\bar{M} > 0\) such that
\(|\Delta(s)| \leq \bar{M}|\alpha(s) - \beta(s)|\), for each \(s > t_i\). Using (7) and Gronwall Inequality, we obtain:
\[
|\alpha(t) - \beta(t)| \leq |\alpha(t_i) - \beta(t_i)|e^{\bar{M}(t-t_i)}, \quad t \geq t_i.
\]
Since \(\alpha(t_i) = \beta(t_i)\), we have \(\alpha(t) = \beta(t)\), for each \(t \geq t_i\), this is, (a) is proved.

Proof (b): Notice that by (a), \((x_A + \eta x_B)/(K_A + \eta K_B) = x_A/K_A\) and \((\eta x_A + x_B)/(\eta K_A + K_B) = x_B/K_B\). Then \(x_C' = r x_C(1 - x_C/K_C)\), for any \(C \in \{A, B\}\), on \([t_i, t_f]\). Therefore,
\[
(x_A + x_B)' = r(x_A + x_B)\left(1 - \frac{x_A + x_B}{K_A + K_B}\right) + r\omega,
\]
where
\[
\omega = \frac{(x_A + x_B)^2}{K_A + K_B} - \left(\frac{x_A^2}{K_A} + \frac{x_B^2}{K_B}\right).
\]
Using \(x_A/K_A = x_B/K_B\), it is straightforward to conclude that \(\omega = 0\).

### 3.2 Equilibria

It is clear that the function \(t \rightarrow (x_A(t), x_B(t)) = (0, 0)\) is a constant solution. Nevertheless, if \((x_A(\cdot), x_B(\cdot))\) is a constant solution such that \(x_A(t)x_B(t) \neq 0\), for any \(t \in [0, 1]\) (for simplicity, in what follows we will identify \([t_i, t_f]\) with \([0, 1]\), then from (6) we have:
\[
\begin{cases}
\eta(t)(K_B - x_B(0)) = x_A(0) - K_A, \\
\eta(t)(x_A(0) - K_A) = K_B - x_B(0).
\end{cases}
\]
From (9), we have that \(x_A(0) = K_A\) if only if \(x_B(0) = K_B\). For other hand, if \(x_B(0) \neq K_B\) or \(x_A(0) \neq K_A\), then combining both identities in (9), \(\eta^3(t) = 1\), for any \(t \in [0, 1]\), i.e., which is a contradiction. Then, the function \((x_A(\cdot), x_B(\cdot))\) equals to \((K_A, K_B)\), is the unique non trivial constant solution.
3.3 Growth

In order to determine some properties of the trajectories defined by (2)–(6), we divided the states set in six zones according: \( Z(1) \): \( x_A \leq K_A \) and \( x_B \leq K_B \); \( Z(2) \): \( x_A \geq K_A \) and \( x_B \geq K_B \); \( Z(3) \): \( x_A \leq K_A, x_B > K_B \) and \( x_A + x_B < K_A + K_B \); \( Z(4) \): \( x_B \leq K_B, x_A > K_A \) and \( x_A + x_B > K_A + K_B \); \( Z(5) \): \( x_B < K_B, x_A > K_A \) and \( x_A + x_B < K_A + K_B \); and \( Z(6) \): \( x_B < K_B, x_A < K_A \) and \( x_A + x_B > K_A + K_B \). See Figure (2).

Given a point \((x_A, x_B)\) in some \( Z(i) \), \( i = 1, \ldots, 6 \), at some time \( t \in [0, 1] \), we present in the table (10) an overview of the signs of \( x'_A(t) \) and \( x'_B(t) \) at the same instant.

\[
\begin{array}{|c|c|c|c|c|c|c|c|c|}
\hline
Z(i) & (1) & (2) & (3) & (3) & (4) & (4) & (5) & (6) \\
\hline
x_A & + & - & + & + & + & - & - & + & + & - \\
x_B & + & - & + & - & - & - & - & + & + & + \\
t \in & [0, 1] & [0, 1] & [0, t_1] & [t_1, 1] & [0, t_1] & [t_1, 1] & [0, t_2] & [t_2, 1] & [0, t_2] & [t_2, 1] \\
\hline
\end{array}
\tag{10}
\]

Notice that reordering the terms of (6), the signs of \( f_t(x_A, x_B) \) and \( g_t(x_A, x_B) \) are given respectively by the signs of the function that follows:

\[ F(x_A, x_B) = (K_A - x_A) + \eta(K_B - x_B) \]

and

\[ G(x_A, x_B) = \eta(K_A - x_A) + (K_B - x_B). \]
If \((x_A(\cdot), x_B(\cdot))\) is in \(\mathbf{Z}(1)\), then \(K_A - x_A > 0\) and \(K_B - x_B > 0\) and is clear that \(F(x_A, x_B)\) and \(G(x_A, x_B)\) are positive numbers. In the \(\mathbf{Z}(2)\) the argument is similar for obtaining that \(F(x_A, x_B)\) and \(G(x_A, x_B)\) are negative numbers.

Now, if the point \((x_A, x_B)\) is in \(\mathbf{Z}(3)\), then \(x_A < K_A, x_B > K_B\) and \((K_A + K_B) - (x_A + x_B) > 0\). Noting that \(F(x_A, x_B)\) is equal to \([(K_A + K_B) - (x_A + x_B)] + (x_B - K_B)\) then: \(x_A' = 0\). In order to determine the sign of \(x_B\), notice that \(\lambda = (x_B - K_B)/(K_A - x_A)\) is a positive number less that one. So, there exists \(t_1 \in [0, 1]\) such that: If \(t < t_1\) (resp. \(t > t_1\)), then \(\eta > \lambda\) (resp. \(\eta < \lambda\)). Since, \(G(x_A, x_B) = (\eta - \lambda)(K_A - x_A)\), we have \(x_B' > 0\) (resp. \(<\)).

Since in \(\mathbf{Z}(4)\) we have that \(F(x_A, x_B)\) equal to \(-\eta[(x_A + x_B) - (K_A + K_B)] - (1 - \eta)(x_A - K_A)(1 - \eta)\), a negative number, then \(x_A' < 0\). For other hand, \(\lambda = (K_B - x_B)/(x_A - K_A) < 1\) and \(G(x_A, x_B) = (\lambda - \eta)(x_A - K_A)\). Notice that \(t < t_1\) (resp. \(t > t_1\)) implies \(\eta > \lambda\) (resp. \(<\)), i.e., \(x_B' < 0\) (resp. \(>\)).

Considering \(q = (K_A - x_A)/(x_B - K_B)\), in \(\mathbf{Z}(5)\), we can express \(F(x_A, x_B) = (K_B - x_B)\) and \(G(x_A, x_B)\) equals to \(-[(x_A + x_B) - (K_A + K_B)] - (1 - \eta)[K_A - x_A]\). Then \(x_B'\) or \(x_A'\) depending if \(q < \eta\) or \(q > \eta\), this is, \(t < t_2\) or \(t > t_2\), where \(\eta(t_2) = q\).

Finally, in \(\mathbf{Z}(6)\) it is convenient to express function \(F\) and \(G\) by \((K_B - x_B)(\eta - q)\) and \((K_B - x_B)(1 - \eta)\). Then \(x_B' > 0\) and \(x_A' > 0\) (resp. \(<\)) if \(t < t_2\) (resp. \(>\)).

Notice that if \((x_A, x_B)\) satisfies \(x_A + x_B = K_A + K_B\) (i.e., is on the antidiagonal line), then:

\[
F = -(1 - \eta)(K_B - x_B) = (1 - \eta)(K_A - x_A) = G.
\]

So that, in the border of:

i) \((3)\) and \((5)\), \(F > 0\) and \(G < 0\), this is, \(x_A' > 0\) and \(x_B' > 0\).

ii) \((4)\) and \((6)\), \(F < 0\) and \(G > 0\), this is, \(x_A' > 0\) and \(x_B' > 0\).

To know in which direction the vector \((x_A', x_B')\) indicates, we need to do the comparison

\[
\frac{|x_A'|}{x_B'} = \frac{x_A \eta K_A + K_B}{x_B K_A + \eta K_B} = \frac{(x_A/K_A) \eta K_A + 1}{(x_B/K_B) 1 + \eta K_B}.
\]

This expression also has the following formulation:

\[
\frac{|x_A'|}{x_B'} = 1 + \frac{x_B K_B - x_A K_A}{x_B (K_A + \eta K_B)}(\bar{\lambda} - \eta),
\]

where \(\bar{\lambda} = (x_A K_B - x_B K_A)/(x_B K_B - x_A K_A)\).

Over the line \(x_A + x_B = K_A + K_B\), we have the possibilities that follows:
If $K_A = K_B$, then $|x'_A/x'_B| = x_A/x_B$. So that, in the common border of $Z(3)$ and $Z(5)$ we have $|x'_A| < |x'_B|$. Similarly, on the border of $Z(4)$ and $Z(6)$ the inequality is $|x'_A| > |x'_B|$.

If $K_A > K_B$, we have two cases. When $x_A > K_A$ and $x_B < K_B$, in whose case, by (11), it is clear that $|x'_A| > |x'_B|$. The case $C: [x_A < K_A, x_B > K_B]$ contains other two possibilities: $[C: x_A < x_B]$ and $[C: x_A > x_B]$. In case $[C: x_A < x_B]$, we have $\tilde{\lambda} > 1$, then by (12) it is concluded $|x'_A| > |x'_B|$.

If $K_A < K_B$, there are also two cases. Firstly, $D: [x_A > K_A, x_B < K_B]$, that will be open in the subcases $[D: x_A < x_B]$ and $[D: x_A > x_B]$. In case $[D: x_A < x_B]$ we have $\tilde{\lambda} > 1$, then by (12) it is concluded $|x'_A| > |x'_B|$. Nevertheless, the case $x_A < K_A$ and $x_B > K_B$, by (11), implies $|x'_A| < |x'_B|$.

In cases $[C: x_A > x_B]$ or $[D: x_A > x_B]$, it can be proved that $\lambda < 1$, so that there exists $t \in [0, 1]$ such that $\eta(t) = \tilde{\lambda}$. So if $t < \tilde{t}$ (resp. $>$), we have $\eta > \tilde{\lambda}$ (resp. $<$) and by (12) it follows $|x'_A| < |x'_B|$ (resp. $>$).

Figure 2: The arrows in zone (1) and (2) indicate the direction of movement of trajectory in the zone. For instance, the arrows forming an “L” in Zone (3) indicate a first direction on a time interval $[0, t_1]$ and the other direction the complementary time, $[t_1, 1]$. Notice that (1) and (2) are repulsor zones. Starting in a particular zone, a trajectory can be remain in this one or continue to another according to the following combinations: $(1) \rightarrow (3) \rightarrow (5)$, $(1) \rightarrow (6) \rightarrow (4)$, $(2) \rightarrow (5) \rightarrow (3)$ or $(2) \rightarrow (4) \rightarrow (6)$. 

©CMMSE
4 Numerical Simulations

Let us consider an interval of fragmentation $[t_i, t_f] = [20, 70]$ a subset of a horizon time $[0, T] = [0, 100]$. We take parameters:

$$
\begin{array}{ccccccc}
\hline
b_0 & d_0 & b & d & m(A) & m(B) \\
2/10 & 1/10 & 2/100 & 1/100 & 60 & 40 \\
\hline
\end{array}
$$

(13)

Notice that with this parameters: $r = 1/10$, $K_A = 600$ and $K_B = 400$.

Considering a linear connectivity $\eta(t) = \xi(t) = (t_f - t)/(t_f - t_i)$, $t \in [t_i, t_f]$, and initial conditions:

$$
\begin{array}{cc}
S1 & x_A(t_i) \quad x_B(t_i) \\
40 & 60 \\
S2 & x_A(t_i) \quad x_B(t_i) \\
50 & 10 \\
\end{array}
$$

we have the behavior showed in figures (3) and (4) for simulations $S1$ and $S2$ respectively, where continuous (resp. dashed) line is the situation with (resp. without) fragmentation.

![Total biomass vs time](image1)

![Planar Trajectory](image2)

Figure 3: With equal initial densities the total abundances are always the same.

Acknowledgment: This work was partially financed by the Fondecyt 1120218.
Figure 4: With independence of function of connectivity, we conjecture that different initial densities imply a total abundance with fragmentation below the unfragmented case.

References


Weighted Tridiagonal Matrix Enhanced Multivariance Products Representation of Finite Interval Data

Emre Demiralp

1 Adobe, 345 Park Ave, San Jose, 95110, California, USA
emails: emre.demiralp@gmail.com

Abstract

In another paper of this symposium we have presented a new matrix decomposer based on consecutive utilization of Enhanced Multivariate Products Representation (EMPR). This method decomposes any square or rectangular matrix to three matrix factors such that the pre and post factors are orthonormal matrices whose columns are recursively generated from a couple of initial vectors. The mid factor of this triple is a tridiagonal rectangular matrix. In this paper, we briefly recall TMEMPR first. Then we extend its definition by introducing weight matrices. Weighted TMEMPR enables us to give different importances to different elements of data.

Another important issue is the fact that each datum in the input matrix elements is generally located in a finite interval. For such cases the TMEMPR approximants may have elements outside the abovementioned finite interval. Then the values outside this finite interval should be renormalized. We present two approaches to this end. Certain illustrations supporting our ideas are also presented.

Key words: Tridiagonal Matrix Enhanced Multivariance Products Representation, Support Vectors, Singular Value Decomposition, Weight Matrices, Renormalizations.

1 Introduction

This work takes its roots from another work which is first companion[1] of this presentation in the same symposium. Hence, we are not going to repeat the introductory portions of that companion paper, but it will be better to revisit certain mainlines of TMEMPR presentation given there. TMEMPR can be formulated as follows for a given matrix \( A \) which may not need to be square. To have more generality, we assume that it is \( \ell \times m \) type where \( \ell \) and
$m$ are positive integers. To facilitate the analysis we start with the case where $\ell < m$. TMEMPR for this case can be written as follows

$$A = \sum_{i=1}^{\ell} \alpha_i u_i v_i^T + \sum_{i=1}^{\ell-1} \beta_i u_{i+1} v_i^T + \sum_{i=1}^{\ell} \gamma_i u_i v_{i+1}^T = U \Sigma V^T, \quad \ell < m$$

where the columns of the orthonormal matrix $U$ are composed of $u$ vectors (in rightward ascending indices) which are mutually orthonormal. This remains valid when $U$ matrix and $u$ vectors are replaced with the orthonormal matrix $V$ and $v$ vectors which are mutually orthonormal. $\Sigma$ which has the type $\ell \times m$ is tridiagonal and its main diagonal elements, $\alpha_i$s are in downwardly increasing index order $\alpha_i$ while the nearest lower and upper main diagonal elements are respectively indexed $\beta$ and $\gamma$ parameters. The highest indices of $\alpha$s and $\gamma$s is $\ell$ while $\beta$s stop at the highest value, $\ell - 1$, when $\ell < m$.

If $m < \ell$ then the $\alpha$ and $\beta$ parameters can have at most $m$ as highest index while $\gamma$s stop at the highest index $m - 1$. The case where $\ell = m$ the highest indices of $\beta$ and $\gamma$ parameters is $\ell - 1 = m - 1$ while $\alpha$ parameter indices climbs up to and including $\ell = m$.

The columns of the matrices $U$ and $V$ which are the mutually orthonormal vector sets $\{u_1, ..., u_\ell\}$ and $\{v_1, ..., v_m\}$ respectively, can be defined through the following recursions

$$u_{j+2} \equiv \frac{1}{\beta_{j+1}} (I_\ell - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} v_{j+1}, \quad v_{j+2} \equiv \frac{1}{\gamma} (I_m - v_{j+1} v_{j+1}^T) A_{1,2}^{(j)} u_{j+1},$$

$$A_{1,2}^{(j+1)} = (I_\ell - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} (I_m - v_{j+1} v_{j+1}^T)$$

$$= \left( I_\ell - \sum_{k=1}^{j+1} u_k u_k^T \right) A \left( I_m - \sum_{k=1}^{j+1} v_k v_k^T \right),$$

which can be initialized as follows

$$A_{1,2}^{(0)} \equiv A, \quad u_1 \equiv u, \quad v_1 \equiv v$$

where $A$ is the target matrix to be decomposed while $u$ and $v$ stand for the given normalized $\ell$ and $m$ element vectors respectively. $u$ and $v$ correspond to support vectors of the bivariate discrete EMPR. One of the general tendencies is to choose these support vector as entities whose all elements are same. In this case we can write

$$u \equiv \left[ \frac{1}{\sqrt{\ell}} \ldots \frac{1}{\sqrt{\ell}} \right]^T, \quad v \equiv \left[ \frac{1}{\sqrt{m}} \ldots \frac{1}{\sqrt{m}} \right]^T.$$  

These are in fact the support vectors of the HDMR for the matrix $A$.

The $\alpha$, $\beta$, and, $\gamma$ parameters can be directly evaluated by using the recursively evaluated $u$ and $v$ vectors through the following formulae

$$\alpha_{j+1} \equiv u_{j+1}^T A_{1,2}^{(j+1)} v_{j+1}, \quad \beta_{j+1} \equiv \left\| (I_\ell - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} v_{j+1} \right\|,$$

$$\gamma_{j+1} \equiv \left\| (I_m - v_{j+1} v_{j+1}^T) A_{1,2}^{(j)} u_{j+1} \right\|.$$
It is not hard to show that the above recursions stop when the vector $u_\ell$ (for $\ell < m$, in the case where $m < \ell$ we need to consider the vector $u_m$ instead) evaluated since first of the following projective equations (the second for $m < \ell$) holds
\[
(I_\ell - \sum_{k=1}^{\ell} u_k u_k^T) = 0_{\ell \times \ell}, \quad (I_m - \sum_{k=1}^{m} v_k v_k^T) = 0_{m \times m}
\]
where $0_{\ell \times \ell}$ and $0_{m \times m}$ denote $\ell \times \ell$ and $m \times m$ type zero matrices respectively. These vanishing properties imply that the matrix $A_{1,2}^{(j+1)}$ identically vanishes when $(j+1)$ matches $\min(\ell, m)$, by leaving no further chance of proceeding through the above recursions. This is the reason why we have a finite tridiagonal structuring.

The matrix $\Sigma$ can be approximated by replacing all $\alpha$, $\beta$, and, $\gamma$ parameters with indices greater than a prescribed positive integer value by zero. We call such approximants “Truncation Approximants” or simply “Truncations”. The approximation quality can be controlled by choosing appropriate truncation levels.

2 Weighted TMEMPR

We have used the weightless inner product definition in the construction of TMEMPR. Therein, the inner product of two vectors, say $x$ and $y$, has been defined as follows
\[
(x, y) \equiv x^T y \equiv y^T x \equiv (x, y)
\]
where $x$ and $y$ belong to the same Cartesian space whose dimension is, say, $n$. If we change this definition to a weighted one then we need to write
\[
(x, y) \equiv x^T W y \equiv y^T W x \equiv (x, y)
\]
where $W$ is a matrix mapping from the space where $x$ and $y$ lay in the same space. Hence it needs to be square first of all. To provide the inner product symmetry, $W$ has to be symmetric. This is compulsory and enables $W$ to have real eigenspectrum. Another “must” in the structure of an inner product is the production of a norm when its both arguments are same. This, however, enforces the positive definiteness of the matrix $W$. Otherwise, if $W$ is positive semi definite then $W$’s nullspaces are not empty. This makes the norm of any vector laying in the nullspace vanishing. Hence positive definiteness is unacceptable. If $W$ would have negative eigenvalue(s) then the norm of certain vectors which are expressible only in terms of eigenvectors corresponding negative eigenvalues, become negative. This is also unacceptable. Thus, as conclusion, $W$ must be symmetric and positive definite.

To introduce weight matrix concept into TMEMPR we need to use two matrices because we have two different Cartesian spaces one of which has the dimension $\ell$ and the vectors $u$s lie therein while the other has the dimension $m$ and the vectors $v$s lie therein. We denote
the weight matrices laying in the $\ell$ and $m$ dimensional Cartesian matrices by $W_L$ and $W_R$ respectively. Hence, the matrices $W_L$ and $W_R$ are of type $\ell \times \ell$ and $m \times m$ respectively and both of them are symmetric and positive definite.

The utilization of the abovementioned weight matrices are reflected to the structures of the previous formulation through the following recursions

$$
\begin{align*}
  u_{j+2} & \equiv \frac{1}{\beta_{j+1}} (I_{\ell} - u_{j+1}u_{j+1}^T W_L) A_{1,2}^{(j)} W_R v_{j+1}, \\
v_{j+2} & \equiv \frac{1}{\gamma} (I_m - v_{j+1}v_{j+1}^T W_R) A_{1,2}^{(j)} W_L u_{j+1}, \\
  A_{1,2}^{(j+1)} & = (I_{\ell} - u_{j+1}u_{j+1}^T W_L) A_{1,2}^{(j)} (I_m - W_R v_{j+1}v_{j+1}^T) \\
  & = \left( I_{\ell} - \sum_{k=1}^{j+1} u_k u_k^T W_L \right) A_{1,2}^{(j)} \left( I_m - \sum_{k=1}^{j+1} W_R v_k v_k^T \right).
\end{align*}
$$

(9)

with the following initializations

$$
A_{1,2}^{(0)} \equiv A, \quad u_1 \equiv u, \quad v_1 \equiv v.
$$

(10)

The $\alpha$, $\beta$, and, $\gamma$ parameters are given for this case through the following equalities

$$
\begin{align*}
  \alpha_{j+1} & \equiv u_{j+1}^T W_L A_{1,2}^{(j)} W_R v_{j+1}, \quad \beta_{j+1} \equiv \left\| (I_{\ell} - u_{j+1}u_{j+1}^T W_L) A_{1,2}^{(j)} v_{j+1} \right\|, \\
  \gamma_{j+1} & \equiv \left\| (I_m - W_R v_{j+1}v_{j+1}^T) A_{1,2}^{(j)} u_{j+1} \right\|.
\end{align*}
$$

(11)

One of the following equalities stops the above recursions depending on the values of $\ell$ or $m$

$$
\begin{align*}
  \left( I_{\ell} - \sum_{k=1}^{\ell} u_k u_k^T W_L \right) = 0_{\ell \times \ell}, \quad \left( I_m - \sum_{k=1}^{m} W_R v_k v_k^T \right) = 0_{m \times m}.
\end{align*}
$$

(12)

3 Varying Finite Data Interval in Truncation Approximants and Renormalization

We can define the following truncation approximants for numerical implementations

$$
T_n \equiv \sum_{i=1}^{j} \left( \alpha_i u_i v_i + \beta_i u_{i+1} v_i + \gamma_i u_i v_{i+1} \right), \quad j = 1, 2, ..., \ell
$$

(13)

where we focused on the case where $\ell < m$, hence $\beta_{\ell} = 0$. The tridiagonal parameters $\alpha$s, $\beta$s, $\gamma$s depend on the given data matrix $A$ and on the initial generators $u$ and $v$. If the data
matrix elements are confined in a finite interval on the real number line. Even though the
above truncation matrices are approximants to the data matrix, this does not guarantee
that the approximant matrix elements will always stay in the finite data interval. There
may be shifts, contractions or swellings in the approximant data interval. All observations
on experimental implementations show that at least one of this changes is encountered.
This urges us to renormalize the approximants. In other words, we need to redefine the
truncation approximants such that the data interval of each approximant matches the finite
interval data.

One way for the renormalization is to construct the truncation approximant to be
focused on and then to change its each element as follows: (i) if the element remains in the
finite data interval then keep it as is; (ii) if the element is greater than the upper end of
the finite data interval then change it to the upper end of the interval; (iii) if the element
is less than the lower end of the finite data interval then change it to the lower end of the
interval. This routine is generally applied to the images when they are processed, and, the
softwares for this processing has this renormalizing feature by default generally.

The other way we want to mention here is the evaluation of the maximum and minimum
of the truncation approximant matrix elements. Then the next thing to do is to use an
affine transformation on the approximant matrix to create a new matrix such that its all
elements remain in the same finite data interval. This change may reduce the quality of
the approximants as it may happen in the previous way. The cost of the construction may
also increase. However, each approximant can be rasterized by any existing tool in software
packages to be used in the implementation.

4 Concluding Remarks

This work has been designed as a companion to another paper of this symposium [1]. We
have introduced the possibility of using weight matrix in the formulation. The goal has been
to give different importances to each data matrix element in the approximant construction.
We have also focused on the case where the data matrix elements are in a finite interval
on the real axis. The interval shifts or size changes in the element finite interval of the
approximants have been renormalized by using two different approaches. We plan to give
more practical details in the post conference publication.

References

Products Representation (TMEMPR) as a Matrix Decomposer The Proceedings of 14th
International Conference of Computational and Mathematical Methods in Science and
Tridiagonal Matrix Enhanced Multivariance Product Representation (TMEMPR) for Matrix Decomposition

Emre Demiralp\(^1\) and Metin Demiralp\(^2\)

\(^1\) Adobe, 345 Park Ave, San Jose, 95110, California, USA

\(^2\) Informatics Institute, İstanbul Technical University, Maslak 34469, İstanbul, Türkiye (Turkey)

emails: emre.demiralp@gmail.com, metin.demiralp@gmail.com

Abstract

Enhanced Multivariate Products Representation (EMPR) is an extended form of High Dimensional Model Representation (HDMR) first proposed by Sobol. It has been developed in Demiralp’s group by increasing the multivariance of each HDMR component via multiplying it by univariate “Support Functions”. EMPR has been used not only for continuous multivariate functions but also for multiway arrays and “Tensors”. In these cases, integrals are replaced with the corresponding sums to handle discreteness. In this work, we focus on the utilization of EMPR for matrices. Even though the representation is composed of only four additive terms, the bivariate component of EMPR can be further expanded to another EMPR with support vector couples whose each member is orthonormal to the original EMPR’s corresponding support vector. This action can be iteratively used until the bivariate remainder becomes a pure outer product of two vectors. We have proven that this iterative method is in fact a tridiagonalization procedure and tightly related to the singular value decomposition of the target matrix. Appropriate choice of the initial support vectors takes us to a diagonal representation which is the singular value decomposition of the matrix under consideration.

**Key words:** High Dimensional Model Representation, Enhanced Multivariance Products Representation, Support Vectors, Singular Value Decomposition.
1 Introduction

Enhanced Multivariance Product Representation (EMPR) of a given multivariate function $f(x_1, ..., x_N)$ can be given as follows

$$f(x_1, ..., x_N) = f_0 \prod_{i=1}^{N} s_i(x_i) + \sum_{i=1}^{N} f_i(x_i) \prod_{j=1}^{N} s_j(x_j) + \sum_{i,j=1}^{N} f_{i,j}(x_i, x_j) \prod_{k=1}^{N} s_k(x_k) + \cdots \quad (1)$$

where $s_i(x_i)$s stand for the given univariate support functions whose norms are assumed to be one. Indexed $f$s are unknown EMPR components and can be determined by using the vanishing conditions which are given as “any univariate integral of an EMPR component except $f_0$ after its multiplication by the relevant support function vanishes over the prescribed integration interval. In the Sobol’s HDMR case each independent variable was considered lying in the interval $[0, 1]$ and the integration’s weight function was the constant function whose integral is 1. Despite, Rabitz has generalized the geometry from unit hypercube to any rectangular hyperprism and brought the nonconstant weight function we are going to use the Sobol’s case geometry and the constant weight here.

Enhanced Multivariate Products Representation above matches High Dimensional Model Representation for the given target function when all support functions are taken as unit constant function. The number of the additive terms in EMPR and therefore HDMR is finite and equals to $2^N$. Each additive term in EMPR and HDMR contains just a single component function denoted by an $f$ indexed in accordance with the multivariance of that component. Hence, in HDMR, there is an ordering in ascending multivariance while EMPR feeds each additive term with appropriate univariate support function factors to get highest multivariance in each additive term.

The finiteness of the additive terms in EMPR does not leave any uniform convergence problem. However, especially when $N$ grows pretty much, the number of the terms may increase too much in practical sense. This brings the truncation approximants to mind. The general tendency is to use constant, constant and univariate, and at most, constant and univariate and bivariate components. This however works well generally when the original function has dominancy in these components. Since this may not be the case always, some different type EMPRs and HDMRs have been developed. However, they are out of the scope of this work.

Abovementioned truncation approximants may not give desired numerical qualities unless the target function does not have quite specific structures. This may urge us to develop new refinement procedures to get desired numerical qualities just by using only EMPR philosophy. This is the main task of us in our recent studies. However, in this work, we start to proceed by focusing on the simplest multivariance and try to develop a scheme which produces approximants whose numerical qualities can be gradually controlled under the warranty of getting ultimate limit. Therefore we confine ourselves into the case...
of bivariance over the discrete value taking independent variables. Ordinary algebraic two dimensional arrays, that is, matrices are chosen our basic target in this work.

2 EMPR for Bivariate Functions and Two-way Arrays

Enhanced Multivariate Products Representation for a bivariate function, \( f(x_1, x_2) \), can be written as follows

\[
f(x_1, x_2) = f_0 u(x_1) v(x_1) + f_1(x_1) v(x_2) + u(x_1) f_2(x_2) + f_{1,2}(x_1, x_2) \tag{2}
\]

where \( u(x_1) \) and \( v(x_2) \) stand for the support functions which satisfy the following unit norm equalities

\[
\int_0^1 dx_1 u(x_1)^2 = 1, \quad \int_0^1 dx_2 v(x_2)^2 = 1 \tag{3}
\]

while the vanishing conditions can be given as follows

\[
\int_0^1 dx_1 f_1(x_1) u(x_1) = 0, \quad \int_0^1 dx_2 f_2(x_2) v(x_2) = 0, \\
\int_0^1 dx_1 f_{1,2}(x_1, x_2) u(x_1) = 0, \quad \int_0^1 dx_2 f_{1,2}(x_1, x_2) v(x_2) = 0. \tag{4}
\]

These conditions are sufficient to get unique solutions for the EMPR components, \( f_0, f_1(x_1), f_2(x_2), f_{1,2}(x_1, x_2) \). The results for these solutions are given below

\[
f_0 = \int_0^1 dx_1 \int_0^1 dx_2 f(x_1, x_2) u(x_1) v(x_2), \\
f_1(x_1) = \int_0^1 dx_2 f(x_1, x_2) v(x_2) - f_0 u(x_1), \\
f_2(x_2) = \int_0^1 dx_1 f(x_1, x_2) u(x_1) - f_0 v(x_2), \\
f_{1,2}(x_1, x_2) = f(x_1, x_2) - f_0 u(x_1) v(x_2) - f_1(x_1) v(x_2) - f_2(x_2) u(x_1). \tag{5}
\]

If the target function varies with the variables taking discrete values then EMPR can be written as follows

\[
a_{i,j} = a_0 u_i v_j + a^{(1)}_{i,j} v_j + u_i a^{(2)}_{j} + a^{(1,2)}_{i,j}, \quad i = 1, 2, \ldots, \ell, \quad j = 1, 2, \ldots, m \tag{6}
\]

where \( a_{i,j} \) stands for the two-way array’s (matrix of ordinary linear algebra) general element while \( a_0 \) denotes the constant EMPR component. \( a^{(1)}_{i} \) and \( a^{(2)}_{j} \) represent the general elements of the univariate EMPR components (vector of ordinary linear algebra). The
rightmost EMPR component is a matrix whose element at the intersection of \( i \)th row and \( j \)th column is denoted by \( a_{i,j}^{(1,2)} \). All these EMPR components are unknowns yet. However the one index entities (vectors) denoted by \( u_i \) and \( v_j \) are the support vectors which are assumed to be in unit norm explicitly given through the following equations

\[
\sum_{i=1}^{\ell} u_i^2 = 1, \quad \sum_{j=1}^{m} v_j^2 = 1. \tag{7}
\]

The vanishing conditions can be written as follows for this case

\[
\sum_{i=1}^{\ell} u_i a_i^{(1)} = 0, \quad \sum_{j=1}^{m} v_j a_j^{(2)} = 0, \quad \sum_{i=1}^{\ell} u_i a_{i,j}^{(1,2)} = 0, \quad \sum_{j=1}^{m} v_j a_{i,j}^{(1,2)} = 0 \tag{8}
\]

which are sufficient for unique determination of the EMPR components. The following results can be given for the EMPR components

\[
a_0 = \sum_{i=1}^{\ell} \sum_{j=1}^{m} u_i v_j a_{i,j}, \quad a_i^{(1)} = \sum_{j=1}^{m} v_j a_{i,j} - a_0 u_i, \quad a_j^{(2)} = \sum_{i=1}^{\ell} u_i a_{i,j} - a_0 v_j,
\]

\[
a_{i,j}^{(1,2)} = a_{i,j} - a_0 u_i v_j - a_i^{(1)} v_j - a_j^{(2)} u_i, \quad i = 1, ..., \ell, \quad j = 1, ..., m. \tag{9}
\]

3 Tridiagonal Enhanced Multivariance Products Representation (TMEMPR)

EMPR given by (6) contains four additive terms giving exactly the original array \( a_{i,j} \) which is in fact an ordinary linear algebraic matrix (we emphasize on this linear algebra to distinguish it from the “Multivariate Linear Algebra” which has quite rigorous properties). For truncation approximants we can take just the constant component including term which is in fact proportional to the outer product of the vectors \( \mathbf{u} \) and \( \mathbf{v} \). The quality of this truncation will apparently be quite poor unless the target matrix \( \mathbf{A} \) is a one-rank entity and beyond that proportional to this outer product. The addition of the univariate component including terms may not make the truncation quality good enough since they are also related to some other outer products. The matrix component involving last term is generally more important portion of the representation if its rank and norm is not sufficiently low. Hence, for the matrices, just EMPR’s itself may not be used efficiently when its matrix component becomes dominant in norm and/or rank. This urges us to develop a new representation which somehow disintegrates the matrix component to rather simple, and preferably, one-rank entities such that the final form of the representation contains sufficiently many components to make the truncations more meaningful.
Now we will prefer to use the concise notations for the vectors and matrices for brevity. In this approach we can rewrite the EMPR of a given matrix $A$ as follows

$$A = a_0 uv^T + a_1 v_1^T + u a_2^T + A_{1,2}$$

(10)

where $u$ and $v$ respectively stand for given $\ell$-element and $m$-element support vectors. The constant EMPR component denoted by $a_0$ is a scalar while the vector EMPR components, $a_1$ and $a_2$ have respectively $\ell$ and $m$ elements. The matrix EMPR component $A_{1,2}$ is of $\ell \times m$ type.

The support vectors have the unit norm properties given as

$$u^T u = 1, \quad v^T v = 1$$

(11)

while the vanishing conditions on the vector and matrix EMPR components can be given as follows

$$u^T a_1 = 0, \quad v^T a_2 = 0, \quad u^T A_{1,2} = 0^T_m, \quad A_{1,2} v = 0_\ell$$

(12)

where the indexed boldface zero symbols stand for the zero vectors whose number of elements are given by their subindices. This can take us to the following explicit results

$$a_0 = u^T A v, \quad a_1 = (I_\ell - uu^T) A v, \quad a_2 = (I_m - vv^T) A^T u,$$

$$A_{1,2} = (I_\ell - uu^T) A (I_m - vv^T)$$

(13)

where $I_\ell$ and $I_m$ respectively stand for the identity matrices on $\ell$ and $m$ dimensional Cartesian spaces.

We can now regard the vector EMPR components as if extra support vectors. To this end we can write the definitions $u_1 \equiv u$ and $v_1 \equiv v$ and

$$u_2 \equiv \frac{1}{\|a_1\|} a_1, \quad v_2 \equiv \frac{1}{\|a_2\|} a_2, \quad \beta_1 \equiv \|a_1\|, \quad \gamma_1 \equiv \|a_2\|$$

(14)

which urges us to rewrite the matrix EMPR as follows

$$A = \alpha_1 u_1 v_1^T + \beta_1 u_2 v_1^T + \gamma_1 u_1 v_2^T + A_{1,2}^{(1)}$$

(15)

where we have used the superscript (1) to imply the first step of a recursive scheme. The first three additive terms of this equation at the right hand side correspond to a tridiagonal decomposition over the orthonormal vectors $u_1$ and $u_2$ together with the orthonormal vectors $v_1$ and $v_2$. On the other hand, the matrix EMPR component has a rank one less than the one for original matrix because of its left and right projective matrix factors. (15) urges us to get a further EMPR decomposition over the last additive term of the right hand
side. We can write the following equation after certain intermediate steps we do not intend to explicitly give

\[ A_{1,2}^{(1)} = \alpha_2 u_2 v_2^T + \beta_2 u_3 v_2^T + \gamma_2 u_2 v_3^T + A_{1,2}^{(2)} \]  \tag{16}

where

\[ \alpha_2 \equiv u_2^T A_{1,2}^{(1)} v_2, \quad \beta_2 \equiv \left\| (I_{\ell} - u_2 u_2^T) A_{1,2}^{(1)} v_2 \right\|, \quad \gamma_2 \equiv \left\| (I_m - v_2 v_2^T) A_{1,2}^{(1)} u_2 \right\|, \]

\[ u_3 \equiv \frac{1}{\beta_2} (I_{\ell} - u_2 u_2^T) A_{1,2}^{(1)} v_2, \quad v_3 \equiv \frac{1}{\gamma} (I_m - v_2 v_2^T) A_{1,2}^{(1)} u_2, \]

\[ A_{1,2}^{(2)} = (I_{\ell} - u_2 u_2^T) A_{1,2}^{(1)} (I_m - v_2 v_2^T) \]

\[ = (I_{\ell} - u_1 u_1^T - u_2 u_2^T) A (I_m - v_1 v_1^T - v_2 v_2^T) \]  \tag{17}

the last equation of which shows that the left and right nullspaces of the second step matrix EMPR components are at least two dimensional and spanned by vectors two of which are \( u_1 \) and \( u_2 \) for the left nullspace, and, \( v_1 \) and \( v_2 \) for the right nullspace. In other words, the rank of the second step matrix EMPR component is less than the rank of the original matrix by two.

Equation (16) has been written for the second step of recursive EMPR. This can be generalized to \( (j + 1) \)th step by writing the following equality

\[ A_{1,2}^{(j)} = \alpha_j u_{j+1} v_{j+1}^T + \beta_j u_{j+2} v_{j+1}^T + \gamma_j u_{j+1} v_{j+2}^T + A_{1,2}^{(j+1)} \]  \tag{18}

where \( A_{1,2}^{(0)} \equiv A \) and

\[ \alpha_{j+1} \equiv \left\| u_{j+1}^T A_{1,2}^{(1)} v_{j+1} \right\|, \quad \beta_{j+1} \equiv \left\| (I_{\ell} - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} v_{j+1} \right\|, \]

\[ \gamma_{j+1} \equiv \left\| (I_m - v_{j+1} v_{j+1}^T) A_{1,2}^{(j)} u_{j+1} \right\|, \]

\[ u_{j+2} \equiv \frac{1}{\beta_{j+1}} (I_{\ell} - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} v_{j+1}, \quad v_{j+2} \equiv \frac{1}{\gamma} (I_m - v_{j+1} v_{j+1}^T) A_{1,2}^{(j)} u_{j+1}, \]

\[ A_{1,2}^{(j+1)} = (I_{\ell} - u_{j+1} u_{j+1}^T) A_{1,2}^{(j)} (I_m - v_{j+1} v_{j+1}^T) \]

\[ = \left( I_{\ell} - \sum_{k=1}^{j+1} u_k u_k^T \right) A \left( I_m - \sum_{k=1}^{j+1} v_k v_k^T \right) \]  \tag{19}

It is not hard to show that the projective matrix equations

\[ \left( I_{\ell} - \sum_{k=1}^{\ell} u_k u_k^T \right) = 0_{\ell \times \ell}, \quad \left( I_m - \sum_{k=1}^{m} v_k v_k^T \right) = 0_{m \times m}, \]  \tag{20}
where $0_{\ell \times \ell}$ and $0_{m \times m}$ denote square zero matrices whose dimensionalities are shown by their subscripts, are satisfied. These equalities imply that the matrix component $A_{1,2}^{(j+1)}$ identically vanishes when $(j + 1)$ matches $\min(\ell, m)$, by leaving no further chance of proceeding. At that point, “this recursion over EMPR” ends, by enabling us to write the following decomposition over 1-rank matrix components

$$A = \sum_{i=1}^{\ell} \alpha_i u_i v_i^T + \sum_{i=1}^{\ell-1} \beta_i u_{i+1} v_i^T + \sum_{i=1}^{\ell} \gamma_i u_i v_{i+1}^T = U \Sigma V^T, \quad \ell < m$$

(21)

where the columns of the orthonormal matrices $U$ and $V$ are respectively $u_1, \ldots, u_\ell$ and $v_1, \ldots, v_m$, and, the tridiagonal matrix $\Sigma$ has diagonal elements which are downwardly ordered as $\alpha_1, \ldots, \alpha_\ell$ while the nearest lower and upper neighbors of the main diagonal are composed of downwardly ordered $\beta_1, \ldots, \beta_{\ell-1}$ and $\gamma_1, \ldots, \gamma_\ell$ parameters respectively. This is valid for the case where $\ell < m$. If $m < \ell$ then the $\alpha$ and $\beta$ parameters are indexed between 1 and $m$ inclusive while the index region of $\gamma$ becomes between 1 and $m - 1$ inclusive. On the other hand, the case where $\ell = m$ has all parameters in the range between 1 and $\ell - 1 = m - 1$ except the existence of additional $\alpha_m$ parameter.

The matrix $\Sigma$ can be approximated by replacing all $\alpha$, $\beta$, and, $\gamma$ parameters with indices greater then a prescribed positive integer value by zero. We call such approximants “Truncation Approximants” or simply “Truncations”. The approximation quality can be controlled by choosing appropriate truncation levels.

## 4 Generators and Singular Value Decomposition

The analysis in the previous section shows that different TMEMPRs can be constructed for the same matrix $A$ by starting with different vector couples, $u$ and $v$. Hence we can call $u$ and $v$ “Left and Right Generators” or simply “Generators” since these two matrices together with the original matrix $A$ suffice to generate everything in the corresponding TMEMPR (REMPR). Perhaps the simplest choices for these generators are the vectors composed of equivalent elements. In mathematical language

$$u_i \equiv \frac{1}{\sqrt{\ell}}, \quad i = 1, 2, \ldots, \ell; \quad v_i \equiv \frac{1}{\sqrt{m}}, \quad i = 1, 2, \ldots, m$$

(22)

which are in fact the support vectors of the HDMR of $A$. This choice of generators may not give high qualities in rather small number of term including truncations. This possibility urges us to construct some other generators. However, it is better somehow to use the target matrix in the construction to get better efficiency (somehow autocoherence).
5 Concluding Remarks

We have developed a new matrix decomposition representation based on recursive utilization of EMPR. This is the tridiagonalization of the target matrix independent of its type and works well for rectangular matrices as well as square ones. Certain important findings and remarks are enumerated below.

1. Presented method works well for moderately sized matrices. As the number(s) of rows and/or columns grow, the computational accuracy increasingly influences the accumulation of error. Even if the scheme is in a rather simple algebraic structure, the number of additions and multiplications may increase enormously leaving us with the digit losses in the results when the dimensionality grows. In other words, the proposed algorithm is in fact ill-conditioned. In order to bypass this negativity, working precision in the calculations should be increased or certain numerical refinement procedures should be taken into action when the matrix dimensionality grows rapidly.

2. The evaluation of the $u$ and $v$ together with the $\alpha$, $\beta$, and $\gamma$ parameters is not iterative but recursive. The lack of iterations drastically reduces the computational complexity in comparison to the spectral and/or singular value decomposition.

3. The tridiagonal parameters, $\alpha$s, $\beta$s, and, $\gamma$s are somehow under the control of the generators. If the generators are coincidentally selected as singular vector couples then the tridiagonal representation becomes diagonal and decomposition matches the singular value decomposition. However this coincidence is almost out of possibility unless certain specifications exist in the matrix to be decomposed.

4. The presented formulation for the method assumes that the rectangular matrix under consideration is full rank, that is, its rank matches $\ell$. However, this may not be the case and the smallest dimensional null space (left null space when $\ell < m$) of the target matrix may not be empty. In that case, it is best to not have components lying in this null space in the left generator. This can be provided by taking the left generator as the image of some other vector under the matrix $AA^T$ since this square matrix annihilates the contributions of the nullspace vectors.

5. For the case where $\ell < m$ all $u$ vectors ($\ell$ number of orthonormal vectors) are required if the matrix to be decomposed is full rank. However, we do not need all $v$ vectors. First $(\ell + 1)$ number of them are sufficient for the decomposition. The remaining $m - \ell - 1$ number of $v$ vectors can be arbitrarily chosen under the limitation such that they must be mutually orthonormal and orthogonal to all of the first $\ell + 1$ vectors mentioned above.
6. This method can be considered as a generalization of the singular value decomposition method and it is related to the Lanczos tridiagonalization method or the utilization of the Krylov basis set for symmetric matrices. However, the presented method is quite far beyond these methods since it can be applied to rectangular matrices without seeking the fulfillment of any symmetry requirement.

7. We are proceeding to extend this approach to the more than two way arrays and also to folding and unfolding issues beside image processing. Some of these types of developments will also be presented in this conference.

8. Readers who are willing to read more are welcome to look at [1–12].

References


Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2014

NnmfPack: a versatile approach to an NNMF parallel library

N. Díaz-Gracia¹, A. Cocaña-Fernández¹, M. Alonso-González¹, F.J. Martínez-Zaldívar², R. Cortina¹, V.M. García-Mollá³, P. Alonso⁴, J. Ranilla¹ and A.M. Vidal³

¹ Departamento de Informática, Universidad de Oviedo, Spain
² Departamento de Comunicaciones, Universitat Politècnica de València, Spain
³ Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València
⁴ Departamento de Matemáticas, Universidad de Oviedo, Spain

emails: noeliadiazgracia@gmail.com, cocanaalberto@gmail.com, monica300876@gmail.com, fjmartin@dcom.upv.es, raquel@uniovi.es, vmgarcia@dsic.upv.es, palonso@uniovi.es, ranilla@uniovi.es, avidal@dsic.upv.es

Abstract

Non-negative Matrix Factorization (NNMF) has become an essential tool in many fields spanning machine learning, data analysis, image analysis or audio source separation, among others. Due to the computational complexity of this factorization and its consequent high execution times, computationally efficient solutions are required. We introduce NnmfPack, a parallel multi-architecture library including multiple NNMF algorithms, designed to work on both sequential and multi-core computers, as well as Graphics Processing Units (GPUs) and Intel Xeon Phi coprocessors. NnmfPack routines are written in C language using, as appropriate, OpenMP, Intel development technologies or the CUDA suite. These routines are also callable from MATLAB/Octave through MEX (MATLAB Executable) interfaces. The current version is designed for Linux operating systems. This document gives an overview of the library, installation and configuration instructions, examples of use and an initial performance evaluation on multiple parallel architectures.

Key words: NNMF, parallel library, GPU, Intel MIC, Multi-core, Many-core

©CMMSE
Page 456 of 1485
ISBN: 978-84-616-9216-3
1 Introduction

In recent years, the Non-negative Matrix Factorization (NNMF) has become an essential tool in many fields such as document clustering, data mining, machine learning, data analysis, image analysis, audio source separation, bioinformatics, etc. [1–6]. This matrix decomposition consists on approximating a matrix $A \in \mathbb{R}^{m \times n}$ with non-negative elements by means of a product of two lower rank matrices with non-negative elements, $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ with $k \leq \min(m, n)$, such that $A \approx WH$. The problem is addressed as the computation of two matrices $W_0$, $H_0$ such that

$$
\|W_0H_0 - A\|_F = \min_{W \geq 0, H \geq 0} \|WH - A\|_F
$$

Throughout this document the Frobenius norm is used. However, other norms can be used instead (see, for instance, [7], where the NNMF is also defined in terms of the Kullback-Leibler divergence). Also, many algorithms have been proposed for NNMF calculation (see [5] and [7–11]).

The relevance of this factorization for building an approximation of the matrix $A$ using a product of lower rank matrices is the dimensionality reduction accomplished that effectively works as a compression tool for many applications. Though useful, because of its complexity, NNMF is also computationally demanding, what encourages to develop efficient routines capable of reducing its high execution times.

Nowadays, computer architectures have evolved from CPUs with a reasonable number of cores (multi-core) to heterogeneous systems where CPUs are aided by hardware accelerators with a huge number of computational cores (many-core).

In this paper we present a first approach to a library for NNMF calculation that we have called NnmfPack. This library has great versatility in the sense that it is designed to work on different current parallel architectures: single-core and multi-core computers, Graphics Processing Units (GPUs) or Intel Xeon Phi coprocessors. Its routines are also callable from MATLAB/Octave through MEX interfaces, what increases NnmfPack’s usability.

This work draws on a wide variety of audio signal processing applications that are the core of the interdisciplinary project DiscoSound (www.discosound.upv.es, TEC2012-38142-C04). These include applications such as audio source decomposition, musical and vocal information restoration, music score alignment, etc. (see, for example, [3], [12] and [13]). Some of these problems have recently inspired works on the parallelization of some algorithms for calculating the NNMF (see, for instance, [14] and [15]).

The remainder of the paper is as follows. Section 2 provides a general description of the library and Section 3 presents its functional characteristics. The experimental advantages of some of the implemented algorithms are analysed on different architectures in Section 4. The article ends with a section devoted to conclusions.
2 General description of NnmfPack

Given the wide variety of problems that NnmfPack can range, it cannot conceivably be presented as a closed and finished product. Thus, this work displays the library's design principles and also a general outline of its functionality. Its incremental and progressive development will allow the necessary feedback to ensure its quality. NnmfPack is an efficient numerical routines library conceived for shared memory heterogeneous parallel systems and it supports, from its conception, both conventional multi-core processors and coprocessors such as Intel MIC (Many Integrated Core) architecture and CUDA compatible [16] Graphics Processing Units (GPUs).

From a software viewpoint, NnmfPack routines are written in C language, using OpenMP [17], Intel development technologies or CUDA suite, as appropriate. Its current version is thought for Linux-compatible operating systems. Support for other operating systems or computing paradigms, such as distributed memory, may be added in future versions.

Even though NnmfPack is aimed at heterogeneous parallel environments, it is not an heterogeneous-coprocessor library, meaning that its routines have been designed to support several architectures but not to solve one single problem using simultaneously cores from different coprocessor architectures.

It should be noted that, currently, the simultaneous use of coprocessors and accelerators would entail information transfer operations, which would have a negative impact on execution times. Moreover, it would require estimating the optimal workload distribution prior to execution, what implies further overload, potentially curtailing NnmfPack efficiency.

Another characteristic feature of NnmfPack is the natural integration with MATLAB/Octave via MEX interfaces, providing flexibility and high user-friendliness. NnmfPack also includes enough examples of use.

Finally, all information regarding NnmfPack is publicly accessible via web [18]. The website is organized as follows. The main page displays a brief description of the library, contact address, acknowledgements and a menu. The documentation generated with Doxygen [19] and the download and licensing section can both be accessed from the menu.

3 Using NnmfPack

NnmfPack installation follows the standard GNU procedure: configuration, compilation and installation. The first step configures the library for the target system hardware and software and checks NnmfPack requirements. Default configuration only includes CPU routines and requires BLAS [21] and LAPACK [22] system libraries. When NnmfPack is configured for GPUs the CUDA compiler (nvcc), cuBLAS [23] and MAGMA [24] libraries are required. When configured for the Intel MIC architecture, the Intel C compiler (icc)
and the MKL [25] library are selected. The second step builds the source code into static and dynamic libraries. The last one installs the static and dynamic libraries, headers, script files, interface files for MATLAB/Octave, etc. into the destination folder.

Once the library is installed, it can be used in two different ways. On the one hand, users can implement applications calling NnmfPack routines, as shown in Algorithm 1.

**Algorithm 1** Using NnmfPack routines in C programs

```c
#include <nnmfpack.h>

int main(int argc, char * argv[])
{
    . . .
    res=dflsa(n, m, k, A, W, H, it, ti);
    if (res != 0)
        . . .
    else
        . . .
    . . .
}
```

On the other hand, it can also be directly called from MATLAB/Octave using the MEX interface files supplied, as shown in Algorithm 2. Examples of use for every architecture can be found in the installation folder.

**Algorithm 2** Using NnmfPack routines in MATLAB/Octave

```matlab
[Wo,Ho]=nnmf(A, k, ... , 'alg' , 'mult');
[Wo, Hc, ic, tc]=mex_dflsa(A, W0, H0, it, ti);
```

4 Experimental results

The results shown in this section were obtained using an ASUS ESC4000 G2 server with the following setup:

- Two Intel Xeon E5-2650 CPUs @ 2.0 GHz with 64 GB of RAM. Hyper-Threading
technology was disabled so the total number of computing cores are 16. The peak performance for this configuration in double precision is 256 GFLOPS.

• One NVIDIA Tesla K20c GPU with 2496 cores @ 706 MHz, 5 GB of DDR5 RAM and a peak performance of 1.17 TFLOPS in double precision.

• One Intel Xeon Phi 5110P coprocessor (Intel Many Integrated Core (MIC) architecture) with 60 cores @ 1.053 GHz, 8 GB RAM and a peak performance of 1.01 TFLOPS in double precision. Each core can execute up to 4 hardware threads concurrently (4-way multithreading).

Intel’s and NVIDIA’s approaches are markedly different. The Intel Xeon Phi runs its own operating system, what conceptually makes it a functionally complete computer. This should be taken into account when running the tests because some of its cores may be being used by the operating system. The manufacturer’s recommendation is to discard (do not use) one of the cores.

Regarding NnMFPack’s setup, the following configuration was used, as is arguably the best performing one:

• CPU and Xeon Phi codes were compiled with the Intel C compiler, version 14.0.2.144, and Intel MKL, version 11.1.2.

• For the Tesla GPU, a NVIDIA NVCC compiler, version 6.0.37, and MAGMA, version 1.4.1 (that also uses Intel MKL due to its installation process), have been used.

• As for MATLAB/Octave, the versions used were 3.4.3 and 8.1.0.604 (R2013a), respectively. MATLAB includes Intel MKL, version 10.3.11, so that for some functionalities it uses all system processor cores (parallel execution).

Among the algorithms currently available in NnMFPack, the overall computational cost of the Alternating Least Squares Algorithm (ALSA), an algorithm based on the alternating application of least squares methods [15], is greater than the Modified Lee and Seung Algorithm (MLSA), an efficient variant of the Lee and Seung algorithm [7] presented in [15], for a given error bound. MLSA is a multiplicative uniform cost algorithm with approximately

\[\frac{4k}{5}mn + 2k(m + n)(2k + 1)\] (2)

flops per iteration, where \(n\), \(m\) and \(k\) are the dimensions of the input matrix \(A \in \mathbb{R}^{m \times n}\) and the two lower rank matrices \(W \in \mathbb{R}^{m \times k}\) and \(H \in \mathbb{R}^{k \times n}\), with \(k \leq \min(m, n)\). On the other hand, ALSA must apply the nonnegative least squares (NNLS) algorithm \((m + n)\) times per iteration, thus can be bounded to

\[(2k^2 + mn)(m + n) + (m + n)F_{NNLS}\] (3)
where $F_{NNLS}$ denotes NNLS flops.

Since the aim of this work is to analyse the NnmfPack behaviour, and not to compare the efficiency of the algorithms, we have focused the experiments on the MLSA algorithm.

The analysis performed in this section has been carried out using as the input matrix $A \in \mathbb{R}^{m \times n}$ uniformly random generated positive square matrices ($m = n$). For a more comprehensive study, for each $m$ (or $n$) value $k$ values ranges from 10% to 100% of $m$ (or $n$) with step 10%. It should be reminded that $k$ is the other dimension for the lower rank matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$.

Figure 1 shows the performance of the MLSA algorithm of NnmfPack and Table 1 presents the theoretical percentage achieved by empirical results over the different architectures for some values of $m$, $n$, and $k$. Readers can access to all numerical results using the URL structpack.uniovi.es/index.php/description.

<table>
<thead>
<tr>
<th>$n = m$</th>
<th>$k$</th>
<th>CPUx1</th>
<th>CPUx16</th>
<th>Xeon Phi</th>
<th>K20c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>500</td>
<td>89.1%</td>
<td>61.5%</td>
<td>37.6%</td>
<td>62.3%</td>
</tr>
<tr>
<td>2500</td>
<td>1250</td>
<td>92.2%</td>
<td>77.1%</td>
<td>61.3%</td>
<td>80.0%</td>
</tr>
<tr>
<td>5000</td>
<td>2500</td>
<td>94.1%</td>
<td>84.4%</td>
<td>68.6%</td>
<td>84.0%</td>
</tr>
<tr>
<td>7500</td>
<td>3750</td>
<td>94.6%</td>
<td>86.6%</td>
<td>70.6%</td>
<td>85.7%</td>
</tr>
<tr>
<td>10000</td>
<td>5000</td>
<td>95.8%</td>
<td>88.8%</td>
<td>75.3%</td>
<td>86.2%</td>
</tr>
</tbody>
</table>

Table 1: Theoretical performance percentage achieved by empirical results.
Figure 1 and Table 1 show that when using the 16 cores available in the testing machine (CPUx16 in Figure 1) the CPU implementation achieves more than the 92% of its peak performance (i.e. with $m = n = k = 10000$), while the sequential algorithm (CPUx1 legend in Figure 1) shows the lowest performance, although it achieves its maximum theoretical computing capacity.

Regarding to the Xeon Phi’s behaviour, it is very closed to the 80% of its peak performance, reaching more than 800 GFLOPS (i.e. with $m = n = k = 10000$). In a similar vein, Tesla K20c reaches up to 1030 GFLOPS (i.e. for $m = n = 10000$ and $k = 8000$), more than the 88% of its peak performance.

Given that the matrix multiplication testbeds provided by CUDA suite and Intel development technologies reach, in ideal conditions (optimal sizes, native execution, etc.), 82% and 90% of the Xeon Phi and K20c’s maximum theoretical performances, respectively, MLSA’s results in these architectures should be considered excellent. This matches the observed behaviour for CPU multi-core executions as well, where the same testbeds are over 96%.

<table>
<thead>
<tr>
<th>$n = m$</th>
<th>$k$</th>
<th>CPUx1</th>
<th>CPUx16</th>
<th>Xeon Phi</th>
<th>K20c</th>
<th>Matlab</th>
<th>Octave</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>1500</td>
<td>16.0</td>
<td>1.2</td>
<td>0.4</td>
<td>0.3</td>
<td>2.5</td>
<td>57.7</td>
</tr>
<tr>
<td>7500</td>
<td>2250</td>
<td>54.1</td>
<td>3.9</td>
<td>1.1</td>
<td>0.8</td>
<td>7.1</td>
<td>193.9</td>
</tr>
<tr>
<td>10000</td>
<td>3000</td>
<td>125.4</td>
<td>8.6</td>
<td>2.7</td>
<td>1.9</td>
<td>15.2</td>
<td>458.3</td>
</tr>
</tbody>
</table>

Table 2: MATLAB/Octave and NnmfPack execution time for MLSA algorithm (seconds).

It is noteworthy that, when establishing a comparison between MATLAB/Octave functions and the MLSA algorithm included in NnmfPack, an initial adjustment of the input parameters is needed, so equity conditions are guaranteed. Table 2 shows the results obtained for three test examples. As can be seen, Octave times are extremely high and MATLAB, even using the intrinsic MKL parallelism, is outperformed by NnmfPack’s CPU implementation (CPUx16).

5 Conclusions

We have presented the first approach to a numerical library that provides algorithms for the non-negative matrix factorization. Despite being an early version, the features available make it an attractive alternative for the NNMF resolution in current multi-core and many-core architectures, providing some interesting performance figures from a computational point of view.

It is also worth noting that this work will lead to future versions extending current features such as the addition of new metrics, more efficient algorithms and the treatment of sparse or structured matrices.
Acknowledgements

This work has been partially supported by “Ministerio de Economía y Competitividad” from Spain, under the projects TEC2012-38142-C04-01 and TEC2012-38142-C04-04.

References


High-Order compact scheme for pricing variance swaps

Mehzabeen J. Dilloo\textsuperscript{1} and Désiré Y. Tangman\textsuperscript{1}

\textsuperscript{1} Department of Mathematics, University of Mauritius
emails: mehzabeen.dilloo1@umail.uom.ac.mu, y.tangman@uom.ac.mu

Abstract

In this paper, we present a high-order compact scheme to price variance swaps under both the one-dimensional Black–Scholes model and the two-dimensional Hestons’s stochastic volatility model. We consider the framework architected by Little and Pant (2001), which involves solving a set of partial differential equations with terminal conditions that depend on the definition of the realised variance. To improve the efficiency of the high-order compact schemes, we coupled them with an implicit-Richardson extrapolation time stepping scheme. Numerical results confirm the fourth-order convergence of our schemes and its superiority compared to the second-order Crank–Nicolson scheme used previously.

Key words: Variance swaps, stochastic volatility, high-order compact, partial differential equations, Richardson extrapolation.

MSC 2000: AMS codes (65B05, 65N06)

1 Introduction

Volatility is a measure for the variations of market prices of an underlying asset over time. Usually to hedge the risks exposure in the directional movements of an equity position in a portfolio, an investor will use an option. But in order to hedge against the magnitude of these movements, then the investor should trade in volatility derivatives. Trading volatility risk has become increasingly important making contracts which provide direct exposure to the volatility more popular. One such contract is a variance swap; it is an over-the-counter financial derivative that permits investors and dealers to hedge and to speculate in the volatility itself. Being forward contracts, no money exchange hands at its initiation. Influential analytical methods to price such derivatives were proposed by [3] and [4]. However, their assumption that the sampling time of a variance swap is continuous is not practical
since almost all variance swap instruments are defined over a set of discrete points in time. So this calls for a more versatile numerical approach that can be adapted to any discretely sampled definition of the realised variance.

One such approach was proposed in [9] where a partial differential equation (PDE) framework was set up to price variance swaps under the Black–Scholes model [1]. They used finite differences to solve a system of PDEs and attain a second-order convergence. However, in financial literature, high-order compact (HOC) schemes (three-points stencil) have been used several times [8, 12, 13, 5] because such schemes defined on the same three-points stencil as for the famous second-order Crank–Nicolson scheme, give more accurate results for the same computational complexity. In this paper, we use the same framework developed in [9] but consider the HOC scheme described in [12] for one-dimensional problems and that derived in [5] for two-dimensional problems. In both cases, the principle lies in differentiating the original PDEs to replace the higher order partial derivatives in the truncation error after applying second-order finite difference approximations.

Considering the temporal discretisation, if a Crank–Nicolson time stepping scheme was to be implemented, we would need the time step to be proportional to the square of the spatial step. This is quite a restrictive condition to attain a general fourth-order convergence and this eventually increases the computational cost of the schemes. So, to circumvent this problem, we choose to implement an implicit-Richardson extrapolation [6] which helps to remove the temporal errors efficiently and rapidly.

In this paper, we first consider pricing of variance swap under the Black–Scholes model. However, it is known that the volatility smile observed on the market plagues the model’s constant volatility assumption. So, to be consistent with the market, we also choose to price variance under the Heston’s stochastic volatility model [7]. Analytical solutions to variance swaps under this model have been derived in [2, 11, 10] using different approaches. In [2], they integrate the SDEs directly and in [11], they use a PDE approach but apply a Fourier inverse transform to solve the system. In [10], the tower property was exploited when finding the risk neutral expectation of the terminal condition. The same expression as in [11] was obtained but with some restrictions. In all these cases, procedures to find the solution have to be done again when the definition of the realised variance changes. This makes the numerical solution of PDEs approach a more versatile one as altering the definition of the realised variance implies only a change in the terminal conditions of the PDE. Therefore, we proposed to price variance swap in the framework designed by [9] using the HOC scheme proposed by [5].

An outline of this paper is as follows: In section 2, we describe the pricing problem under the Black–Scholes and the Heston’s stochastic volatility models. The HOC schemes for both the one-dimensional and two-dimensional schemes are given in section 3 and section 4 respectively. In section 5, we describe how we discretise the temporal derivative. We give our numerical experiments showing the computational efficacy of the new approach in
section 6 and section 7 contains our conclusion.

2 Variance swaps

A variance swap with maturity $T$ has payoff $L(\sigma^2_R - K)$, where $L$ is the notional amount of the swap per annualised volatility point squared and $\sigma^2_R$ is the realised variance of the asset $S_t$ as defined in the contract. The realised variance $\sigma^2_R$ can be given many definitions—the most popular one and used in this paper being

$$\sigma^2_R = \frac{AF}{N} \sum_{i=0}^{N-1} \left( \frac{S_{i+1} - S_i}{S_i} \right)^2 100^2,$$

where $AF$ is the annualisation factor, $N$ is the number of observations during the lifetime $T$ of the variance swap, $S_i = S_{t_i}$, $t_i = i\Delta t$ and $\Delta t = 1/AF = T/N$. Therefore, the pricing problem resides in finding the fair value of the strike $K$ such that no money exchange hands at initiation of the contract. Given that $r$ is the risk free interest rate, this implies that

$$e^{-rT} L \{ \mathbb{E}[\sigma^2_R] - K \} = 0,$$

and this sets the value of $K$ as

$$K = \mathbb{E}[\sigma^2_R],$$

$$= \frac{AF}{N} \sum_{i=0}^{N-1} \mathbb{E} \left[ \left( \frac{S_{i+1} - S_i}{S_i} \right)^2 \right] 100^2,$$

where $\mathbb{E}[\cdot]$ is the expectation under the risk neutral measure.

So, the value of $K$ depends on the dynamics followed by the asset $S_t$. In this paper, we investigate the case where the asset follows the celebrated Black–Scholes model and the Heston’s stochastic volatility model.

2.1 Pricing variance swap under the Black–Scholes model

The Black–Scholes model assumes that the underlying asset $S_t$ follows the SDE

$$\frac{dS_t}{S_t} = (r - q) dt + \sigma dW_t,$$

where $r$ is the interest rate, $\sigma$ the volatility, $q$ is the dividend and the stochastic driving force $W_t$ is a Wiener process.

From (1), we can see that the problem is to find $N$ expectations of the form

$$\mathbb{E} \left[ \left( \frac{S_{i+1} - S_i}{S_i} \right)^2 \right],$$

(2)
for some fixed period \( \Delta t \) and \( N \) different tenors \( \tilde{t} = i\Delta t, \ i = 1, \cdots, N \). Following [9], we break the problem of evaluating (1) into two parts; one for tenors \( \tilde{t} = i\Delta t, \ i = 2, \cdots, N \) and one for \( \tilde{t} = \Delta t \).

2.1.1 For case \( \tilde{t} = i\Delta t \), where \( i = 2, \cdots, N \)

In this case, we introduce another variable [9, 11]

\[
I_t = \int_0^t \delta(\tilde{t} - \Delta t - \tau) S_\tau \, d\tau,
\]

since the expectation (2) depends on two unknowns \( S_{\tilde{t}} \) and \( S_{\tilde{t} - \Delta t} \). The Dirac delta function \( \delta(\tilde{t} - \Delta t - \tau) \) makes \( I_t \) jumps at \( \tilde{t} - \Delta t \) so that

\[
I_t = \begin{cases} 
0, & 0 \leq t < \tilde{t} - \Delta t, \\
S_{\tilde{t} - \Delta t}, & t \geq \tilde{t} - \Delta t.
\end{cases}
\]

Now, using Ito’s lemma and dynamic hedging, we can show that the value of a contingent claim \( P \) depending on \( t, S_t \) and \( I_t \) satisfies

\[
\frac{\partial P}{\partial t} + \delta(\tilde{t} - \Delta t - t) \frac{\partial P}{\partial I} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + (r - q) S \frac{\partial P}{\partial S} - rP = 0.
\]

Solving the above PDE with the terminal condition

\[
P(\tilde{t}; S, I) = \left( \frac{S_{\tilde{t}}}{I_{\tilde{t}}} - 1 \right)^2,
\]

(since \( I_t = S_{\tilde{t} - \Delta t} \) when \( t \geq \tilde{t} - \Delta t \)) and using the Feynman–Kac theorem, the solution to (2) can be expressed as

\[
\mathbb{E}\left[ \left( \frac{S_{\tilde{t}} - S_{\til{t} - \Delta t}}{S_{\til{t} - \Delta t}} \right)^2 \right] = e^{r\tilde{t}} P(0; S_0, I_0).
\]

From the definition of the Dirac delta function in PDE (4), we notice that away from \( \tilde{t} - \Delta t \), the PDE is reduced to the Black–Scholes PDE

\[
\frac{\partial P}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + (r - q) S \frac{\partial P}{\partial S} - rP = 0.
\]

As in [9], we apply the transformations \( x = \ln S \) and \( z = \ln I \) so that we now deal with a constant coefficient PDE which is easier to manipulate. Hence, we need \( P(0; x_0, z_0) \) from

\[
\frac{\partial P}{\partial t} + \frac{1}{2} \sigma^2 \frac{\partial^2 P}{\partial x^2} + \left( r - q - \frac{1}{2} \sigma^2 \right) \frac{\partial P}{\partial x} - rP = 0,
\]

(5)
with terminal condition

$$P(\tilde{t}; x, z) = (e^{x-z} - 1)^2.$$  \hspace{1cm} (6)

More importantly, we notice that our two-dimensional PDE (4) have been transformed into a one-dimensional PDE (5). Therefore, we exploit this particular property which considerably improves the computational efficiency of our problem. However, the presence of the variable \( z \) in the terminal conditions (6) still defines the problem over a two-dimensional space. Fortunately, the absence of any derivative in \( z \) in the PDE (5) indicates that \( z_t \) is constant away from \( t = \tilde{t} - \Delta t \).

Bearing in mind that values should be continuous in the absence of arbitrage, our claim \( P(t; x_t, z_t) \) should satisfy the jump condition

$$\lim_{t \uparrow \tilde{t} - \Delta t} P(t; x_t, z_t) = \lim_{t \downarrow \tilde{t} - \Delta t} P(t; x_t, z_t).$$  \hspace{1cm} (7)

Hence, our solution \( P(0; x_0, z_0) \) can be obtained by solving the Black–Scholes PDE (5) subject to terminal condition (6) and jump condition (7).

Before any further descriptions, consider the interval \([x_{\min}, x_{\max}]\) and build the computational grid

$$\Omega_{\Delta x} = \left\{ x_j \in \mathbb{R}^+ : x_j = x_{\min} + (j - 1)\Delta x, j = 1, \cdots, j_m + 1, \Delta x = \frac{x_{\max} - x_{\min}}{j_m} \right\}.$$  \hspace{1cm} (8)

From our jump condition (7), it implies that

$$P(\tilde{t} - \Delta t^--; x^-, z^-) = P(\tilde{t} - \Delta t^++; x^+, z^+),$$

and because \( z^- = z_0 = \ln 0 = -\infty (I_t = 0 \text{ for } t < \tilde{t} - \Delta t) \) and \( z^+ = x^+ (I_t = S_{\tilde{t} - \Delta t} \text{ for } t \geq \tilde{t} - \Delta t) \), we have

$$P(\tilde{t} - \Delta t^-; x^-, -\infty) = P(\tilde{t} - \Delta t^++; x^+, x^+),$$

and on the grid structure (8), we set

$$P(\tilde{t} - \Delta t^-; x_j^-, -\infty) = P(\tilde{t} - \Delta t^++; x_j^+, x_j^+).$$  \hspace{1cm} (9)

So, due to the jump condition at \( t = \tilde{t} - \Delta t \), for each \( \tilde{t} = i\Delta t \) \( (i = 2, \cdots, N) \), we need to go through two stages to obtain our solution. Firstly, we need to solve for \( P(\tilde{t} - \Delta t^++; x_j^+, x_j^+) \) for each \( j \) when \( P_j(\tilde{t}; x) = P(\tilde{t}; x, x_j) \) satisfies the PDE (5) and terminal condition (6) with \( z = x_j \), when \( t \in (\tilde{t} - \Delta t, \tilde{t}] \). For this, we extend the domain such that for each \( j \), \( x_j \in [x_{\minext}, x_{\maxext}] \), and a new computational domain is created with the same spatial step \( \Delta x \) so that

$$\Omega_{\text{ext}} = \left\{ x_{\text{jext}} \in \mathbb{R}^+ : x_{\text{jext}} = x_{\text{minext}} + (j_{\text{ext}} - 1)\Delta x, j_{\text{ext}} = 1, \cdots, j_{\text{maxext}} + 1 \right\}.$$
where \( j_{\text{next}} = (x_{\text{minext}} - x_{\text{maxext}}) / \Delta x \). Therefore, for each \( x_j \), we solve the PDE (5) on \( \Omega_{\text{ext}} \) and interpolate at \( x_j \) to obtain \( P(\tilde{t} - \Delta t^+; x_j^+, x_j^+) \). In so doing, from (9), we also have the values \( P(\tilde{t} - \Delta t^-; x_j^-, -\infty) \) for each \( j \). In turn, these values will act as the terminal condition for our second stage which involves solving the same PDE (5) on \( \Omega_{\Delta x} \) for \( P(0; x_0) \equiv P(0; x_0, z_0) \) subject to

\[
P(\tilde{t} - \Delta t; x) = P(\tilde{t} - \Delta t^-; x^-, -\infty),
\]

when \( t \in [0, \tilde{t} - \Delta t] \).

### 2.1.2 For case \( \tilde{t} = \Delta t \) where \( i = 1 \)

In this case \( S_{\tilde{t} - \Delta t} = S_0 \) is known and to find the expectation which reads as

\[
\frac{1}{S_0^2} E \left[ (S_{\tilde{t}} - S_0)^2 \right],
\]

we only need to solve the constant coefficient Black–Scholes PDE (5) subject to the terminal condition

\[
P(\Delta t; x) = (e^{x-x_0} - 1)^2.
\]

Hence, for \( \tilde{t} = \Delta t \) (\( i = 1 \)), the expectation (2) is given as

\[
E \left[ \frac{(S_{\Delta t} - S_0)^2}{S_0^2} \right] = e^{r \Delta t} P(0; x_0).
\]

The analytical solution for pricing variance swap can be obtained as in [9] in a few steps. First, we can show that (1) simplifies to

\[
E[\sigma_R^2] = AF \left( E \left[ \left( \frac{S_{\tilde{t}}}{S_{\tilde{t} - \Delta t}} \right)^2 - 2 \left( \frac{S_{\tilde{t}}}{S_{\tilde{t} - \Delta t}} \right) + 1 \right] \right) 100^2,
\]

and using Ito’s lemma, we have the following expectations

\[
E \left[ \left( \frac{S_{\tilde{t}}}{S_{\tilde{t} - \Delta t}} \right)^2 \right] = e^{2(r-q)\Delta t + \sigma^2 \Delta t},
\]

\[
E \left[ \left( \frac{S_{\tilde{t}}}{S_{\tilde{t} - \Delta t}} \right) \right] = e^{2(r-q)\Delta t},
\]

so that the closed-form solution postulates as

\[
E[\sigma_R^2] = AF \left[ e^{2(r-q)\Delta t + \sigma^2 \Delta t} - 2e^{2(r-q)\Delta t} + 1 \right] 100^2.
\]

We remark that the analytical solution depends on the time to maturity \( T \) of the variance swap and on the number of observations \( N \) during its lifetime through the annualisation factor \( AF \) which represents the number of observations in one year.
2.2 Heston’s stochastic volatility model

Now, we consider the dynamics of the Heston’s stochastic volatility model given as

\[
\frac{dS_t}{S_t} = \mu dt + \sqrt{v_t} dW^S_t, \quad \text{and} \quad dv_t = \kappa (\theta - v_t) dt + \xi_v \sqrt{v_t} dW^v_t,
\]

where we think of \(v_t\) as the variance of the asset, of \(\theta\) as the long run mean, of \(\kappa\) as the mean reversion rate, and of \(\xi_v\) as the volatility of the variance. We also denote the correlation between the two Brownian motions as \(\rho\), that is, \(<dW^S_t, dW^v_t> = \rho dt\).

Pricing variance swap under the Heston’s model, we need to find the strike \(K\) from (1). Similar procedures as under the Black–Scholes model are to be followed. Firstly for the tenors \(\tilde{t} = i \Delta t, i = 2, \cdots, N\), we introduce the variable \(I_t\) as defined in (3) and consider the problem away from \(\tilde{t} - \Delta t\). Given that our claim is denoted by \(P(t; S, v)\) and \(\lambda(S,v,t)\) is the market risk parameter, we have the PDE

\[
\frac{\partial P}{\partial t} + \frac{1}{2} S^2 \frac{\partial^2 P}{\partial S^2} + rS \frac{\partial P}{\partial S} - rV + \frac{1}{2} \xi_v^2 \frac{\partial^2 P}{\partial v^2} + [\kappa(\theta - v) - \lambda(S,v,t)] \frac{\partial P}{\partial v} + \rho \xi_v S_v \frac{\partial^2 P}{\partial S \partial v} = 0.
\]

After applying the transformations \(x = \ln S, z = \ln I, P = e^{(T-t)\tilde{P}}\) and \(y = v/\xi_v\). These transformations prove to be essential when deriving our HOC scheme as they make both the diffusion coefficients identical and remove the reaction term in the PDE so that less terms are involved [5]. As such we have

\[
\frac{\partial P}{\partial t} + \frac{1}{2} y \xi_v \left( \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} \right) + \rho \xi_v y \frac{\partial^2 P}{\partial x \partial y} - \left( \frac{1}{2} \xi_v y - r \right) \frac{\partial P}{\partial x} + \kappa \left( \frac{\theta - \xi_v y}{\xi_v} \right) \frac{\partial P}{\partial y} = 0. \quad (10)
\]

After applying the jump conditions (7) in a similar manner as for the Black–Scholes model and considering the domain \(\Omega_{\text{ext}}\), we find an approximation for \(P_j(\tilde{t} - \Delta t; x, y)\) using our first stage problem which involves solving the PDE (10) subject to the terminal condition (6) with \(z = x_j\), for \(t \in (\tilde{t} - \Delta t, \tilde{t}]\). Hence, we are able to find \(P(\tilde{t} - \Delta t^+; x_j^+, y, x_j^+)\) and consequently, \(P(\tilde{t} - \Delta t^-; x_j^-, y, -\infty)\) from the jump condition. Having obtained these values for each \(j\), we move on to the second stage where we solve the same PDE (10) while satisfying the terminal condition

\[
P(\tilde{t} - \Delta t^-; x, y) = P(\tilde{t} - \Delta t^-; x^-, y, -\infty),
\]

for \(t \in [0, \tilde{t} - \Delta t]\).

Again, for \(\tilde{t} = \Delta t\), the problem simplifies to solving the PDE (10) for \(P(0; x_0, y_0)\) with terminal condition

\[
P(\Delta t; x, y) = (e^{x - x_0} - 1)^2.
\]
Now, using the values \( P(0; x_0, y_0, z_0) \equiv P(0; x_0, y_0) \) and Feynman–Kac theorem, the expectations for each \( i \) can be obtained as shown below

\[
E \left[ \left( \frac{S_t - S_{t-\Delta t}}{S_{t-\Delta t}} \right)^2 \right] = e^{-\tilde{r} t} P(0; S_0, v_0, I_0),
\]

\[
= e^{-\tilde{r} t} (e^{-r \tilde{t}} P(0; x_0, y_0, z_0)).
\]

Our approximate solution is compared to the closed-form solution given in [11].

3 High-order compact scheme for the Black–Scholes problem

For any claim \( P \), the Black–Scholes PDE (5) can be represented as

\[
P_t + a P_{xx} + b P_x + c P = 0,
\]

where \( a = \sigma^2/2, b = r - q - \sigma^2/2 \) and \( c = -r \). Considering the implementation of the usual second-order approximations

\[
\delta_x P_j = \frac{-P_{j-1} + P_{j+1}}{2\Delta x}, \quad \text{and} \quad \delta_x^2 P_j = \frac{P_{j-1} - 2P_j + P_{j+1}}{\Delta x^2},
\]

the semi-discretised system at \( x = x_j \) can be given as

\[
\frac{\partial P}{\partial t} + a \delta_x^2 P + b \delta_x P + c P - \frac{1}{12} \Delta x^2 (2bP_{xxx} + aP_{xxxx}) + O(\Delta x^4) = 0,
\]

after dropping the subscript on \( j \). Solving problems with the above discretised system will give only a second-order accurate solution due to the leading truncation error term being of order \( O(\Delta x^2) \). The idea of obtaining a high-order compact scheme from the above system is to replace the higher order partial derivatives in the truncation error term with lower order ones. This is because approximations for higher order partial derivatives do not exist on compact (three-points) stencils.

Hence, to obtain expressions for \( P_{xxx} \) and \( P_{xxxx} \) in terms of lower order partial derivatives \( P_x \) and \( P_{xx} \), we differentiate the original PDE (11). As such, we have

\[
P_{xxx} = -\frac{cP_x + bP_{xx} + P_{x,t}}{a}, \quad \text{and} \quad P_{xxxx} = -\frac{cP_{xx} + bP_{xxx} + P_{xx,t}}{a}.
\]

Replacing the above in the truncation error term and multiplying the resulting system throughout by \( a \), we have a fourth-order compact semi-discretised system. Further, given that the problem is defined for \( t \in [t_{\min}, t_{\max}] \), we apply a time reverse transformation \( \tau = t_{\max} - t_{\min} - t \), so that the system reads as

\[
\frac{\partial}{\partial \tau} (AP) = BP,
\]

\[ \text{(12)} \]
where

\[ A = a + \frac{\Delta x^2}{12} b \delta_x + \frac{\Delta x^2}{12} a \delta^2_x, \quad \text{and} \quad B = ab \delta_x + \left( a^2 + \frac{\Delta x^2}{12} b^2 \right) \delta^2_x + c A. \]

4 High-order compact scheme for the Heston’s problem

In the case of the problem defined under the Heston’s model, we point out that the two-dimensionality of the PDE requires some special conditions to be satisfied. First, we must have identical diffusion coefficients, which was made possible in [5] with the transformations discussed above. Secondly, considering the computational domain \( \Omega_{\Delta x} \) in (8) and \( \Omega_{\Delta y} \) defined as

\[ \Omega_{\Delta y} = \left\{ y_k \in \mathbb{R}^+ : y_k = y_{\min} + (k - 1) \Delta y, \ k = 1, \ldots, k_m + 1, \Delta y = \frac{y_{\max} - y_{\min}}{k_m} \right\}, \]

we must impose the condition that \( \Delta y = \Delta x \) so that terms can be grouped together and a scheme of order \( O(\Delta x^4) \) can be derived.

Similar procedures as in [5], which reflects the same methodology as discussed for the Black–Scholes problem, are to be followed. Firstly, we apply the usual second-order finite difference approximations

\[
\begin{align*}
\delta_x P_{j,k} &= \frac{-P_{j-1,k} + P_{j+1,k}}{2\Delta x}, && \delta^2_x P_{j,k} = \frac{P_{j-1,k} - 2P_{j,k} + P_{j+1,k}}{\Delta x^2}, \\
\delta_y P_{j,k} &= \frac{-P_{j,k-1} + P_{j,k+1}}{2\Delta y}, && \delta^2_y P_{j,k} = \frac{P_{j,k-1} - 2P_{j,k} + P_{j,k+1}}{\Delta y^2}, \\
\delta_x \delta_y P_{j,k} &= \frac{P_{j+1,k+1} - P_{j-1,k+1} - (P_{j+1,k-1} - P_{j-1,k-1})}{4\Delta x \Delta y},
\end{align*}
\]

to the PDE (10). The leading truncation error terms resulting from the above discretisations are

\[
\frac{1}{24 \xi_v} \Delta x^2 [4r(\theta - \xi_v y) P_{yyy} + \xi_v [y \xi_v (P_{gyyy} + P_{xxxx}) + 4\rho \xi_v y (P_{xyy} + P_{xxxy}) + (4r - 2y \xi_v) P_{xxx}]].
\]

So, following the same principle, we need to find the higher order partial derivatives in terms of lower order partial derivatives by differentiating the PDE (10) as

\[
\begin{align*}
P_{xxx} &= -P_{xyy} - 2\rho P_{xx} - 2r + \xi_{v} y P_{xx} + \frac{2r(\xi_{v} y - \theta)}{\xi_{v} y} P_{xy} + \frac{2}{\xi_{v} y} P_{x,t}, \\
P_{yyy} &= -P_{xx} - 2\rho P_{xyy} - \frac{1}{y} P_{xx} - \frac{2r(\theta - \xi_{v} y) + \xi_{v}^2 }{\xi_{v}^2 y} P_{yy} - \frac{2\rho \xi_{v} + 2r - \xi_{v} y}{\xi_{v} y} P_{xy} + \frac{1}{y} P_{x}, \\
\quad + \frac{2\xi_{v} y}{\xi_{v} y} P_{y} + \frac{2}{\xi_{v} y} P_{y,t}. 
\end{align*}
\]
Further differentiating (13) and (14) with respect to $y$ and $x$ respectively, and adding the resulting expressions gives

$$P_{xyyy} + P_{xxx} = \frac{1}{y^2 \xi_v^2} \left[ 2 \theta \kappa P_{xy} + 2 \kappa y (\xi_v y - \theta) P_{xyy} \right] + \frac{1}{y^2 \xi_v} \left[ 2 r P_{xx} + y (\xi_v y - 2 r) P_{xxy} - 2 (y^2 \rho \xi_v P_{xxyy} + P_{x,t} - y P_{xy,t}) \right].$$

Similarly, we differentiate the PDE twice with respect to $x$ and $y$ and we add up the two equations to get

$$P_{xxxx} + P_{yyyy} = -2 \rho (P_{xyyy} + U_{xxy}) - 2 P_{xyy} + \frac{2 (\kappa \xi_v y - \xi_v^2 - \kappa \theta)}{\xi_v^2 y} U_{xyy} - \frac{(2 r - \xi_v y) U_{xxx}}{\xi_v y} + \frac{2 (\kappa \xi_v y - \xi_v^2 - \kappa \theta)}{\xi_v^2 y} U_{yy} - \frac{(\xi_v y + 4 \rho \xi_v + 2 r)}{\xi_v y} P_{xyy} + \frac{2 \kappa}{\xi_v y} P_{yy} + \frac{2}{\xi_v y} (P_{xx,t} + P_{yy,t}).$$

So, replacing the above in the truncation error term, multiplying throughout by $y$ to avoid division by zero, and applying the time reversed transformation $\tau = t_{\text{max}} - t_{\text{min}} - t$ leads to a semi-discretised PDE of the form (12) at the point $(x_j, y_k)$ where this time

$$A = \frac{\Delta x^2}{6} \rho y \delta_x \delta_y \frac{y g}{\xi_v^2} + \frac{\Delta x^2}{12} 2 \rho \xi_v + f \delta_x + \frac{\Delta x^2}{12} y \delta^2_x + \frac{\Delta y^2}{12} \delta^2_y,$$

$$B = \frac{1}{24} \Delta x^2 \left( f^2 - 4 \rho \xi_v r - 2 \kappa g - 2 \xi_v^2 \right) + \frac{1}{12} \left( \xi_v + \frac{4 \rho \xi_v + 2 r}{\xi_v^2} \right) \delta^2_x \delta_y + \frac{1}{12} \left( 2 \kappa - \kappa \theta y - \xi_v^2 y - \xi_v^4 \right) + \frac{6 \xi_v^4 y^2}{\xi_v^2} \delta^2_y,$$

where $f = \xi_v y - 2 r$ and $g = \xi_v y - \theta$.

### 5 Implicit-Richardson extrapolation

Having discretised the spatial derivatives for both problems under the Black–Scholes and Heston's models, we now consider the first-order implicit scheme for the temporal discretisation giving

$$(A - \Delta \tau B) P_{l+1} = AP_l,$$

for $l = 1$ to $l_m$. 

©CMMSE
Page 475 of 1485
ISBN: 978-84-616-9216-3
where \( \Delta \tau = (\tau_{\text{max}} - \tau_{\text{min}})/l_m \), and \( l_m \) is the number of time steps. Then the truncation error for the approximation \( P \) to the exact solution \( P \) at \( \tau_l \) is of order \( O(\Delta \tau) \). We eliminate the temporal error terms, \( O(\Delta \tau^2), O(\Delta \tau^3), \cdots \), by combining different stages \( s \) of the same solution evaluated using different refinement time steps. Generally, for doubling the number of steps at each stage, an \( s \)–stage extrapolation tableau is constructed as [6]

\[
P_{p,q} = \frac{2^{p-1}P_{p,q-1} - P_{p-1,q-1}}{2^{p-1} - 1}, \quad \text{for } p = 2, \ldots, s \quad \text{and} \quad q = 2, \ldots, p.
\]

(15)

So, to obtain an overall fourth-order scheme, we perform a repeated extrapolation with \( s = 4 \) so that temporal errors up to the term in \( O(\Delta \tau^4) \) are eliminated.

### 6 Numerical results

Now, we conduct some numerical experiments to test the efficiency of our schemes. Parameters for the problem of pricing variance swaps under the Black–Scholes model are taken from [9] and those for the problem defined under the Heston’s model are taken from [11]. We compare the results with those obtained when the usual second-order discretisations are implemented in space together with the second-order Crank–Nicolson time stepping scheme.

Table 1: Error and convergence when pricing variance swap under the Black–Scholes model when \( AF = 12 \), \( x_0 = \log(1.2), \sigma = 0.25, \ T = 1, \ r = 0.05 \) and \( x \in [x_0 - 6 * 0.70 * \sqrt{T}, x_0 + 6 * 0.70 * \sqrt{T}] \). For Richardson extrapolation 10 initial time steps is used for 4 stages. The exact solution is 633.9662.

<table>
<thead>
<tr>
<th>( j_m = l_m )</th>
<th>( 2^{\text{nd}} \text{ order scheme} )</th>
<th>( \text{HOC scheme} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approx.</td>
<td>Error</td>
</tr>
<tr>
<td>20</td>
<td>736.1595</td>
<td>102.1933</td>
</tr>
<tr>
<td>40</td>
<td>658.8958</td>
<td>24.9296</td>
</tr>
<tr>
<td>80</td>
<td>640.1605</td>
<td>6.1943</td>
</tr>
<tr>
<td>160</td>
<td>635.5124</td>
<td>1.5462</td>
</tr>
</tbody>
</table>

Table 1 illustrates the results when pricing variance swaps under the Black–Scholes model for the case when \( AF = 12 \). Firstly, we can observe that the theoretical fourth-order of our scheme is recovered numerically. Secondly, bearing in mind that we are using the same number of grid points to calculate our solutions, errors for the second-order scheme are much more larger than those for our HOC scheme. Hence the HOC scheme is much more efficient than the Crank–Nicolson scheme used in [9]. Even on the coarse grid, \( j_m = 20 \), the error for the fourth-order compact scheme is about 25 times smaller than that for the second-order scheme. Moreover, we can observe that errors can go up to the order of \( e^3 \) for the HOC scheme while a dollar accurate price is not obtained for the second-order scheme.
Similarly, we record values in Table 2 for the Heston’s model. For comparison, we take $\Delta x = \Delta y$ when implementing the second-order scheme, which we recall to be a necessity for the HOC scheme. Here also, we can observe that errors for the HOC scheme are much smaller than for the second-order scheme.

Table 2: Error and convergence when pricing variance swap under the Heston’s model when $AF = 12$, $x_0 = \log(1.2)$, $v_0 = 0.04$, $\xi_v = 0.618$, $T = 0.5$, $r = 0.1$, $\kappa = 11.35$, $\rho = -0.64$, $\theta = 0.022$, $x \in [-1, 1]$ and $y \in [0, 0.5]$. For Richardson extrapolation 10 initial time steps is used for 4 stages. The exact solution is 242.7208.

<table>
<thead>
<tr>
<th>$j_m = t_m$</th>
<th>2nd order scheme</th>
<th>HOC scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approx.</td>
<td>Error</td>
</tr>
<tr>
<td>10</td>
<td>278.6019</td>
<td>35.8809</td>
</tr>
<tr>
<td>20</td>
<td>252.0407</td>
<td>9.3199</td>
</tr>
<tr>
<td>40</td>
<td>245.2118</td>
<td>2.4910</td>
</tr>
<tr>
<td>80</td>
<td>243.3819</td>
<td>0.6610</td>
</tr>
</tbody>
</table>

For the same parameters except for $AF$ which is now taken to be 52, we graphically represent our results in Figure 1 for pricing variance swap under the Black–Scholes and the Heston’s model. Here, the two diverging lines show that as $\log(j_m)$ increases, $\log(\text{error})$ decreases much faster for the HOC scheme than for the second-order scheme, showing its higher convergence. Hence, we can conclude that the HOC schemes are much more efficient than the second-order scheme.

![Figure 1: log(error) against log($j_m$).](image-url)
7 Conclusion

In this paper, we have applied both the Black–Scholes and the Heston’s stochastic volatility model to price discretely-sampled variance swaps using high-order compact schemes in space and an implicit-Richardson extrapolation time stepping scheme. This is a new approach in pricing variance swaps and our numerical experiments conducted in this work have proved its efficiency in comparison with the second-order Crank–Nicolson scheme proposed in [9]. Knowing that we have been using a PDE approach and that changing the definitions of the realised variance leads to only a change in the terminal conditions of our pricing problems, we can conclude that our approach is a practical tool to price variance swaps under the Black–Scholes and Heston’s stochastic volatility models. In a future work, we shall investigate the pricing of variance swaps under the Lévy driven models for which no analytical prices exist.

Acknowledgement

The work of Mehzabeen Dilloo has been supported by the University of Mauritius Postgraduate Scholarship.

References


A fourth order accurate finite difference solution of a multipoint nonlocal problem for the Laplace equation

A.A. Dosiyev

1 Department of Mathematics, Eastern Mediterranean University
emails: adiguzel.dosiyev@emu.edu.tr

Abstract

We consider the multipoint nonlocal boundary value problem for the two-dimensional Laplace equation in a rectangular domain. The solution of this problem is defined as a 9-point finite difference solution with the fourth order gluing operator of the local Dirichlet boundary value problem, by constructing a special method to find a function as the boundary value on the side of the rectangle, where the nonlocal condition is given. Numerical experiments are illustrated to support the analysis made.

Key words: Bitsadze-Samarskii problem, Elliptic equation, Nonlocal boundary value problems, Difference scheme
MSC 2000: 35A35, 65N06, 65N15

1 Introduction

After the paper of A.V Bitsadze and A.A. Samarskii [1], which stated a simple version of nonlocal boundary value problems, different generalizations of the nonlocal conditions and their approximate solutions were investigated by numerous authors (see [2] and references therein). As follows from the existing papers, difficulties arise in the analysis of both the exact and numerical solutions due to the existence of the nonlocal conditions.

In [3] and [4], a new constructive method for the solution of the Poisson equation with Bitsadze-Samarskii nonlocal boundary condition have been proposed. In [5], by the method of contraction mapping, the existence and uniqueness of the classical solution of the multipoint nonlocal problem is proved.

In this paper by the 9-point approximation of the Laplace equation with the fourth order interpolation operator, the approximate method used in [3] and [4] is generalized for
the multipoint nonlocal problem. The 9-point solution of the multipoint nonlocal problem is defined as 9-point solution of the Dirichlet problem by finding a function given as the boundary value on the side of the rectangle where the nonlocal condition is given. Finally, the numerical experiments are illustrated to support the obtained theoretical results.

2 Multipoint Nonlocal Boundary Value Problem

Let

\[ R = \{(x, y) : 0 < x < 1, 0 < y < 2\} \]  

be an open rectangle, \( \gamma^p, p = 1, 2, 3, 4 \) be its sides including the ends, numerated in the clockwise direction, starting with the side which lies on the y-axis, and let \( \gamma = \bigcup_{p=1}^{4} \gamma^p \) be the boundary of \( R, \overline{R} = R \cup \gamma \).

Let \( \eta_1, \eta_2, ..., \eta_m, \) and \( \alpha_1, \alpha_2, ..., \alpha_m \) be given numbers satisfying for some fixed number \( \delta > 0 \) the inequalities

\[ 0 < \delta \leq \eta_1 < \eta_2 < ... < \eta_m \leq 2 - \delta < 2, \]  

\[ \left(1 - \frac{\eta_1}{2}\right) \sum_{k=1}^{m} |\alpha_k| < 1. \]  

We consider the following nonlocal boundary value problem on \( R \):

\[ \Delta u = 0 \text{ on } R, \ u = 0 \text{ on } \gamma^1 \cup \gamma^3, \ u = \tau(x) \text{ on } \gamma^2, \]  

\[ \sum_{k=1}^{m} \alpha_k u(x, \eta_k) = u(x, 0), \ 0 \leq x \leq 1, \]  

where \( \Delta \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2 \) is the Laplace operator, \( \tau(x) \) is the given continuous function on \([0, 1]\) and \( \tau(0) = \tau(1) = 0 \).

The existence and uniqueness of the classical solution \( u \in C(\overline{R}) \cap C^2(R) \) of the problem (4), (5) is given in [6], [5].

3 Approximate solution of the nonlocal problem by the finite difference method

We say that \( \phi \in C^{k,\lambda}(D) \), if \( \phi \) has \( k \)-th derivatives on \( D \) satisfying the Hölder condition with exponent \( \lambda \). Let \( \tau(x) \in C^{4,\lambda}(\gamma^2), 0 < \lambda < 1 \).
We assign a square mesh \( D_h \) obtained with the lines \( x, y = 0, h, 2h, \ldots \) where \( h = \frac{1}{N} \) is the step size, \( N > 2 \) is an integer, such that \( h \) is less than half of the minimum length of the intervals \([0, \eta_1], [\eta_1, \eta_2], \ldots, [\eta_m, 2]\), and we denote by \( j_l \) the number for which
\[ j_l h \leq \eta_l < (j_l + 1)h. \]

We denote by \( R_h = D_h \cap R \), \( \gamma^p \) is the set of grids on \( \gamma^p \), \( p = 1, \ldots, 4 \), \( \gamma_h = \bigcup_{p=1}^{4} \gamma^p \), \( \overline{R}_h = R_h \cup \gamma_h \), and for each integer \( \mu, 1 \leq \mu \leq m \), the set of intersection points of the line \( y = \eta_\mu \) with the grid lines \( x = ih, i = 0, 1, \ldots, N \), is denoted by \( Y^\mu_h \).

Let \([0,1]_h = \{x = x_i, x_i = ih, i = 0, 1, \ldots, N; \ h = \frac{1}{N}\}\) be the set of nodes on the interval \([0,1] \) with the step size \( h \). Let \( C^0_h \) be the set of grid functions \( f_h \) on \([0,1]_h \) for which \( f_h(0) = f_h(1) = 0 \). We define the norm \( \| f_h \|_{C^0_h} = \max_{x \in [0,1]_h} |f_h| \). It is obvious that the space \( C^0_h \) is complete equipped with this norm.

Let \( \bar{v}_h \) be a solution of the finite difference problem
\[ \bar{v}_h = B \bar{v}_h \text{ on } R_h, \quad \bar{v}_h = \tau_h \text{ on } \gamma^2_h, \quad \bar{v}_h = 0 \text{ on } \gamma_h \setminus \gamma^2_h, \quad (6) \]
where \( \tau_h \) is the trace of \( \tau \) on \( \gamma^2_h \), and
\[ Bu(x,y) = (u(x+h,y) + u(x-h,y) + u(x,y+h) + u(x,y-h))/5 \]
\[ + (u(x+h,y+h) + u(x+h,y-h) + u(x-h,y+h) + u(x-h,y-h))/20. \]

On the basis of maximum principle, the problem (6) has a unique solution.

We define
\[ \varphi_h = \sum_{l=1}^{m} \alpha_l (S^4 \bar{v}_{hl}) \in C^0_h, \quad x \in [0,1]_h, \quad (7) \]
where \( S^4 \) is a fourth order gluing operator constructed in [7], for which \( S^4 \bar{v}_{hl} = \sum_{k=0}^{5} \mu_{k,l} \bar{v}_h(P_{k,l}) \), \( P_{0,l} = (x, lh_i), P_{1,l} = (x, (j_l + 1)h), P_{2,l} = (x - h, (j_l + 1)h), P_{3,l} = (x - h, j_l h), P_{4,l} = (x + h, (j_l + 1)h), P_{5,l} = (x + h, j_l h) \), and \( \mu_{k,l} \geq 0, \sum_{k=0}^{5} \mu_{k,l} = 1 \).

We consider the following problem
\[ w_h = Aw_h \text{ on } R_h, \quad w_h = 0 \text{ on } \gamma^m_h, \quad w_h = f_h \text{ on } \gamma^4_h, \quad (8) \]
where \( f_h \in C^0_h \), is an arbitrary function. On the basis of maximum principle for any \( f_h \) the problem (8) has a unique solution.

We introduce a linear operator \( B^h_l : C^0_h \to C^0_h \), and let for any grid function \( f_h \) \( f_h(x) \in C^0_h \)
\[ B^h_l f_h = S^4_h w_{hl} \in C^0_h, \quad l = 1, 2, \ldots, m, \quad (9) \]
where \( w_h \) is the solution of the problem (8).
On the basis of maximum principle, we obtain
\[ \left\| B^h_l f_h \right\|_{C^0_h} \leq \left\| f_h \right\|_{C^0_h} \left( 1 - \frac{\eta_l}{2} \right), \quad l = 1, 2, ..., m, \] (10)
i.e., the norm of operator $B^h_l$ does not exceed $q_l = 1 - \frac{\eta_l}{2}$.

Let $\tilde{\psi}_{l,h}$, $l = 1, 2, ..., m$ be the solution of the system of equations
\[ \tilde{\psi}_{l,h} = B^h_l \left( \tilde{\varphi}_h + \sum_{k=1}^{m} \alpha_k \tilde{\psi}_{k,h} \right), \quad l = 1, 2, ..., m. \] (11)

We seek the solution of system (11) by the following fixed-point iteration
\[ \tilde{\psi}_{0,l,h} = 0, \quad \tilde{\psi}_{n,l,h} = B^h_l \left( \tilde{\varphi}_h + \sum_{k=1}^{m} \alpha_k \tilde{\psi}_{n-1,k,h} \right), \quad l = 1, 2, ..., m; \quad n = 1, 2, ... \] (12)
On the basis of inequality (10) follows that the iteration converges to the unique solution of the equation (11).

We define the function
\[ \tilde{f}^n_h = \tilde{\varphi}_h + \sum_{l=1}^{m} \alpha_l \tilde{\psi}^n_{l,h}, \quad \text{on } \gamma_4, \] (13)
where $\tilde{\varphi}_h$ is defined by (7), $\tilde{\psi}^n_{l,h}$ is the $n$-th element of the sequence (12).

For the approximate solution of problem (4), (5), we take the solution of the following finite difference problem
\[ \tilde{u}^n_h = A \tilde{u}^n_h \text{ on } R_h, \quad \tilde{u}^n_h = 0 \text{ on } \gamma_2, \quad \tilde{u}^n_h = \tau_h \text{ on } \gamma_2, \] (14)
\[ \tilde{u}^n_h = \tilde{f}^n_h \text{ on } \gamma_4, \] (15)
where $\tilde{f}^n_h$ is defined by (13).

**Theorem 1** Let the boundary function $\tau(x)$ in the problem (4), (5) on $\gamma^2$ and $\tau^{(2)}(0) = \tau^{(2)}(1) = 0$. The next estimation holds
\[ \max_{(x,y) \in \Omega_h} |\tilde{u}^n_h - u_h| \leq c h^4 + c_1 \frac{q^{n+1}}{1 - q}, \quad n \geq 2, \] (16)
where $\tilde{u}^n_h$ is a solution of the problem (14), (15), $u_h$ is the trace of the exact solution of the nonlocal problem (4), (5), $c$ and $c_1$ are constants independent of $n$ and $h$, $q = \max \left\{ 1 - \frac{\eta_1}{2}, \left( 1 - \frac{\eta_2}{2} \right) \sum_{k=1}^{m} |\alpha_k| \right\} < 1.$
4 Numerical experiments

Let

\[ R = \{(x, y) : 0 < x < 1, 0 < y < 2\} . \]

We consider the next problem:

\[ \Delta u = 0 \text{ on } R, \quad u(0, y) = u(1, y) = 0, \]

\[ u(x, 0) = \frac{1}{2} u \left( x, \frac{3}{5} \right) + \frac{1}{4} u \left( x, \frac{6}{5} \right) + \frac{1}{4} u \left( x, \frac{9}{5} \right), \quad 0 \leq x \leq 1. \]

The exact solution of this problem is unknown. The approximate solution obtained by the proposed method is given on the lines \( y = \frac{3}{5}, y = \frac{6}{5} \) and \( y = \frac{9}{5} \). According to the repeated digits, for the decreasing mesh steps \( h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128} \) it follows that the maximum error on these lines decreases as \( O(h^4) \). To achieve this accuracy just 11 iterations are needed.

References


Improving the Solution of Band Linear Systems on Hybrid CPU+GPU Platforms

Ernesto Dufrechou\textsuperscript{1}, Pablo Ezzatti\textsuperscript{1}, Enrique S. Quintana-Ortí\textsuperscript{2} and Alfredo Remón\textsuperscript{3}

\textsuperscript{1} Instituto de la Computación, Universidad de la República, Uruguay
\textsuperscript{2} Depto. de Ingeniería y Ciencia de los Computadores, Universidad Jaime I, Spain
\textsuperscript{3} Max Planck Institute for Dynamics of Complex Technical Systems, Germany

emails: edufrechou@fing.edu.uy, pezzatti@fing.edu.uy, quintana@icc.uji.es, remon@mpi-magdeburg.mpg.de

Abstract

We address the solution of band linear systems on heterogeneous CPU-GPU platforms proposing a tuned routine that concurrently computes the LU factorization and solves the subsequent lower triangular system. This routine is complemented with an efficient band triangular system solver which is specially appealing to tackle several systems with the same coefficient matrix. The experimental evaluation performed on a platform equipped with an NVIDIA K20 GPU and an INTEL i3-3220 processor shows the convenience of our approach over the solutions provided by the MKL library and previous approaches.

Key words: band linear system, GPU

1 Introduction

The solution of linear systems with band coefficient matrix is required in several scientific areas such as finite element analysis in structural mechanics, domain decomposition methods for partial differential equations in civil engineering, and as part of matrix equations solvers in control and systems theory. In these applications, exploiting the banded structure of the matrix is mandatory since it allows to save both memory and arithmetic operations, by not storing neither operating with the null entries that lie out of the band [5].
The relevance of solving band linear systems motivates the inclusion of this operation in the LAPACK [1] specification. Moreover, when linked to a (multi-threaded) implementation of BLAS, LAPACK provides an efficient means to solve this problem on general-purpose (multicore) processors.

In a previous work [2] we presented tuned computational kernels that leverage the concurrency provided by graphics processors in order to accelerate the solution of band linear systems. In this work, we extend the aforementioned kernels to further exploit the computational capabilities of hybrid hardware platforms. In particular, we modify the algorithm to overlap computations on both devices so that the solution of the band matrix LU factorization and the lower band triangular linear system are computed concurrently. An experimental evaluation on a platform equipped with an NVIDIA K20 GPU and an INTEL i3-3220 multicore processor shows the superior performance attained by the new hybrid approach over the solver provided by the Intel MKL (Math Kernel Library) [6], as well as our previous proposal.

The rest of the paper is structured as follows. In Section 2 we review the use of LAPACK for the solution of band linear systems, and summarize our previous approach. Section 3 describes the different strategies followed by the new proposal. After that, in Section 4, the experimental evaluation of the new hybrid CPU-GPU solver is presented; and finally, Section 5 offers a few concluding remarks.

2 Solution of Band Linear Systems

2.1 LAPACK approach

LAPACK provides specific routines for the solution of a linear system with band coefficient matrix. In particular it defines a packed storage format for band matrices, and routines to compute the LU factorization and solve the subsequent triangular systems.
The storage format for band matrices employed by the BLAS and LAPACK routines permits a regular access pattern to the elements and, at the same time, allows important memory savings; Figure 1 shows an example of this storage format.

2.1.1 Factorization of band matrices

LAPACK includes two routines for the computation of the LU factorization of a band matrix: GBTF2 and GBTRF, which encode, respectively, unblocked and blocked algorithmic variants of the operation. In general, GBTRF is more efficient for large matrices; therefore we will focus hereafter on that particular algorithmic variant.

Given a band matrix $A \in \mathbb{R}^{n \times n}$ with lower and upper bandwidth $k_l$ and $k_u$, respectively, routine GBTRF computes the LINPACK-style LU factorization with partial pivoting

$$ L_{n-2}^{-1} \cdot P_{n-2} \cdots L_{1}^{-1} \cdot P_{1} \cdot L_{0}^{-1} \cdot P_{0} \cdot A = U $$

(1)

where $P_0, P_1, \ldots, P_{n-2} \in \mathbb{R}^{n \times n}$ are permutation matrices, $L_0, L_1, \ldots, L_{n-2} \in \mathbb{R}^{n \times n}$ represent Gauss transforms, and $U \in \mathbb{R}^{n \times n}$ is upper triangular with upper bandwidth $k_l + k_u$.

GBTRF presents two important drawbacks regarding its implementation in parallel architectures: (1) it involves some operations with triangular matrices requiring additional storage, memory copies and operations with null elements; (2) it requires the execution of small operations with a reduced inner parallelism.

2.1.2 Solution of triangular systems

The routine GBTRS from LAPACK tackles the subsequent band triangular systems that result from performing the LU factorization of the $A$ matrix by means of the GBTRF routine, as given in (1). This operation belongs to the BLAS-3 level.

Unlike GETRS, which is the equivalent LAPACK routine to solve general triangular linear systems, the implementation of GBTRS is based on BLAS-2 kernels, which limits its performance. Specifically, given a linear system $AX = B$ where $X, B \in \mathbb{R}^{n \times m}$ the routine proceeds as follows:

1. For $i = 0, 1, \ldots, n-2$, apply the permutation matrix $P_i$ to the right-hand side term $B$, and update this matrix with the corresponding multipliers in $L_i$:

   $B := P_i B, \quad (\text{SWAP})$

   $B := L_i^{-1} B, \quad (\text{GER})$

(2)

2. For $j = 1, 2, \ldots, m$ solve a triangular system with coefficient matrix $U$ and the right-hand side vector given by the $j$-th column of $B$ (denoted as $B_j$)

   $B_j := U^{-1} B_j, \quad (\text{TBSV})$

(3)
2.2 Our previous approach

Our routine in [2] includes a solver that does not suffer the limitations encountered in the LAPACK routines. In particular, we introduce minor changes to the matrix storage scheme that, at the cost a moderate increase in the memory requirements, allow to overcome the limitations from routine GBTRF and report relevant gains in terms of performance and execution time.

Specifically, our solution pads the packed data structure containing $A$ with $b$ extra rows at the bottom, where all the entries in this additional space initially set to zero. Then, the $n$ steps of the original implementation of GBTRF are transformed as follows:

1. Compute the LU factorization with partial pivoting

$$
P_1 \begin{pmatrix}
A_{11} \\
A_{21} \\
A_{31}
\end{pmatrix} = 
\begin{pmatrix}
L_{11} \\
L_{21} \\
L_{31}
\end{pmatrix} U_{11}, \quad (4)
$$

where the blocks of $L$ and $U$ overwrite the corresponding blocks of $A$.

2. Apply the permutations in $P_1$ to the remaining columns of the matrix:

$$
\begin{pmatrix}
A_{12} & A_{13} \\
A_{22} & A_{23} \\
A_{32} & A_{33}
\end{pmatrix} := 
P_1 \begin{pmatrix}
A_{12} & A_{13} \\
A_{22} & A_{23} \\
A_{32} & A_{33}
\end{pmatrix} \quad \text{(LASWP).} \quad (5)
$$

A single call to LASWP suffices now as the zeros at the bottom of the data structure and the additional $k_l$ superdiagonal set to zero in the structure ensure that fill-in may only occur in the elements in the lower triangular part of $A_{13}$.

3. Compute the updates:

$$
\begin{pmatrix}
A_{12} & A_{13}
\end{pmatrix} := L_{11}^{-1} \begin{pmatrix}
A_{12} & A_{13}
\end{pmatrix} \quad \text{(TRSM),} \quad (6)
$$

$$
\begin{pmatrix}
A_{22} & A_{23} \\
A_{32} & A_{33}
\end{pmatrix} := 
\begin{pmatrix}
A_{22} & A_{23} \\
A_{32} & A_{33}
\end{pmatrix} - 
\begin{pmatrix}
L_{21} \\
L_{31}
\end{pmatrix} \begin{pmatrix}
U_{12} & U_{13}
\end{pmatrix} \quad \text{(GEMM).} \quad (7)
$$

The lower triangular system in (6) returns a lower triangular block in $A_{13}$.

4. Undo the permutations on $[L_{11}^T, L_{21}^T, L_{31}^T]^T$ so that these blocks store the multipliers used in the LU factorization and, therefore $L_{31}$ is upper triangular.
3 Hybrid CPU-GPU Solver

The proposal is an extension of our previous solver for band linear systems based on the LU factorization. The central idea of the new approach is a reorganization of the procedure followed by a routine that leverages the computational power of heterogeneous CPU-GPU hardware platforms by overlapping some computations on both devices. With this aim, the novel method computes the update step of the LU factorization in the GPU concurrently with the solution of the first triangular band linear system in the CPU.

In our previous approach [2], the solution of both triangular band linear systems were executed after the factorization stage. In other words, \( L \) and \( U \) were completely obtained before the triangular systems \( LY = B \) and \( UX = Y \) were solved.

The new approach merges certain part of the solution of the band triangular linear systems together with the LU factorization. Thus, at each step of the loop that computes \( L \) (and \( U \)), the corresponding blocks of \( L \) matrix are obtained via (4). When this is completed, the procedure concurrently computes the update of the rest of the matrix \( A \) (equations (6) and (7)) in the GPU and the partial solution and update of \( B \) matrix (equation (2)) in the CPU.

By applying this new approach we can benefit of both:

- Data locality: we anticipate the solution of the linear system involving \( L \), as this operation is performed when the procedure is still manipulating these blocks.

- Concurrent execution: we can overlap two operations on two different processors, and therefore we can hide the computational cost of the least time-consuming one.

4 Experimental evaluation

The following experimental evaluation, on a platform equipped with an NVIDIA K20 GPU and an INTEL i3-3220 multicore processor, shows the superior performance of our approach over the solver provided by MKL and our previous proposal. We use a well-known benchmark problem [4] from the field of control theory as a case of study. This case is defined by a square sparse matrix with dimension 79,841, \( \text{RAIL}_{79841} \). In this case, the coefficient matrix is symmetric and sparse, but it does not present a band structure. By means of the RCM algorithm [3] we obtain a reordered band matrix with bandwidth \( k_u = k_l = 550 \). Additionally, we employ 5 right-hand-side matrices \( B \), each with a different number of columns \( m \), which mimic the dimensions that appear regularly on matrix equation solvers such as the LR-ADI algorithm [7] for Lyapunov equations.

Table 1 summarizes the runtime (in seconds) for the three solvers: the MKL solver (referred to as LAPACK), our previous approach, and the novel proposal.

The results obtained show that the new approach yields significant reductions on the execution time. It should be noted that our new proposal outperforms both solvers (MKL
and our previous method) for all the systems evaluated, even when the number of columns $m$ of $X$, $B$ is small.

5 Concluding Remarks

This work presented an extension of our previous solver for band linear system with the aim of leveraging the computational power of hybrid hardware platforms. Specifically, we designed and implemented a new version of this solver which maximizes the hardware usage overlapping the execution of some key computations on both devices.

The experimental evaluation conducted on an NVIDIA K20 GPU and an Intel i3-3220 multicore processor showed that the novel implementation significantly reduces the runtime required to solve band linear systems, even if the number of unknown vectors is small.

As part of future work, we plan to study the use of several GPUs to accelerate the band matrix linear system solver. Also, we plan to study the impact of the new hybrid algorithm on energy consumption.

Acknowledgements

Ernesto Dufrechou and Pablo Ezzatti acknowledge support from Programa de Desarrollo de las Ciencias Básicas, and Agencia Nacional de Investigación e Innovación, Uruguay. Enrique S. Quintana-Ortí was supported by project TIN2011-23283 of the Ministry of Science and Competitiveness (MINECO) and EU FEDER, and project P1-1B2013-20 of the Fundació Caixa Castelló-Bancaixa and UJI.
References


The influence of plotting positions on the correlation coefficient based on a Normal Q-Q Plot

María Dolores Estudillo-Martínez¹, Sonia Castillo-Gutiérrez¹ and Emilio Lozano-Aguilera¹

¹ Department of Statistics and Operations Research, University of Jaén
emails: mdestudi@ujaen.es, socasti@ujaen.es, elozano@ujaen.es

Abstract

In this contribution we highlight the influence of plotting positions used to construct a Normal Q-Q Plot to detect non-normality of a data set on the correlation coefficient obtained from this Normal Q-Q Plot. In particular, we will focus on the comparison of the value of the correlation coefficient obtained when Hazen’s plotting positions and Weibull’s plotting positions are used in a Normal Q-Q Plot.

Key words: Normal Q-Q Plot, Correlation Coefficient, Hazen’s plotting positions, Weibull’s plotting positions.

1 Introduction

A Normal Q-Q Plot allow us to detect non-normality of a data set. If the plotted points present a linear configuration, we can say that the observations come from a Normal distribution. To construct a Q-Q Plot, we use a key element: plotting positions. In the literature there are several definitions of plotting positions [1]. We will focus on plotting positions proposed by Hazen in 1930 [3]:

\[ p_i = \frac{i - 0.5}{n} \quad i = 1, ..., n \]

and plotting positions proposed by Weibull in 1939 [6]:

\[ p_i = \frac{i}{n + 1} \quad i = 1, ..., n \]
2 Correlation Coefficient based on a Normal Q-Q Plot

In 1975, Filliben [2] introduces the normal probability plot correlation coefficient as a test statistic in complete samples for the composite hypothesis of normality. We are mentioned above that if the plotted points present a linear configuration, then the observations come from a Normal distribution. The correlation coefficient is a simple and straightforward measure of linearity of a Normal Q-Q Plot.

Therefore, we have to calculate the correlation coefficient between the ordered observations $x_{(i)}$ $i = 1, ..., n$ and $\Phi^{-1}(p_i)$ $i = 1, ..., n$ where $\Phi$ is the standard normal cumulative distribution function and $p_i$ are the selected plotting positions.

Therefore, the value of the correlation coefficient for a same set of observations will be different if we change the choice of plotting positions.

3 Example

In this section, we present an example where we can observe the influence of plotting positions on the correlation coefficient obtained from a Normal Q-Q Plot.

Table 1 shows a simulated size 10 sample of a Cauchy distribution.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-7.3969323</td>
<td>-0.6987647</td>
<td>-1.3263917</td>
<td>0.9041168</td>
<td>-3.0412808</td>
</tr>
<tr>
<td>2.5359498</td>
<td>0.3047345</td>
<td>0.5952836</td>
<td>10.8483261</td>
<td>0.7673032</td>
</tr>
</tbody>
</table>

Table 1: Simulated sample of a Cauchy distribution

Figure 1 shows a Normal Q-Q Plot constructed from the above observations. Plotting positions considered, $p_i$, are Hazen’s definition and Weibull’s definition. We can also represent the correlation coefficient value obtained using both definitions of plotting positions.

We can observe that exists differences between plotted points obtained using both definitions of plotting positions and the value of the correlation coefficient is different too. By mean of a simulation study, 10000 samples for each size $n = 3(1)100$ were generated from a Normal distribution to obtain the empirical percentage of rejections under the null hypothesis of normality. For a sample of size 10 and a significance level at 5%, this values are: 0.9183118 for Hazen’s plotting positions and 0.9131884 for Weibull’s plotting positions.

If we compare both values, the empirical percentage of rejections under the null hypothesis of normality and the correlation coefficient on the graph, we can observe that the hypothesis of normality is rejected according to Weibull’s plotting positions and it is not rejected according to Hazen’s plotting positions.
The influence of plotting positions on the correlation coefficient based on a Normal Q-Q Plot

Figure 1: Sample observations in a Normal Q-Q Plot using two different plotting positions

References


Multiresolution analysis for two-dimensional interpolatory schemes on uniform grids

L. Fernández¹, M. A. Fortes¹ and M. L. Rodríguez¹

¹ Departamento de Matemática Aplicada, Universidad de Granada
emails: lidiafr@ugr.es, mafortes@ugr.es, miguelrg@ugr.es

Abstract

We develop a multiresolution interpolatory scheme in a two dimensional frame based on Harten’s work in a one-dimensional context ([2]). We obtain the decomposition and reconstruction algorithms for different interpolatory schemes and we apply them in several examples regarding data compression and discontinuities detection.

Key words: Multiresolution analysis, decomposition-reconstruction algorithms, compression data, discontinuities detection.

1 Introduction

In [1] Beam and Warming propose to extend Harten’s interpolatory multiresolution representation ([2]) to include Hermite interpolation in the unit interval [0, 1]. The authors develop decomposition and reconstruction algorithms and, as application, they show the compression features of the proposed method. In this work we extend the work of Beam and Warming to the two-dimensional frame of the unit square [0, 1] × [0, 1]. We consider different interpolatory schemes on uniform grids of [0, 1] × [0, 1], we develop the corresponding algorithms and we apply them in several examples to check the higher degree of accuracy in compressing data and discontinuities detection.

2 Notation and preliminaries

Let us consider, for each $k \in \mathbb{N} \cup \{0\}$, two sequences $\{X^k\}$ and $\{Y^k\}$ of partitions of the unit interval $[0, 1]$ defined inductively as follows:
Let now be
\[ X^k = \{ x_1^k, x_2^k, \ldots, x_{2^k+1}^k \} \quad \text{and} \quad Y^k = \{ y_1^k, y_2^k, \ldots, y_{2^k+1}^k \}, \]
with \( x_1^k = y_1^k = 0; \) \( x_{2^i+1}^k = y_{2^i+1}^k = 1; \) \( x_i^k < x_{i+1}^k \) and \( y_i^k < y_{i+1}^k \) for all \( i = 1, \ldots, 2^k. \)

Then, for each \( i = 1, \ldots, 2^k, \) we consider two points \( x_{2i-1}^k = x_i^k \) and \( y_{2i-1}^k = y_i^k \) for all \( i = 1, \ldots, 2^k + 1, \) and we consider the new partitions
\[ X^{k+1} = \{ x_1^{k+1}, x_2^{k+1}, \ldots, x_{2^{k+1}+1}^{k+1} \} \quad \text{and} \quad Y^{k+1} = \{ y_1^{k+1}, y_2^{k+1}, \ldots, y_{2^{k+1}+1}^{k+1} \}. \]

Observe that \( X^k \) and \( Y^k \) are therefore obtained from \( X^{k+1} \) and \( Y^{k+1} \) by removing the knots with even subindices.

For all \( k \geq 0 \) and for all \( i = 1, \ldots, 2^k, \) let \( h_x^{k,i} = x_{i+1}^k - x_i^k \) and \( h_y^{k,i} = y_{i+1}^k - y_i^k, \) in such a way that
\[ h_x^{k,i} = h_x^{k+1,2i-1} + h_x^{k+1,2i} \quad \text{and} \quad h_y^{k,i} = h_y^{k+1,2i-1} + h_y^{k+1,2i} \quad \text{for all} \quad i = 1, \ldots, 2^k. \]

Let us consider, for any \( k \geq 0, \) the partition \( T^k \) of \([0,1] \times [0,1]\) whose set of knots is \( D^k = X^k \times Y^k. \) We will denote the elements of \( D^k \) by \( d_i^k = (x_i^k, y_i^k) \) for \( i = 1, \ldots, 2^k + 1, \) (in Figure 1 it is shown \( D^2 \)). Observe that the knots \( d_{i,j}^{k+1} \) of \( D^{k+1} \) with \( i, j \) odd are exactly the ones of \( D^k \) (more precisely, \( d_{i,j}^{k+1} = d_{2i+1,2j+1}^k \) for \( i, j = 1, \ldots, 2^k + 1). \)

At each resolution level \( k \geq 0, \) and at each knot \( d_{i,j}^k, \) for \( i, j = 1, \ldots, 2^k + 1, \) we will associate a vector values
\[ u_{i,j}^k = (u_{1,1}^k, \ldots, u_{2^k,2^k}^k)' \in \mathbb{R}^{N^2}. \]

\[ \vdots \]
and we will consider the column vector
\[ \mathbf{U}^k = \begin{pmatrix} \mathbf{u}^{k+1}_{1,1}, \ldots, \mathbf{u}^{k+1}_{2k+1,1} \\ \mathbf{u}^{k+1}_{1,2}, \ldots, \mathbf{u}^{k+1}_{2k+1,2} \\ \vdots \\ \mathbf{u}^{k+1}_{1,2k+1}, \ldots, \mathbf{u}^{k+1}_{2k+1,2k+1} \end{pmatrix} \in \mathcal{M}_{N(2^k+1)^2}, \]

where \( \cdot \)' denotes the transposition operation.

The values in \( \mathbf{U}^k \) may come from an empirical experiment or from a function \( u \) defined in \([0, 1] \times [0, 1] \).

Our main objective in this work is to establish a multiresolution process in order to transfer information between different resolution levels. More precisely, we want to obtain both the values in \( \mathbf{U}^k \) at resolution \( k \) knowing the values in \( \mathbf{U}^{k+1} \) at resolution \( k + 1 \) (decomposition process), and the values in \( \mathbf{U}^{k+1} \) knowing the values in \( \mathbf{U}^k \) (reconstruction process).

All throughout this work, \( \mathbf{I}_n \) will denote the identity matrix of order \( n \) and \( \mathbf{0}_{m \times n} \) will denote the zero matrix of order \( n \times m \).

### 3 Decomposition algorithm

We want to get the values in \( \mathbf{U}^k \) knowing the ones in \( \mathbf{U}^{k+1} \). To this end we just have to decimate the values in \( \mathbf{U}^{k+1} \) associated to knots not belonging to \( \mathbf{U}^k \):

\[ \mathbf{u}^{k+1}_{i,j} = \mathbf{u}^{k+1}_{2i-1,2j-1} \text{ for } i, j = 1, \ldots, 2^k + 1. \]

The matricial expression of the decomposition process is then

\[ \mathbf{U}^k = \mathbf{L}^D \mathbf{U}^{k+1}, \]

being

\[ \mathbf{L}^D = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{I} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mathbf{I} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \mathbf{I} \end{pmatrix} \in \mathcal{M}_{N(2^k+1)^2 \times N(2^{k+1}+1)^2}, \]

where \( \mathbf{0} = \mathbf{0}_{N(2^k+1) \times N(2^{k+1}+1)} \), and \( \mathbf{I} \), appearing \( 2^k + 1 \) times in \( \mathbf{L}^D \), is defined as

\[ \mathbf{I} = \begin{pmatrix} \mathbf{I}_N & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_N \end{pmatrix} \in \mathcal{M}_{N(2^k+1) \times N(2^{k+1}+1)}, \]

where \( \mathbf{0} = \mathbf{0}_{N \times N} \).
4 Reconstruction algorithm

We want now to obtain an approximation \( \hat{u}_{i,j}^{k+1} \) of all the vectors \( u_{i,j}^{k+1} (i, j = 1, \ldots, 2^{k+1} + 1) \) in \( U^{k+1} \) knowing the ones in \( U^k \). To this aim, first observe that those values in \( U^{k+1} \) associated to the knots in \( D^{k+1} \) that also belong to \( D^k \) can be exactly reconstructed, i.e.:

\[
\hat{u}_{i,j}^{k+1} = u_{i,j}^{k+1} = u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k}, \quad \forall i, j = 1, \ldots, 2^{k+1} + 1; \quad i, j \text{ odd.} \tag{3}
\]

The remaining approximations \( \hat{u}_{i,j}^{k+1} \), for \( i, j \) non both odd, will be obtained by using some of nearest vectors \( u^k \)'s in \( U^k \). More precisely:

- The vector values \( u_{i,j}^{k+1} \), for \( i \) even, \( j \) odd (corresponding to points \( t_{i,j}^{k+1} \) lying on horizontal segments of \( T^k \)) will be obtained by using the two adjacent vector values \( u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k} \) and \( u_{i+\frac{1}{2}, j+\frac{1}{2}+1}^{k} \) by means of a linear interpolation formula that in matricial form will be expressed as

\[
\hat{u}_{i,j}^{k+1} = \text{Int}_{m}^{H, i,j} u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k} + \text{Int}_{m}^{H, i,j} u_{i+\frac{1}{2}, j+\frac{1}{2}+1}^{k}, \tag{4}
\]

where \( \text{Int}_{m}^{H, i,j} \) is a square matrix of order \( N \) for \( m = 1, 2 \).

- The vector values \( u_{i,j}^{k+1} \), for \( i \) odd, \( j \) even (corresponding to points \( t_{i,j}^{k+1} \) lying on vertical segments of \( T^k \)) will be obtained by using the two adjacent vector values \( u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k} \) and \( u_{i+\frac{1}{2}+1, j+\frac{1}{2}}^{k} \) by means of a linear interpolation formula that in matricial form will be expressed as

\[
\hat{u}_{i,j}^{k+1} = \text{Int}_{m}^{V, i,j} u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k} + \text{Int}_{m}^{V, i,j} u_{i+\frac{1}{2}+1, j+\frac{1}{2}}^{k}, \tag{5}
\]

where \( \text{Int}_{m}^{V, i,j} \) is a square matrix of order \( N \) for \( m = 1, 2 \).

- The vector values \( u_{i,j}^{k+1} \), for \( i, j \) even (corresponding to points \( t_{i,j}^{k+1} \) inside the rectangles of \( T^k \)) will be obtained by using the four adjacent vector values \( u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k}, u_{i+\frac{1}{2}+1, j+\frac{1}{2}}^{k}, u_{i+\frac{1}{2}, j+\frac{1}{2}+1}^{k} \) and \( u_{i+\frac{1}{2}+1, j+\frac{1}{2}+1}^{k} \) by means of a linear interpolation formula that in matricial form will be expressed as

\[
\hat{u}_{i,j}^{k+1} = \text{Int}_{m}^{S, i,j} u_{i+\frac{1}{2}, j+\frac{1}{2}}^{k} + \text{Int}_{m}^{S, i,j} u_{i+\frac{1}{2}+1, j+\frac{1}{2}}^{k} + \text{Int}_{m}^{S, i,j} u_{i+\frac{1}{2}, j+\frac{1}{2}+1}^{k} + \text{Int}_{m}^{S, i,j} u_{i+\frac{1}{2}+1, j+\frac{1}{2}+1}^{k}, \tag{6}
\]

where \( \text{Int}_{m}^{S, i,j} \) is a square matrix of order \( N \) for \( m = 11, 12, 21, 22 \).

Observe that the fact that the interpolation matrices in the three cases depend on \( i, j \) allows to consider different interpolation formulas to obtain each of the vectors \( \hat{u}_{i,j}^{k+1} \) (for \( i, j = 1, \ldots, 2^{k+1} + 1, \) non both odd).
It is possible to write this reconstruction algorithm in matricial form as

$$\tilde{U}^{k+1} = L^R \tilde{U}^k,$$

where $$\tilde{U}^{k+1} \in \mathcal{M}_{N(2^{k+1}+1)^2}$$ is defined as

$$\tilde{U}^{k+1} = \begin{pmatrix}
\begin{pmatrix}(a_{1,1}^{k+1}, \ldots, a_{2^{k+1}+1}^{k+1} ) \\
(a_{1,2}^{k+1}, \ldots, a_{2^{k+1}+1}^{k+1} ) \\
\ddots \\
(a_{2^{k+1}+1}, \ldots, a_{2^{k+1}+1, 2^{k+1}+1}^{k+1} )
\end{pmatrix} \\
\cdots \\
\begin{pmatrix}(a_{1,1}^{k}, \ldots, a_{2^{k+1}+1}^{k} ) \\
(a_{1,2}^{k}, \ldots, a_{2^{k+1}+1}^{k} ) \\
\ddots \\
(a_{2^{k+1}+1}, \ldots, a_{2^{k+1}+1, 2^{k+1}+1}^{k} )
\end{pmatrix}
\end{pmatrix},$$

and $$L^R \in \mathcal{M}_{N(2^{k+1}+1)^2 \times N(2^{k+1}+1)^2}$$ is the structured block matrix:

$$L^R = \begin{pmatrix}
L_1^R & 0 & \cdots & 0 \\
L_2^R & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
L_{2^{k+1}+1}^R & 0 & \cdots & 0
\end{pmatrix},$$

where:

$$\sim 0 = 0_{N(2^{k+1}+1) \times N(2^{k+1})};$$

$$\sim$$ for $$j$$ odd, $$L_{j}^R = \in \mathcal{M}_{N(2^{k+1}+1) \times N(2^{k+1})}$$ is the matrix defined as

$$L_{j}^R = \begin{pmatrix}
\text{Id}_N & 0 & \cdots & 0 \\
\text{Int}_{H1}^{1,2,j} & 0 & \cdots & 0 \\
0 & \text{Id}_N & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & \cdots & \text{Id}_N \\
0 & 0 & \cdots & \text{Int}_{H1}^{1,2^{k+1}+1,j} \\
\text{Int}_{V1}^{k+1,2^{k+1}+1,j} & 0 & \cdots & 0
\end{pmatrix},$$

with $$0 = 0_{N \times N};$$

$$\sim$$ for $$j$$ even, $$L_{j}^R = (L_{j,1}^R \mid L_{j,II}^R) \in \mathcal{M}_{N(2^{k+1}+1) \times 2N(2^{k+1})}$$ is the matrix defined as

$$L_{j,1}^R = \begin{pmatrix}
\text{Int}_{V1}^{k+1,2^{k+1}+1,j} & 0 & \cdots & 0 \\
\text{Int}_{S1}^{k+1,2^{k+1}+1,j} & 0 & \cdots & 0 \\
0 & \text{Int}_{V1}^{k+1,2^{k+1}+1,j} & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & \cdots & \text{Int}_{S1}^{k+1,2^{k+1}+1,j} \\
0 & 0 & \cdots & \text{Int}_{V1}^{k+1,2^{k+1}+1,j}
\end{pmatrix},$$
Multiresolution analysis for two-dimensional interpolatory schemes on uniform grids

\[ L^R_{j,II} = \begin{pmatrix}
\text{Int}V_{k,1,j} & \ldots & 0 \\
0 & \text{Int}S_{k,1,j} & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \text{Int}V_{k,2^{k+1}-1,j}
\end{pmatrix},
\]

with \( 0 = 0_{N \times N} \).

5 Detail coefficients

We will save an extra information in the decomposition process (following Harten multiresolution approach) in order to the reconstruction be exact. More precisely we define the detail coefficients

\[ r^k_{i,j} = u^{k+1}_{i,j} - u^{k+1}_{i,j}, \quad \text{for} \ k \geq 0; \ i,j = 1, \ldots, 2^{k+1} + 1. \]

Observe that, due to (3), \( r^k_{i,j} = 0 \) for \( i,j \) both odd. Let us consider the vector \( R^k \) formed by the non necessarily zero detail coefficients \( r^k_{i,j} \) (i.e., with \( i,j \) non both odds):

\[ R^k = \begin{pmatrix}
\begin{pmatrix} r^k_{2,1}, r^k_{3,1}, \ldots, r^k_{2^{k+1},1} \end{pmatrix} \\
\begin{pmatrix} r^k_{2,1}, r^k_{3,1}, \ldots, r^k_{2^{k+1},1} \end{pmatrix} \\
\vdots \\
\begin{pmatrix} r^k_{2,1}, r^k_{3,1}, \ldots, r^k_{2^{k+1},1} \end{pmatrix}
\end{pmatrix} \in \mathcal{M}_{N2^k(2+32^k)}.
\]

Then, taking into account (3), (4), (5) and (6) we obtain

\[ R^k = H^D u^{k+1}, \quad (9) \]

\( H^D \in \mathcal{M}_{N2^k(2+32^k) \times N(2^{k+1}+1)^2} \) being the matrix defined as

\[ H^D = \begin{pmatrix}
H^D_{1,1} & \ldots & 0 & \ldots & 0 \\
\ldots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \ldots & 0 \\
0 & \ldots & 0 & \ldots & 0 \\
0 & \ldots & 0 & \ldots & 0 \\
\end{pmatrix},
\]

where:

\[ \sim 0 = 0_{N2^k \times N(2^{k+1}+1)}. \]
\[ H^D_j = \begin{pmatrix} -\text{Int}H^k_{1,1} & \text{Id}_N & 0 & 0 & \ldots & 0 & 0 & 0 \\ \text{Id}_N & -\text{Int}H^k_{1,1} & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}H^k_{2,2} & \text{Id}_N \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}H^k_{2,2} & \text{Id}_N \\ \end{pmatrix}, \]

with \(0 = 0_{N \times N};\)

\[ H^{D_{j,I}} = \begin{pmatrix} -\text{Int}V^k_{1,1} & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \text{IntH}^k_{1,1} & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}V^k_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}V^k_{2,2} & 0 \\ \end{pmatrix}, \]

with \(0 = 0_{N \times N};\)

\[ H^{D_{j,II}} = \begin{pmatrix} -\text{Int}V^k_{1,1} & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \text{IntH}^k_{1,1} & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}V^k_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & -\text{Int}V^k_{2,2} & 0 \\ \end{pmatrix}, \]

with \(0 = 0_{N \times N}.\)

Hence the **decomposition algorithm** is

\[
\begin{align*}
U^k &= L^D U^{k+1} \\
R^k &= H^D U^{k+1} 
\end{align*}
\] (11)

Finally, the **reconstruction algorithm** is

\[
U^{k+1} = U^{k+1} + H^R R^k = L^R U^k + H^R R^k,
\] (12)

where \(H^R = M_{N(2k+1)^2 \times N2^k(2+3-2k)}\) is the matrix defined as

\[
H^R = \begin{pmatrix} H^R_1 & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{Id}_N(2k+1+1) \\ \end{pmatrix},
\]
where
\[ 0 = 0_{N(2^k+1+1) \times N(2^k+1+1)}; \]
\[ H^R_0 \in \mathcal{M}_{N(2^k+1+1) \times N2^k} \] is the matrix defined as
\[
H^R_0 = \begin{pmatrix}
0_{N(2^k+1+1) \times N(2^k+1+1)} & \ldots & 0 \\
0 & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & 0
\end{pmatrix},
\]
with \( 0 = 0_{N \times N}. \)

Of course, the reconstruction and decomposition process are inverse one each other, and hence, it holds:
\[ U^{k+1} = L^R L^D U^{k+1} + H^R H^D U^{k+1} \implies L^R L^D + H^R H^D = \text{Id}_{N(2^k+1+1)^2} \]
and
\[
\begin{pmatrix} L^D \\ H^D \end{pmatrix} \begin{pmatrix} L^R & H^R \end{pmatrix} \begin{pmatrix} U^k \\ \mathcal{R}^k \end{pmatrix} = \begin{pmatrix} U^k \\ \mathcal{R}^k \end{pmatrix} \implies \begin{pmatrix} L^D L^R = \text{Id}_{N(2^k+1)^2} \\ L^D H^R = 0_{N(2^k+1) \times N2^k(2+3 \cdot 2^k)} \\ H^D L^R = 0_{N2^k(2+3 \cdot 2^k) \times N(2^k+1)^2} \\ H^D H^R = \text{Id}_{N2^k(2+3 \cdot 2^k)} \end{pmatrix}
\]

6 Applications

In order to develop the applications we consider \( N = 1 \) in Equation (1), i.e., we will consider just one single value \( u_{i,j}^k = ((u_{i,j}^k)_1) := u_{i,j}^k \) associated to each of the \( d_{i,j}^k \)'s. In this case, in order to interpolate the values \( u_{i,j}^{k+1} \), for \( i, j \) non both odd, from the values in \( \mathcal{U}^k \), we will consider the following schemes:

\( \triangleright \) The values \( u_{i,j}^{k+1} \), for \( i \) even, \( j \) odd (corresponding to horizontal segments of \( \mathcal{T}^k \)) will be approximated by linear interpolation of \( u_{\frac{i}{2}+1, \frac{j+1}{2}}^k \) and \( u_{\frac{i}{2}+1, \frac{j+1}{2}}^k \), i.e.
\[
\tilde{u}_{i,j}^{k+1} = \frac{h_{x}^{k+1, i} \cdot u_{\frac{i}{2}+1, \frac{j+1}{2}}^k + h_{x}^{k+1, i-1} \cdot u_{\frac{i}{2}+1, \frac{j+1}{2}}^k}{h_{x}^{k, \frac{i}{2}}}, \tag{13}
\]
in such a way that the \( 1 \times 1 \)-matrices \( \text{Int} H_{m,i}^{k,i,j} \) appearing in (4) become
\[
\text{Int} H_1^{k,i,j} = \left( \frac{h_{x}^{k+1, i}}{h_{x}^{k, \frac{i}{2}}} \right) \quad \text{and} \quad \text{Int} H_2^{k,i,j} = \left( \frac{h_{x}^{k+1, i-1}}{h_{x}^{k, \frac{i}{2}}} \right).
\]
The values $u_{i,j}^{k+1}$, for $i$, $j$ even (corresponding to vertical segments of $T^k$) will be approximated by linear interpolation of $u_{i+1/2,j}^k$ and $u_{i+1/2,j+1}^k$, i.e.:

$$\tilde{u}_{i,j}^{k+1} = \frac{h_y^{k+1,j} \cdot u_{i+1/2,j}^k + h_y^{k+1,j-1} \cdot u_{i+1/2,j+1}^k}{h_y^{k+1,j} + h_y^{k+1,j-1}},$$

(14)

in such a way that the $1 \times 1$-matrices $\text{Int}V_{m}^{k,i,j}$ appearing in (5) become

$$\text{Int}V_1^{k,i,j} = \left( \frac{h_y^{k+1,j}}{h_y^{k+1,j}} \right) \quad \text{and} \quad \text{Int}V_2^{k,i,j} = \left( \frac{h_y^{k+1,j-1}}{h_y^{k+1,j}} \right).$$

Observe that $\text{Int}H_{m}^{k,i,j}$ (resp. $\text{Int}V_{m}^{k,i,j}$) do not depend on $j$ (resp. on $i$) since the way to obtain the value $\tilde{u}_{i,j}^{k+1}$ from $u_{i+1/2,j}^k$ and $u_{i+1/2,j+1}^k$ (resp. $u_{i+1/2,j}^k$, $u_{i+1/2,j+1}^k$) do not depend on component $j$ (resp. $i$), just on component $i$ (resp. $j$).

The values $u_{i,j}^{k+1}$, for $i$, $j$ even (corresponding to points inside the squares of $T^k$) will be approximated by weighted average of $u_{i+1/2,j}^k$, $u_{i+1/2,j+1}^k$, $u_{i+1/2,j}^k$, and $u_{i+1/2,j+1}^k$, i.e.:

$$\tilde{u}_{i,j}^{k+1} = \frac{h_y^{k+1,j} \left( h_x^{k+1,i} \cdot u_{i+1/2,j}^k + h_x^{k+1,i-1} \cdot u_{i+1/2,j+1}^k \right)}{h_x^{k+1,i} \cdot h_y^{k+1,j}} \quad + \quad \frac{h_y^{k+1,j-1} \left( h_x^{k+1,i} \cdot u_{i+1/2,j}^k + h_x^{k+1,i-1} \cdot u_{i+1/2,j+1}^k \right)}{h_x^{k+1,i} \cdot h_y^{k+1,j}},$$

(15)

in such a way that the $1 \times 1$-matrices $\text{Int}S_{m}^{k,i,j}$ appearing in (6) become

$$\text{Int}S_{11}^{k,i,j} = \left( \frac{h_y^{k+1,j} \cdot h_x^{k+1,i}}{h_y^{k+1,j} \cdot h_x^{k+1,i}} \right) \quad ; \quad \text{Int}S_{21}^{k,i,j} = \left( \frac{h_y^{k+1,j-1} \cdot h_x^{k+1,i-1}}{h_y^{k+1,j} \cdot h_x^{k+1,i}} \right);$$

$$\text{Int}S_{12}^{k,i,j} = \left( \frac{h_y^{k+1,j-1} \cdot h_x^{k+1,i}}{h_y^{k+1,j} \cdot h_x^{k+1,i}} \right) \quad ; \quad \text{Int}S_{22}^{k,i,j} = \left( \frac{h_y^{k+1,j-1} \cdot h_x^{k+1,i-1}}{h_y^{k+1,j} \cdot h_x^{k+1,i}} \right).$$
6.1 Data compression

In this application we will check the high accuracy of the algorithm when compressing data. We have considered the function \( g(x, y) = \sin(\pi(x^2 + y^2)) \) and, for different values of \( k \), we have applied the following algorithm:

1. We have computed \( u_{i,j}^k = g(d_{i,j}^k) \),
2. We have decomposed \( \mathcal{U}^k \) into the approximation coefficients \( \mathcal{U}^{k-1} \) at lower level \( k \), and the detail coefficients \( \mathcal{R}^{k-1} \).
3. We have computed the truncated detail coefficients \( \tilde{\mathcal{R}}^{k-1} \) by applying the following threshold criteria:

\[
\tilde{r}_{i,j}^{k-1} = \begin{cases} 
    r_{i,j}^{k-1} & \text{if } |r_{i,j}^{k-1}| > \tau \cdot \text{Max} \{ |r_{i,j}^{k-1}| \}_{i,j} \\
    0 & \text{if } |r_{i,j}^{k-1}| \leq \tau \cdot \text{Max} \{ |r_{i,j}^{k-1}| \}_{i,j}
\end{cases}
\]

where \( \tau \in [0, 1] \) is the threshold parameter to be chosen.
4. We have reconstructed an approximation \( \tilde{\mathcal{U}}^k \) of \( \mathcal{U}^k \) by using \( \mathcal{U}^{k-1} \) and \( \tilde{\mathcal{R}}^{k-1} \).
5. We have computed an estimation of the error committed by means of the expression

\[ E_k = \frac{\sum_{i,j}(\tilde{u}_{i,j}^k - u_{i,j}^k)^2}{\sum_{i,j}(u_{i,j}^k)^2} \]

In the next table we show, for different values of \( k \) and \( \tau \), the results obtained for \( E_k \) as well as the percentage \( P_{k-1} = \frac{\text{card} \{ r_{i,j}^{k-1} : r_{i,j}^{k-1} \neq 0 \} }{\text{card} (\mathcal{R}^{k-1})} \) of coefficients used in the reconstruction process.

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
<th>( k = 4 )</th>
<th>( k = 5 )</th>
<th>( k = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau = 0.25 )</td>
<td>( E_k )</td>
<td>( 0 )</td>
<td>( 9.22 \cdot 10^{-4} )</td>
<td>( 3.38 \cdot 10^{-4} )</td>
<td>( 2.16 \cdot 10^{-4} )</td>
<td>( 1.53 \cdot 10^{-4} )</td>
</tr>
<tr>
<td></td>
<td>( P_{k-1} )</td>
<td>100%</td>
<td>62.5%</td>
<td>35.71%</td>
<td>33.66%</td>
<td>31.88%</td>
</tr>
<tr>
<td>( \tau = 0.5 )</td>
<td>( E_k )</td>
<td>0</td>
<td>2.07 \cdot 10^{-2}</td>
<td>1.56 \cdot 10^{-2}</td>
<td>1.07 \cdot 10^{-3}</td>
<td>7.18 \cdot 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>( P_{k-1} )</td>
<td>100%</td>
<td>50%</td>
<td>5.36%</td>
<td>4.81%</td>
<td>4.38%</td>
</tr>
<tr>
<td>( \tau = 0.75 )</td>
<td>( E_k )</td>
<td>0.67</td>
<td>5.11 \cdot 10^{-2}</td>
<td>1.73 \cdot 10^{-2}</td>
<td>1.32 \cdot 10^{-3}</td>
<td>8.44 \cdot 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>( P_{k-1} )</td>
<td>20%</td>
<td>37.5%</td>
<td>1.79%</td>
<td>1.44%</td>
<td>1.75%</td>
</tr>
<tr>
<td>( \tau = 1 )</td>
<td>( E_k )</td>
<td>1</td>
<td>0.27</td>
<td>2.18 \cdot 10^{-2}</td>
<td>1.6 \cdot 10^{-3}</td>
<td>1.07 \cdot 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>( P_{k-1} )</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
6.2 Discontinuities detection

In this second application, we propose to detect the discontinuity of the function

\[ f(x, y) = \begin{cases} 
\frac{4}{10} - \sqrt{(x - 0.5)^2 + (y - 0.5)^2} & \text{if } (x - 0.5)^2 + (y - 0.5)^2 \leq 0.25^2, \\
0 & \text{if } (x - 0.5)^2 + (y - 0.5)^2 > 0.25^2.
\end{cases} \]

Fig. 2 shows the graphic of \( f \) and of the discontinuity line.

![Image](image.png)

Figure 2.

We have considered \( X^k = Y^k \) as the uniform partition of \([0, 1]\) into \( 2^k \) subintervals, for \( k \geq 1 \). The procedure to detect the discontinuity consists of the following steps:

1. For a given value of \( k \), we consider \( u_{i,j}^k = f(d_{i,j}^k) \).
2. Next we decompose \( U^k \) into \( U^{k-1} \) and \( R^{k-1} \).
3. We have plotted the points of \( D^{k-1} \) in which the detail coefficients are higher.

The results are shown in Fig. 3.
Multiresolution analysis for two-dimensional interpolatory schemes on uniform grids

Figure 3. Detection of the discontinuity line for different resolution levels \( k \).

References


Fractional modelling of Pennes’ bioheat equation using distributed order differential equations

L. L. Ferrás\textsuperscript{1}, N. J. Ford\textsuperscript{2}, M. L. Morgado\textsuperscript{3}, J. M. Nóbrega\textsuperscript{1} and M. Rebelo\textsuperscript{4}

\textsuperscript{1} Institute for Polymers and Composites, University of Minho, Campus de Azurém 4800-058 Guimarães, Portugal
\textsuperscript{2} Department of Mathematics, University of Chester, CH1 4BJ, UK
\textsuperscript{3} Department of Mathematics, University of Trás-os-Montes e Alto Douro, UTAD, Quinta de Prados 5001-801, Vila Real, Portugal
\textsuperscript{4} Department of Mathematics, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal

emails: luis.ferras@dep.uminho.pt, njford@chester.ac.uk, luisam@utad.pt, mnobrega@dep.uminho.pt, mjr@fct.unl.pt

Abstract

In this work we provide a new mathematical model for the Pennes’ bioheat equation, assuming a fractional time derivative of distributed order. Different versions of the bioheat equation are considered, that take into account the temperature-dependent variability in the tissue perfusion, and that comprise both finite and infinite speed propagation of heat signals. The bioheat proposed model is solved numerically using an implicit difference method. Different weight functions for the order of integration are used and tested, aiming to optimize the numerical approach. The results obtained with the distributed order fractional model, are compared with the original models that use classical (integer order) derivatives and with the fractional bioheat model (without distributed order).

Key words: Fractional differential equations, distributed order differential equation, Caputo derivative, bioheat equation, implicit numerical scheme.
1 Introduction

The way temperature diffuses in our body, has been a subject of interest for a long time. From the practical method of measuring the body temperature with our own hands, to the use of highly sophisticated measuring devices, we can find a diverse number of alternative possibilities and intense theoretical and experimental research work that resulted in great advances and increased knowledge on temperature distribution inside the human body.

The pioneering work of Harry H. Pennes [1] in 1948 is the cornerstone of the mathematical modelling of temperature diffusion in tissues, but, as it happens with most of the initial modelling approaches, it requires some improvements. Moreover, this model was originally derived for modelling the temperature in a human forearm, but, it is extensively used by several authors for modelling temperature diffusion in different tissues (such as the analysis of hyperthermia in cancer treatment [2]).

Pennes’ [1] bioheat transfer equation (see also [3, 4, 5, 6, 7, 8]), which describes the thermal distribution in human tissue, taking into account the influence of blood flow, (see figure (1)) is given by,

\[ \rho c_t \frac{\partial T(x,t)}{\partial t} = k \frac{\partial^2 T(x,t)}{\partial x^2} + W_b c_b (T_a - T) + q_m, \quad t > 0, \quad 0 \leq x \leq L, \]  

(1)

where \( \rho_t, c_t \), and \( k \) are the density \([kg/m^3]\), the specific heat \([J/(kg\,°C)]\), and the tissue thermal conductivity\([W/(m\,°C)]\), respectively; \( W_b \) is the mass flow rate of blood per unit volume of tissue \([kg/(s\,m^3)]\); \( c_b \) is the blood specific heat; \( q_m \) is the metabolic heat generation per unit volume \([W/m^3]\); \( T_a \) represents the temperature of arterial blood \(°C\); \( T \) is the temperature and the term \( W_b c_b (T_a - T) \) represents the perfusion of the blood. It is worth mentioning that the \( W_b \) constant was experimentally obtained by Pennes for a human forearm (he adjusted \( W_b \) until the temperature theoretical results matched the experimental ones).

In order to overcome Pennes’ bioheat model limitations, other models were proposed. Since in Eq. 1 the blood velocity field is not taken into account, in 1974, Wulff [9] and Klinger [10] considered the local blood mass flux to account the blood flow direction. Also, Pennes assumed that thermal equilibration occurs in the capillaries, but in 1980 Chen and Holmes [11] showed that the major heat transfer processes occur in the 50 to 500 \( \mu m \) diameter vessels. Based on the Klinger model [10], Chen and Holmes proposed a new model by adding dispersion and microcircularatory perfusion terms, and, in 1984, Weinbaum, Jiji and Lemons [12] presented a new vascular bioheat model by considering the countercurrent blood flow (this way, the blood leaving the tissue can also influence the temperature of the medium).

All these models, although sophisticated, do not take into account the role of thermoregulation. Therefore, in 2010, Zolfaghari and Maerefat [13] developed the simplified thermoregulatory bioheat (STB) model that takes into account the thermoregulatory mech-
anisms of the human body (shivering, regulatory sweating and vasomotion). The model is a combination of Pennes’ bioheat equation and Gagge’s two-node model (thermal comfort model) \cite{14, 15}. This model proved to be reasonably accurate, showing a good fit of experimental data \cite{16, 17}.

Although these models are more complete, and, theoretically more accurate than the classic Pennes’ bioheat equation, their complexity makes them quite complicated to handle (some of the field variables needed for the models to work, are difficult to obtain), and, adjust to acquired experimental data. On the other hand, Pennes’ equation is simple, with a small number of physical parameters, thus attracting researchers from different fields, encouraging the continued improvement of the model.

Different versions of this model have been proposed in the literature, that take into account the temperature-dependent variability in the tissue perfusion \cite{7, 8, 18, 19}, and, with the thermal conductivity being either depth-dependent or temperature-dependent,

\[
\rho_c c_t \frac{\partial (T(x,t) - T_s)}{\partial t} = \frac{\partial}{\partial x} \left( k(\cdot) \frac{\partial (T(x,t) - T_s)}{\partial x} \right) - W_b c_b (T(x,t) - T_s) - \omega_1 \left( T(x,t) - T_s - \frac{q_m}{c_b \omega_b} \right)^2 + q_m \quad t > 0, \quad 0 \leq x \leq L, \tag{2}
\]

and

\[
\frac{\partial T(x,t)}{\partial t} = k \frac{\partial^2 T(x,t)}{\partial x^2} + \left[ \omega_0 + \omega_1 \frac{T(x,t) - T_a}{T_a} \right] (T_a - T(x,t)) \quad t > 0, \quad 0 \leq x \leq L, \tag{3}
\]

(with Eq. 3 a simplified version of Eq. 2) where $T(x,t)$ is the local tissue temperature, $T_s$ represents the skin steady-state temperature, $k(\cdot)$ is the thermal diffusivity that is either a
Fractional modelling of Pennes’ equation using distributed order differential equations

function of \( x \), or, a function of the temperature increase \( T(x,t) - T_s \), \( \omega_b \) is the blood perfusion rate, and \( \omega_0 \) and \( \omega_1 \) are the temperature independent and dependent (respectively) perfusion coefficients. Note that these two equations are nonlinear while the original Pennes’ bioheat equation is linear.

Since fractional calculus showed improvement in the modelling of most physical phenomena (specially processes with memory), Damor et al. [20] proposed a fractional version of the bioheat equation, by replacing the first-order time derivative with a derivative of arbitrary positive real order \( \alpha \),

\[
\rho c_t \partial_t^\alpha T(x,t) = k \partial_x^2 T(x,t) + W_b c_b (T_a - T) + q_m, \quad t > 0, \quad 0 \leq x \leq L, \quad (4)
\]

More recently, Ezzat et al. [21] (see also [22]) presented a new mathematical model for the Pennes’ bioheat equation using a fractional version of the Fourier law for temperature. They constructed a model that comprises both the classic (parabolic) and hyperbolic Pennes’ bioheat equation (note that the hyperbolic equation [23] ensures a finite speed pulse propagation while an infinite speed is obtained with the classic one), being given by,

\[
\rho c_t \partial_t \left( T(x,t) + \frac{\tau_0}{\alpha} \partial_t^\alpha T(x,t) \right) = k \partial_x^2 T(x,t) + \left( q_m (x,t) + \frac{\tau_0}{\alpha} \partial_t^\alpha q_m(x,t) \right), \quad t > 0, \quad 0 \leq x \leq L, \quad (5)
\]

Other versions of the model exist in the literature that basically are tailored versions of the original bioheat equation (Eq. 1), with the purpose of modelling very specific cases.

The paper is organized as follows: in the next section we describe the fractional differential equation of distributed order used in this work; in Section 3 we describe the numerical scheme, and, in Section 4 we perform numerical studies with the modified Pennes’ model. The paper ends with Section 5, where we provide some conclusions and plans for further investigation.

2 Fractional differential equations of distributed order

The different versions of the bioheat equation presented before, specifically Eqs. 1, 2, 3, and 4, are now written, using the time-fractional derivative [24] of distributed order instead of the first-order time derivative, \( \partial T(x,t)/\partial t \), generalising, in this way, not only the model originally derived by Pennes, but also the models containing time-fractional derivatives. The equations are given in a compact form as a diffusion equation with distributed order in time, with a source term that depends on the temperature, \( T(x,t) \), and a diffusion coefficient, \( k(\cdot) \), that may be a function of temperature and the spatial variable, \( x \).

\[
\int_0^1 c(\alpha) \partial_t^\alpha T(x,t) \, d\alpha = \frac{k(\cdot) \partial_x^2 T(x,t)}{\rho c_t} + \frac{f(x,t,T(x,t))}{\rho c_t} \quad t > 0, \quad 0 \leq x \leq L, \quad (6)
\]
where $\frac{\partial^\alpha}{\partial t^\alpha}$ is the fractional Caputo derivative of arbitrary real order $\alpha$ given by [25],

$$\frac{\partial^\alpha T(x, t)}{\partial t^\alpha} = \frac{1}{\Gamma (1 - \alpha)} \int_0^t (t - s)^{-\alpha} \frac{\partial T(x, s)}{\partial s} \, ds$$

(with $0 < \alpha < 1$), and $c(\alpha)$ is a weight function for the order of differentiation with support somewhere in the interval $[0, 1]$, such that [26],

$$c(\alpha) \geq 0 \text{ and } \int_0^1 c(\alpha) \, d\alpha = C > 0$$

In terms of physical and analytical aspects, a justification for the time-fractional diffusion equation of distributed order can be found in [24, 26, 27]. In particular, for Eq. 6 with $f(x, t, T(x, t)) \equiv 0$, $\alpha = 1$, $\frac{k(\cdot)}{\rho c_0} = 1$, initial condition $T(x, t) = \delta(x)$ and decay to zero conditions for $|x| \to \infty$, the solution is given by the Gaussian probability density function with a second moment (variance) growing linearly with time (a characteristic of normal diffusion) [24].

For the time-fractional diffusion equation of distributed order the weight function $c(\alpha)$ plays an important role when modelling physical phenomena. For $c(\alpha) = 1$ ($0 \leq \alpha \leq 1$) we obtain a super-slow diffusion process (logarithmic growth) and for $c(\alpha) = b_1 \delta(\alpha - \alpha_1) + b_2 \delta(\alpha - \alpha_2)$ (with $0 < \alpha_1 < \alpha_2 \leq 1$, $b_1 > 0$, $b_2 > 0$, $b_1 + b_2 = 1$) a slow diffusion of power law growth is obtained (see [24, 28]). In [28], the case, $c(\alpha) = (\alpha_2 - \alpha_1)^{-1}$ if $0 \leq \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$ and 0 otherwise, was also considered.

It seems that the type of weight functions ($c(\alpha)$) proposed in the literature, is limited to simple functions, due to the difficult task of finding analytical solutions for the time-fractional diffusion equation of distributed order. In practical terms, if we imagine the integral on the left-hand-side of Eq. 6 as a discrete operator given by the finite sum of the integrand functions (a sum of various modes), the concept of distributed order can be seen, in terms of modelling, as a multi-mode model, that, through linear (or nonlinear) combination of its various modes, $c(\alpha_i) \frac{\partial^{\alpha_i} T(x, t)}{\partial t^{\alpha_i}}$, increases the probability of getting a better modelling of the physical phenomena, since a higher number of parameters is used. Therefore, in this work we consider different weight functions $c(\alpha)$:

$$c(\alpha) = 1$$

$$c(\alpha) = x$$

$$c(\alpha) = 1 + x$$
### 3 Numerical solution

For the numerical solution of Eq. 6, we will use a method similar to the one proposed in [29], that seems to be the only numerical method existing in the literature for the solution of this type of equations (see also [30]). Note that in [29], the source term is given by a function $f(x,t)$, that is only a function of time and space. Here we are dealing with a reaction-diffusion equation since the source term also depends on the unknown temperature field. Moreover, the boundary conditions here are not of the same kind as the ones considered in [29].

For the numerical solution of Eq. 6, first we need to approximate the time and spatial derivatives, and, the integral on the left-hand-side of the equation. Following [29], the integral is approximated using the midpoint rule (second order approximation),

$$
\int_0^1 c(\alpha) \frac{\partial^{\alpha} T(x,t)}{\partial \alpha} d\alpha \approx h \sum_{j=1}^{N} c(\alpha_j) \frac{\partial^{\alpha_j} T(x,t)}{\partial \alpha_j},
$$

the diffusive term (first term on the right-hand-side of Eq. 6) is approximated using a second order finite difference formula,

$$
\frac{\partial^2 T(x_i,t)}{\partial x^2} \approx \frac{T(x_{i+1},t) - 2T(x_i,t) + T(x_{i-1},t)}{(\Delta x)^2}
$$

and for the fractional derivative we use the backward finite difference formula provided by Diethelm ($O((\Delta t)^{2-\alpha})$),

$$
\frac{\partial^{\alpha} T(x_i,t_l)}{\partial \alpha} \approx \frac{(\Delta t)^{-\alpha_j}}{\Gamma(2-\alpha_j)} \sum_{m=0}^{l} a^{(\alpha_j)}_{m,l} (T(x_i,t_{l-m}) - T(x_i,0))
$$

The source term, $f(x,t,T(x,t))$, is simply given by $f(x_i,t_l,T(x_i,t_l))$, with $f(x,t,T(x,t))$ a linear or nonlinear function of $T(x,t)$.

As shown in Fig. 2, we assume uniform meshes for the midpoint rule ($h = 1/N$) and both spatial ($\Delta x = L/K$) and time discretizations ($\Delta t = T/R$), with $N$, $K$ and $R$ the number of divisions of each grid.

Denoting $T(x_i,t_l)$ by $T_i^l$, the finite difference scheme is then given by,

$$
h \sum_{j=1}^{N} c(\alpha_j) \frac{(\Delta t)^{-\alpha_j} \Gamma(2-\alpha_j)}{1} \sum_{m=0}^{l} a^{(\alpha_j)}_{m,l} (T_{i-m}^l - T_i^0) = \frac{T_{i+1}^l - 2T_i^l + T_{i-1}^l}{(\Delta x)^2} + f(x_i,t_l,T_i^l)
$$

for $i = 1, ..., K - 1$, $l = 1, ..., R$. 

©CMMSE ISBN: 978-84-616-9216-3
3.1 Boundary conditions

In their paper [29], Ford et al. proved that their numerical scheme is convergent and unconditionally stable, assuming an initial condition

$$T(x, 0) = g(x),$$

and the Dirichlet boundary conditions,

$$T(0, t) = T_0, \quad T(L, t) = T_L,$$

Here, we consider the following Neumann conditions (constant heat flux)

$$-k \frac{\partial T(x, t)}{\partial x} \bigg|_{x=L} = 0,$$

$$-k \frac{\partial T(x, t)}{\partial x} \bigg|_{x=0} = q_0$$

and a initial condition,

$$T(x, 0) = T_a,$$

4 Validation and numerical results

Single order

For validation of the numerical method, we compared the results obtained by the method described in the previous section with the analytical solution derived by [5] for $\alpha = 1,$
Fractional modelling of Pennes’ equation using distributed order differential equations

\[ T(x,t) = T_a + \frac{q_0}{\sqrt{4kW_bC_b}} \left[ e^{-\frac{W_bC_b}{4k}x} \operatorname{erfc}\left(\frac{x}{\sqrt{4kC_b\rho_t}}\right) - e^{\frac{W_bC_b}{4k}x} \operatorname{erfc}\left(\frac{x}{\sqrt{4kC_b\rho_t}}\right) \right] \]  

(21)

For that purpose, we chose the same coefficients as the ones given in [20], but, we set the metabolic heat generation to zero. Therefore, we have used \( \rho_t = 1050 \), \( c_t = 4180 \), \( k = 0.5 \), \( W_b = 0.5 \), \( C_b = 3770 \), \( T_a = 37 \), \( L = 0.02 \) and \( q_m = 0.999 \).

In Fig. (3) (a) we can see the results obtained for \( \Delta x = 0.0001 \) and \( \Delta t = 0.25 \) for three different time intervals. Since the analytical solution was derived for \( \alpha = 1 \), we used a value of \( \alpha \) such that \( \alpha \to 1 (\alpha = 0.999) \). As shown, a good agreement was obtained between the numerical and the analytical solutions. These numerical results were obtained without the distributed order, since, for this case, \( \alpha \) is a constant value. For that, we drop the summation operator in Eq. 15.

We see that the temperature increases with time, due to the influence of the source term that represents the perfusion of the blood. This feature can be easily seen if we plot the variation of temperature, \( T \), with time, \( t \), for \( x = \text{cte} \), as shown in Fig. (3) (b). We can also see the influence of \( \alpha \) on the temperature. After a certain time, when \( \alpha \) increases the temperature decreases.

Distributed order

As explained before, a generalization of the classic and fractional bioheat model, can be performed by using a fractional derivative of distributed order. To validate the numerical
Figure 4: Comparison of the numerical results with the analytical solution for $t = 0.9$ s.

Figure 5: (a) Variation of temperature with time for different weight functions. The solid line represents the classic bioheat equation. (b) Comparison between the temperature evolution with time, for the time-fractional single order and distributed order bioheat equations.

As shown in Fig. (4) we can see a good agreement between the numerical results and the exact solution.

We also solved numerically the Pennes’ bioheat with fractional time derivative of distributed order, using different weight functions, Eqs. 9, 10 and 11 (see Fig. (5)) (a). We see that the weight functions have a huge influence on the temperature distribution. As shown
in Fig (5) (b), the distributed order bioheat equation with $c(\alpha) = 1$ shows a stronger subdiffusive effect, when compared with the fractional single order distribution.

5 Conclusions

We derived a numerical method for the solution of the fractional bioheat equation of single and distributed order. For the single order, we observed that for a fixed, $x$, and varying $t$ from zero to a certain point $t_\alpha \sim 1$, when the order of the time derivative increases, the temperature decreases; after that point $t_\alpha$, the opposite behavior is observed. For the distributed order we concluded that the weight functions have a huge influence on the velocity distribution, therefore, a right choice of these functions can lead to a better modelling. In a future work, the proofs of convergence and stability of the numerical method will be provided, and, we will also study the properties of the weight functions, so that the theoretical results can match the experimental data.

Acknowledgements


References


Fractional modelling of Pennes’ equation using distributed order differential equations


The effect of reversible binding sites on drug release from drug eluting stents

José A. Ferreira¹, Jahed Naghipoor¹ and Paula de Oliveira¹

¹ CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal

emails: ferreira@mat.uc.pt, jahed@mat.uc.pt, poliveir@mat.uc.pt

Abstract

A coupled model of a cardiovascular drug delivery system using a biodegradable drug eluting stent is proposed. A reversible reaction between the drug and the binding sites in the arterial tissue is taken into account. The proposed model is sensitive to the different nature of the therapeutic compounds used. The numerical results are obtained using an IMEX finite element method in the variational form.

Key words: reversible binding, non-Fickian coupled model, cardiovascular drug delivery, drug eluting stent, numerical simulation.

1 Introduction

A drug eluting stent (DES) is a stent placed into a narrowed vessel to release anti-proliferative drug into its walls with a programmed pharmacokinetics. It consists of a metallic stent strut coated with a polymeric layer that encapsulates a therapeutic drug. The drug is used to reduce smooth muscle cell growth and to prevent inflammatory response which are the predominant causes of neointimal proliferation and in-stent restenosis.
During the last years, a number of studies have proposed mathematical models to simulate drug delivery in the cardiovascular tissues. We refer without being exhaustive to [1, 3, 4, 5, 6, 7, 9]. Most of these studies address the release of drug and its numerical behavior while the behaviour of the biodegradable materials is disregarded. In [2], we proposed a coupled reaction diffusion cardiovascular drug delivery system where biodegradation of polylactic acid was taken into account.

In this paper, we develop the model proposed in [2] to consider reversible binding between the drug and specific sites inside the arterial wall ([4]). We can distinguish between hydrophobic drugs, which are retained within the tissue and hydrophilic which are rapidly cleared and sometimes are ineffective.

The paper is organized as follows. Section 2 is devoted to the description of reversible binding reactions and to setup the model and its initial, boundary and interface conditions. The effect of reversible binding sites, in the presence of drug in the arterial wall, and the behaviour of drugs with different reversible binding properties are discussed in Section 3. Finally in Section 4 some conclusions are presented.

2 Reaction-Diffusion-Convection Model

When a DES is implanted in a vessel, the coated stent will be gradually covered by neo-intima. For a sake of simplicity, DES is considered to be already embedded in the arterial wall (see Figure 1). These are reasonable assumptions because of the complex dynamics of tissue healing and regrowth which take place immediately after DES implantation in the arterial wall. However, since growing of neo-intima and endothelium cells around DES are not well studied in the literature, the evolution of neo-intima around the stent is considered negligible ([4]).

The complex multi-layer structure of the arterial wall is lumped into a homogeneous porous material whose physical properties corresponds to the intermediate layer, that is the media. We also neglect the dynamics of the drug into the blood flow. We assume that the drug that is conveyed by the blood flow is immediately transported away without influencing the downstream region of the artery.

Let us consider a DES coated with polylactic acid (PLA) where the drug is distributed
Figure 1: Drug eluting stent embedded in the arterial wall (left: http://www.ibmt.med.uni-rostock.de/nachwuchsgruppe.html).

uniformly. PLA is a biodegradable polymer and three main reactions are responsible for its degradation into lactic acid and oligomers both in the coated stent and in the arterial wall. In the first reaction, the hydrolysis of the PLA occurs resulting in small molecules with smaller molecular weights i.e. oligomers (with molecular weight $M_W$ such that $2 \times 10^4 \text{ g/mol} \leq M_W \leq 1.2 \times 10^5 \text{ g/mol}$) and lactic acid (with molecular weight $M_W \leq 2 \times 10^4 \text{ g/mol}$). The second and the third reactions are the hydrolysis of the oligomers resulting in lactic acid occurring in the coating and in the arterial wall respectively.

In what follows the subscript $S$ stands for the coated stent while the subscript $V$ stands for the arterial wall. Let $C_{1,S}$ and $C_{1,V}$ be the concentrations of plasma in the stent and in the arterial wall respectively. The concentrations of oligomers in the stent and in the arterial wall are denoted by $C_{3,S}$ and $C_{3,V}$ respectively. By $C_{4,S}$ and $C_{4,V}$ we denote the concentrations of lactic acid in the stent and in the arterial wall respectively. By $C_{2,S}$ we represent the concentration of PLA while concentration of drug in the stent is represented by $C_{5,S}$.

The reactions are represented schematically by

\[
C_{1,S} + C_{2,S} \xrightarrow{\kappa_{1,S}} C_{3,S} + C_{4,S}, \quad C_{1,S} + C_{3,S} \xrightarrow{\kappa_{2,S}} C_{4,S}, \quad C_{1,V} + C_{3,V} \xrightarrow{\kappa_{1,V}} C_{4,V},
\]

where $\kappa_{1,S}$ and $\kappa_{2,S}$ denote the reaction rates of the hydrolysis of PLA and oligomers in the
stent and $\kappa_{1,V}$ denotes the reaction rate of the hydrolysis of oligomers in the arterial wall.

Binding occurs when a ligand (drug) and a receptor (binding site) collide due to diffusion forces and when the collision has the correct orientation and enough energy ([4]). When binding has occurred, drug and binding sites remain bound together for an amount of time depending on their affinity. After dissociation, the drug and the binding site keep the properties they have before the binding. The drug-binding site reaction is schematically represented by

\[
\text{Drug + Binding sites} \xrightarrow{\text{association}} \text{Drug-binding complex}.
\]  

(2)

To define the mathematical kinetic model associated to (2), the following assumptions are made: all the binding sites are equally accessible to the drug; all the binding sites are either free or bound to the drug, there are not states of partial binding; neither drug nor binding site are altered by binding ([4]).

The concentration of free drug in the arterial wall is represented by $C_{5,V}$ with initial concentration $C_{5,V}^0 = 0$, while $C_{6,V}$ represents the concentration of free binding sites in the arterial wall with initial concentration $C_{6,V}^0 \neq 0$. The concentration of activated drug-binding sites is represented by $C_{7,V}$, and we assume that its initial concentration is null. The drug-binding reaction is schematically represented by

\[
C_{5,V} + C_{6,V} \xrightarrow{\kappa_{b,V}} C_{7,V},
\]  

(3)

where $\kappa_{b,V}$ is the association rate between the drug and the binding sites and $\kappa_{u,V}$ is the dissociation rate.

The drug assumes two different states: the dissolved state where drug moves by convection and diffusion and the bound state where drug attaches reversibly to specific sites inside the arterial wall and is no longer diffused or transported by plasma.

It should be noted that $K_b = \frac{C_{6,V}^0 \kappa_{b,V}}{\kappa_{u,V}} \gg 1$ corresponds to drugs that have high affinity for their target binding sites.

The coupled nonlinear reaction-diffusion-convection model that describes the evolution of PLA and its compounds, the drug and the free and the activated drug-binding sites, is
defined by

\[
\begin{align*}
\frac{\partial C_{m,S}}{\partial t} &= -\nabla \cdot J_{m,S}(C_S) + F_{m,S}(C_S) \quad \text{in } S \times \mathbb{R}^+, \ m = 1, \ldots, 5, \\
\frac{\partial C_{m,V}}{\partial t} &= -\nabla \cdot J_{m,V}(C_V) + F_{m,V}(C_V) \quad \text{in } V \times \mathbb{R}^+, \ m = 1, \ldots, 5, \ m \neq 2, \\
\frac{\partial C_{6,V}}{\partial t} &= F_{6,V}(C_V) \quad \text{in } V \times \mathbb{R}^+, \\
\frac{\partial C_{7,V}}{\partial t} &= F_{7,V}(C_V) \quad \text{in } V \times \mathbb{R}^+, \\
\end{align*}
\]

(4)

where \(C_S = (C_{m,S})_{m=1,\ldots,5}\) and \(C_V = (C_{m,V})_{m=1,\ldots,7}\), and the mass fluxes in the stent and in the arterial wall are defined respectively by

\[
\begin{align*}
J_{m,S}(C_S) &= -D_{m,S}\nabla C_{m,S} + u_SC_{m,S}, \ m = 1, \ldots, 5, \\
J_{m,V}(C_V) &= -D_{m,V}\nabla C_{m,V} + u_VC_{m,V}, \ m = 1, \ldots, 5, \ m \neq 2. \\
\end{align*}
\]

(5)

In [6], \(D_{m,S}, \ m = 1, \ldots, 5\), represent the diffusion coefficients of the species in the stent and are defined by

\[
D_{m,S} = D^0_{m,S}e^{\alpha_{m,S}\frac{C^0_{2,S} - C_{2,S}}{c^0_{2, S}}} \quad \text{in } S \times \mathbb{R}^+, \ m = 1, \ldots, 5, 
\]

(6)

where \(D^0_{m,S}, \ m = 1, \ldots, 5\), are the diffusion coefficients of the corresponding species in the unhydrolyzed PLA and \(C^0_{2,S}\) is the concentration of the unhydrolyzed PLA at \(t = 0\). For a sake of simplicity, we assume that the diffusion coefficients in the arterial wall \(D_{m,V}, \ m = 1, \ldots, 5, \ m \neq 2\), are constants.

In [4], \(F_{m,S}, \ m = 1, \ldots, 5\), introduced in [6], are reaction terms defined by

\[
F_{m,S}(C_S) = \begin{cases} 
-\sum_{i=1,2} F_{i,S}(C_S), & m=1, \\
-F_{1,S}(C_S), & m=2, \\
\sum_{i=1,2} (-1)^{i-1} F_{i,S}(C_S), & m=3, \\
\sum_{i=1,2} F_{i,S}(C_S), & m=4, \\
0, & m=5,
\end{cases}
\]

(7)

that describe the degradation of PLA.

We assume that the degradation of oligomers and also the binding and unbinding of
the drug take place in the arterial wall and are defined by

\[ F_{m,V}(C_V) = \begin{cases} 
    -F_{1,V}(C_V), & m=1, \\
    -F_{1,V}(C_V), & m=3, \\
    F_{1,V}(C_V), & m=4, \\
    -F_{2,V}(C_V), & m=5, \\
    -F_{2,V}(C_V), & m=6, \\
    F_{2,V}(C_V), & m=7. 
\end{cases} \]  

(8)

In (7) and (8) the following definitions are used

\[ \begin{align*}
F_{1,S}(C_S) &= \kappa_{1,S}C_{1,S}C_{2,S}(1 + \alpha C_{4,S}), \\
F_{2,S}(C_S) &= \kappa_{2,S}C_{1,S}C_{3,S}(1 + \beta C_{4,S}), \\
F_{1,V}(C_V) &= \kappa_{1,V}C_{1,V}C_{3,V}(1 + \gamma C_{4,V}), \\
F_{2,V}(C_V) &= \kappa_{6,V}C_{5,V}C_{6,V} - \kappa_{u,V}C_{7,V},
\end{align*} \]  

(9)

where \( \alpha, \beta \) and \( \gamma \) are some positive dimensional constants.

The velocities \( u_j, \ j=S,V \), in (4) are obtained by solving Darcy’s equations

\[ \begin{align*}
u_V &= -\frac{k_V}{\mu_V} \nabla p_V \quad \text{in } V, \\
\nabla \cdot u_V &= 0 \quad \text{in } V, \\
p_V &= p_{\text{lumen}} \quad \text{on } \Gamma_{\text{lumen}}, \\
p_V &= p_{\text{adv}} \quad \text{on } \Gamma_{\text{adv}}, \\
u_V \cdot \eta_V &= 0 \quad \text{on } \Gamma_{\text{wall}},
\end{align*} \]  

(10)

in the arterial wall and

\[ \begin{align*}
u_S &= -\frac{k_S}{\mu_S} \nabla p_S \quad \text{in } S, \\
\nabla \cdot u_S &= 0 \quad \text{in } S, \\
u_S \cdot \eta_S &= 0 \quad \text{on } \Gamma_{\text{strut}},
\end{align*} \]  

(11)

in the stent, where \( \eta_S \) and \( \eta_V \) represent exterior unit normals.

Equations (10) and (11) are completed with the interface condition

\[ \begin{align*}
p_S &= p_V \quad \text{on } \Gamma_{\text{coat}}, \\
u_S \cdot \eta_S &= -u_V \cdot \eta_V \quad \text{on } \Gamma_{\text{coat}}.
\end{align*} \]  

(12)

In (10) and (11), we assume that the permeabilities \( k_j, \ j=S,V \), and the viscosities \( \mu_j, \ j=S,V \), are constants.
To complete the coupled problem (4)-(12), we define in what follows the initial, the boundary and the interface conditions. The initial conditions in the coating and in the arterial wall are given by

\[
\begin{align*}
C_{m,S}(0) &= 0, \quad m = 1, 3, 4, \\
C_{m,S}(0) &= C_0^m, \quad m = 2, 5, \\
C_{m,V}(0) &= C_0^m, \quad m = 1, 6, \\
C_{m,V}(0) &= 0, \quad m = 3, 4, 5, 7.
\end{align*}
\]  
(13)

The boundary and interface conditions are defined by

\[
\begin{align*}
J_{m,S} \cdot \eta_S &= 0 \quad \text{on } \Gamma_{\text{strut}} \times \mathbb{R}^+, \quad m = 1, \ldots, 5, \\
J_{2,S} \cdot \eta_S &= 0 \quad \text{on } \Gamma_{\text{coat}} \times \mathbb{R}^+, \\
C_{m,S} &= C_{m,V} \quad \text{on } \Gamma_{\text{coat}} \times \mathbb{R}^+, \quad m = 1, \ldots, 5, \quad m \neq 2, \\
J_{m,S} \cdot \eta_S &= -J_{m,V} \cdot \eta_V \quad \text{on } \Gamma_{\text{coat}} \times \mathbb{R}^+, \quad m = 1, \ldots, 5, \quad m \neq 2, \\
J_{1,V} \cdot \eta_V &= \gamma_{1,V}(C_{1,V}^{\text{out}} - C_{1,V}) \quad \text{on } \Gamma_{\text{lumen}} \times \mathbb{R}^+, \\
J_{m,V} \cdot \eta_V &= -\gamma_{m,V} C_{m,V} \quad \text{on } \Gamma_{\text{lumen}} \times \mathbb{R}^+, \quad m = 3, 4, 5, \\
J_{m,V} \cdot \eta_V &= 0 \quad \text{on } (\Gamma_{\text{wall}} \cup \Gamma_{\text{adv}}) \times \mathbb{R}^+, \quad m = 1, \ldots, 5, \quad m \neq 2.
\end{align*}
\]  
(14)

3 Numerical Experiments

We solve problem (4), (13) and (14) with the finite element method. We fix \( h > 0 \) and define in \( \Omega = S \cup V \) (Figure 1) an admissible triangulation \( T_h \), depending on \( h > 0 \), such that the corresponding admissible triangulations in \( S \) and \( V \), respectively \( T_{hS} \) and \( T_{hV} \), are compatible in \( \Gamma_{\text{coat}} \) (see the zoomed part of Figure 2). We represent by \( \Delta_1 \) a typical element of \( T_{hS} \) and by \( \Delta_2 \) a typical element of \( T_{hV} \).

Figure 2: Triangulations in the stent and in the arterial wall.
Several choices of finite element spaces can be made, but we use here the piecewise linear finite element space \( P_1 \) to approximate the concentration of drug.

All numerical experiments have been done with the open source PDE solver freeFEM++ considering the triangulation plotted in Figure 2 with 3688 elements (1968 vertices) for the arterial wall and 100 elements (83 vertices) for each stent and using an implicit-explicit backward integrator with time step size \( \Delta t = 10^{-3} \).

We define the mass in the coated stent and in the arterial wall by

\[
\mathcal{M}_{m,S,h}(t_n) = \int_{S_h} C_{m,S,h}(t_n) dS, \; m = 1, \ldots, 5,
\]

\[
\mathcal{M}_{m,V,h}(t_n) = \int_{V_h} C_{m,V,h}(t_n) dV, \; m = 1, \ldots, 7, \; m \neq 2,
\]

respectively, where \( \mathcal{M}_{m,j,h}(t_n), \; j = S, V, \) are the numerical approximation for \( \mathcal{M}_{m,j}(t_n), \; j = S, V, \) respectively at time interval.

The following values for the parameters have been considered in the numerical simulations ([6, 9]):

\[
\kappa_{1,S} = \kappa_{2,V} = 1 \times 10^{-6} \text{ cm}^2\text{g}^{-1}\text{s}^{-1}, \; \kappa_{2,S} = 1 \times 10^{-7} \text{ cm}^2\text{g}^{-1}\text{s}^{-1}, \; \gamma_{m,V} = 1 \times 10^{10} \text{ cm.s}^{-1}, \; D_{1,S}^0 = 1 \times 10^{-8} \text{ cm}^2\text{s}^{-1}, \; D_{2,S}^0 = 1 \times 10^{-15} \text{ cm}^2\text{s}^{-1}, \; D_{3,S}^0 = 1 \times 10^{-10} \text{ cm}^2\text{s}^{-1}, \; D_{4,S}^0 = 2 \times 10^{-10} \text{ cm}^2\text{s}^{-1}, \; D_{5,S}^0 = 1 \times 10^{-8} \text{ cm}^2\text{s}^{-1}, \; k_S = 2 \times 10^{-14} \text{ cm}^2, \; k_V = 1 \times 10^{-15} \text{ cm}^2, \; \mu_S = 0.72 \times 10^{-2} \text{ g.cm}^{-1}\text{s}^{-1}, \; \mu_V = 0.5 \times 10^{-2} \text{ g.cm}^{-1}\text{s}^{-1}, \; D_{1,V} = 1 \times 10^{-8} \text{ cm}^2\text{s}^{-1}, \; D_{3,V} = 1 \times 10^{-10} \text{ cm}^2\text{s}^{-1}, \; D_{4,V} = 2 \times 10^{-10} \text{ cm}^2\text{s}^{-1}, \; \alpha = 1 \text{ s.cm}^{-2} - \; \beta = \gamma = 10 \text{ s.cm}^{-2}, \; p_{\text{lumen}} = 100 \text{ mmHg} \; \text{and} \; p_{\text{adv}} = 0 \text{ mmHg}, \; C_{0,6,V}^0 = 1 \times 10^{-5} \text{ mol}.
\]

Parameters in Table 1 for a hydrophilic drug (heparin) and a hydrophobic drug (paclitaxel) have been considered in our numerical simulations ([1, 7, 8]).

<table>
<thead>
<tr>
<th>Drug</th>
<th>( D_{5,S}^0 ) [cm(^2)s(^{-1})]</th>
<th>( D_{5,V} ) [cm(^2)s(^{-1})]</th>
<th>( k_{b,V} ) [Mol(^{-1})s(^{-1})]</th>
<th>( k_{u,V} ) [s(^{-1})]</th>
<th>( K_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heparin</td>
<td>( 1 \times 10^{-10} )</td>
<td>( 7.7 \times 10^{-8} )</td>
<td>( 9.2 \times 10^4 )</td>
<td>( 15 \times 10^{-3} )</td>
<td>60</td>
</tr>
<tr>
<td>Paclitaxel</td>
<td>( 5.7 \times 10^{-9} )</td>
<td>( 2.6 \times 10^{-8} )</td>
<td>( 3.6 \times 10^6 )</td>
<td>( 9 \times 10^{-2} )</td>
<td>400</td>
</tr>
</tbody>
</table>

Table 1: Properties of heparin and paclitaxel.

Figure 3 illustrates the evolution of heparin with and without binding. We observe that the concentration of drug in the arterial wall with binding is higher than without binding. As it is well known, the existence of binding sites increases the residence time of the drug in the arterial wall.
Hydrophilic drugs like heparin are known to be rapidly cleared and ineffective. Nowadays they have been discarded from clinical use in favour of the more persistent hydrophobic drugs such as paclitaxel, sirolimus and everolimus ([1]). Taxus™ paclitaxel eluting stent from Boston Scientific, Natick, MA, USA, applies paclitaxel, a fairly hyrophobic drug ($K_b = 400$), as therapeutic agent to control migration of smooth muscle cells from endothelium caused by in-stent restenosis. Heparin, a hydrophilic drug ($K_b = 60$), is used in Carmeda BioActive Surface (CBAS) heparin coating made by Carmeda, Upplands Vasby, Stockholm, Sweden.

Distribution of two different drugs, heparin and paclitaxel, released from drug eluting
stents in the arterial wall are compared in Figure 4. We observe that residence time of paclitaxel is higher than the residence time of the heparin. This means that heparin leaves the arterial wall faster than the paclitaxel.

Figure 4: Concentration of heparin (top left) and paclitaxel (top right) in the arterial wall in day 30 and the evolution of the mass of drugs (bottom).

4 Conclusions

In this paper we updated the model presented in [2] by including binding of drug to immobilized sites in the arterial wall. From the modeling and the numerical viewpoints the effect of reversible binding sites in the evolution of different drugs is taken into account.
We show that the hydrophobic drug paclitaxel stays longer in the arterial wall than the hydrophilic drug heparin. These observations can give manufacturers useful guidelines in the production of DES.

Acknowledgements

This work was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0324/2013 and also FCT-Grant SFRH/BD/51167/2010.

References


The decay of solutions of a wave equation with memory

J. A. Ferreira\textsuperscript{1}, Paula de Oliveira\textsuperscript{1} and G. Pena\textsuperscript{1}

\textsuperscript{1} CMUC, Department of Mathematics, University of Coimbra

emails: ferreira@mat.uc.pt, poliveir@mat.uc.pt, gpena@mat.uc.pt

Abstract

In this paper the qualitative behaviour of a damped linear wave equation with memory is studied. A new energy functional is introduced which generalizes some of the most common energy functionals found in the literature for this type of problem. We establish an upper bound for this functional that, under suitable regularity conditions, converges to zero as time increases. Numerical waves that mimic their continuous counterpart are also computed using the finite element approach.

Key words: wave equation, memory, viscoelasticity, energy estimate

1 Introduction

In this paper we consider the following damped wave equation with memory

\[
\frac{d^2u}{dt^2}(t) + c \frac{du}{dt}(t) - D_1 \Delta u(t) = -D_2 \int_0^t K_{er}(t-s) \Delta u(s) ds + f(t), \quad t \in \mathbb{R}^+, \tag{1}
\]

where

\[ K_{er}(s) = \frac{1}{\tau} e^{-s/\tau}, \quad \tau > 0, \]

and \( u(t) \) denotes a function defined from \( \Omega \subset \mathbb{R}^n \) into \( \mathbb{R} \), \( c \) is a function depending only on spatial variables and that accounts for the damping of the wave, \( D_1, D_2 \) and \( \tau \) are positive constants and \( f \) denotes a source term.

Equation (1) can be used to model the displacement of a viscoelastic material under the action of an external force when the stress tensor \( \sigma(t) \) and the strain tensor \( \epsilon(t) \) are related by the following constitutive equation

\[
\sigma(t) = E(0)D\epsilon(t) - \int_0^t \frac{\partial}{\partial s} E(t-s)D\epsilon(s) ds, \tag{2}
\]
where $D$ is an elastic tensor and the stress relaxation function, $E$, is nonnegative and monotone decreasing. Assuming that the viscoelastic behaviour is described by Maxwell-Wiechert model (with only one Maxwell arm), then $E(t) = E_0 + E_1 e^{-\alpha_1 t}$, where $E_0$ is the Young modulus of the spring arm, $E_1$ is the Young modulus of the Maxwell arm and $\alpha_1 = \frac{E_1}{\mu_1^2}$ being $\mu_1$ the associated viscosity. The relation between the displacement $u$, the stress $\sigma$ and the external force $f$ is given by Newton’s second law

$$\rho \frac{d^2 u}{dt^2} = \nabla \cdot \sigma + f,$$

where $\rho$ is the mass density of the body and

$$\nabla \cdot \sigma = \left[ \sum_{j=1}^{n} \frac{\partial \sigma_{ij}}{\partial x_j} \right]_{i,j=1}^{n}.$$ 

Assuming that the relation between the strain and the displacement is given by

$$\epsilon(t) = \frac{1}{2} \left( \nabla u(t) + \nabla u(t)^T \right),$$

we obtain from (3) the following second order integro-differential equation

$$\rho \frac{d^2 u}{dt^2}(t) - D_1 \Delta u(t) = -D_2 \int_{0}^{t} K_{er}(t-s) \Delta u(s) \, ds \, ds + f,$$

with $D_1 = D(E(0) + E_1), D_2 = \frac{E_1}{\mu_1^2}$ and $\tau = \alpha_1^{-1}$.

In what follows we consider homogeneous Dirichlet boundary conditions and the following initial conditions

$$u(0) = u_0, \quad \frac{du}{dt}(0) = u_1.$$ 

A quasilinear problem of type of (1) was also introduced for instance in [5], [9] and [11] to describe a viscoelasticity physical problem. Without being exhaustive, we mention [1], [2], [7], [8], [10] and [12] for the study of qualitative properties of partial differential problems defined by equations of type of (1).

The initial boundary value problem (IBVP) (1), (5) with homogeneous Dirichlet boundary conditions is now replaced by its weak formulation. To define such formulation we introduce the functional context needed. Let $L^2(\Omega), L^\infty(\Omega)$ and $H^1_0(\Omega)$ be the usual Sobolev spaces. In $L^2(\Omega)$ we consider the usual inner product $(\cdot, \cdot)$ and the norm induced by this inner product is denoted by $\|\cdot\|_0$. In $H^1_0(\Omega)$ we consider the usual norm $\|\cdot\|_1$. Let $L^2(\mathbb{R}^+; H^1_0(\Omega))$ be the space of functions $v : \mathbb{R}^+ \to H^1_0(\Omega)$ such that $\int_{0}^{T} \|v(t)\|^2_1 \, dt < \infty$, $\forall T > 0$. Let $H^1(\mathbb{R}^+; H^1_0(\Omega))$ be the subspace of $L^2(\mathbb{R}^+; H^1_0(\Omega))$ of all functions $v$ such
that its weak derivative \( \frac{dv}{dt} : \mathbb{R}^+ \rightarrow H^1_0(\Omega) \) belongs to \( L^2(0, \infty; H^1_0(\Omega)) \). By \( L^\infty(\mathbb{R}^+; L^2(\Omega)) \), we represent the space of all functions \( v : \mathbb{R}^+ \rightarrow L^2(\Omega) \) such that

\[
\text{ess sup}_{t \in [0,T]} \| v(t) \|_0 < \infty, \forall T > 0.
\]

Let \( V = L^2(\Omega) \) or \( V = H^1_0(\Omega) \). By \( C^m(\mathbb{R}^+; V) \), \( m \in \mathbb{N} \), we represent the space of functions \( v : \mathbb{R}_0^+ \rightarrow V \) with continuous derivatives \( \frac{d^j v}{dt^j} : \mathbb{R}_0^+ \rightarrow V \), for \( j = 0, \ldots, m \).

The weak solution of the problem (1), (5) with homogeneous Dirichlet boundary conditions is the solution of the variational problem:

Find \( u \in H^1(\mathbb{R}^+; H^1_0(\Omega)) \) with \( \frac{d^2 u}{dt^2} \in L^\infty(\mathbb{R}^+; L^2(\Omega)) \) and such that, for all \( T > 0 \), holds the following

\[
\begin{aligned}
&\left( \frac{d^2 u}{dt^2}(t) + c \frac{du}{dt}(t), w \right) + D_1(\nabla u(t), \nabla w) = D_2 \int_0^t K_{er}(t-s) (\nabla u(s), \nabla w) \, ds + (f(t), w), \\
&\text{a. e. in } (0,T), \forall w \in H^1_0(\Omega),
\end{aligned}
\]

\( \lim_{t \to \infty} \frac{du}{dt} = 0 \text{ in } L^2(\Omega), \)

\( \lim_{t \to \infty} u(t) = 0 \text{ in } H^1(\Omega), \)

\( \lim_{t \to \infty} \int_0^t K_{er}(t-s) \nabla u(s) \, ds = 0 \text{ in } L^2(\Omega). \)

In (6) the inner products in \( L^2(\Omega) \) and \( (L^2(\Omega))^n \) are denoted indifferently by \( (\cdot, \cdot) \). Their norms will be also represented indifferently by \( \| \cdot \|_0 \).

The main objective of this paper is the analysis of an energy functional under general assumptions and the illustration of the qualitative behaviour of numerical solutions for several different choices of the parameters. We shall prove that a suitable energy functional converges to zero as \( t \to \infty \). Numerical wave equations that mimic their continuous counterpart will be also considered and their behaviour will be explored.

The paper is organized as follows. In Section 2 we introduce the new energy functional and we prove that under convenient assumptions we have

A finite element method is introduced in Section 3 that mimics the energy behaviour of the IBVP studied in this paper. The behaviour of the IBVP (1), (5) with homogeneous Dirichlet boundary conditions is studied in Section 4. In Section 5 we summarize some conclusions.
2 Energy behaviour

The energy functional that we introduce here extends several definitions presented before in the literature. For instance, in [4] and [16] the authors considered the classical energy functional
\[ E(u)(t) = \|u(t)\|_0 + \|\nabla u(t)\|_0^2, \]
while in [10] a term induced by the boundary conditions was added to the last energy functional. In [3], for a quasilinear problem, a term related with the reaction term was added. In [12] the energy functional
\[ E(u)(t) = \frac{1}{2} \left\| \frac{du}{dt}(t) \right\|_0^2 + \frac{1}{2} \left( 1 - \int_0^t K_{er}(t-s) ds \right) \|\nabla u(t)\|_0^2 \]
\[ + \int_0^t K_{er}(t-s) \|\nabla u(t) - \nabla u(s)\|_0^2 ds, \] (7)
was introduced. A similar definition was considered in [16] but with the last term of (7) replaced by
\[ \int_0^t K_{er}(t-s) \|u(t) - u(s)\|_0^2 ds. \]

We establish in what follows an estimate for the energy functional
\[ E(u)(t) = \left\| \frac{du}{dt}(t) \right\|_0^2 + \|u(t)\|_1^2 + \left\| \int_0^t K_{er}(t-s) \nabla u(s) ds - \nabla u(t) \right\|_0^2 \]
\[ + \left\| \int_0^t K_{er}(t-s) \nabla u(s) ds \right\|_0^2, \] (8)
for \( t > 0 \), where \( u \) is a solution of (6). Under suitable regularity conditions and using the energy method, it is straightforward to show the following result.

**Theorem 1.** Let \( u \in C^2(0, \infty, L^2(\Omega)) \cap C^1(0, \infty, H^1_0(\Omega)) \) be a solution of (6) for \( D_1 > D_2 \), \( K_{er}(s) = Ke^{-\beta s}, c \in L^\infty(\Omega) \) satisfying
\[ c \geq c_0 > 0 \text{ on } \bar{\Omega}, \] (9)
and \( f = 0 \). If there exists a positive constant \( \gamma \) such that \( \gamma > \min \{\|c\|_\infty, \beta + K\} \) and \( C_1 - 2\gamma < 0 \) where
\[ C_1 = \frac{\max \left\{ 2(2\gamma - c_0), 2D_2 \frac{(\gamma - \beta)^2}{K} \right\}}{\min \left\{ 1, \gamma^2 - \gamma \|c\|_\infty, D_1 - D_2, D_2, D_2 \frac{\gamma - \beta - K}{K} \right\}}, \] (10)
then there exists a constant $C_2 > 0$, independent of $t$, such that

$$E(u)(t) + \int_0^t e^{-2\gamma(t-s)} \|\nabla u(s)\|_0^2 \, ds \leq C_2 e^{(C_1 - 2\gamma)t} \left( \|u_1 + \gamma u_0\|_0^2 + \gamma^2 - \gamma c_0 \right)$$

and consequently,

$$\lim_{t \to \infty} E(u)(t) = 0.$$  \hfill (12)

3 Decay decreasing of numerical waves

The study of numerical methods to solve numerically the IBVP (1), (5) with homogeneous Dirichlet boundary conditions was presented for instance in [6], [14], [13], [15] and some of the references in these papers.

In this section we show that the numerical solutions of the IBVP (1), (5) with homogeneous Dirichlet boundary conditions present the same qualitative behaviour as the continuous solutions. Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain and let $h > 0$ be a fixed parameter and let $T_h$ be an admissible triangulation of $\Omega$ with diameter $h$, that is,

$$h = \max_{\Delta \in T_h} \text{diam}(\Delta),$$

where $\text{diam}(\Delta)$ denotes the diameter of $\Delta$. Let $V_h$ be the space of piecewise polynomials of degree $m$ defined in $T_h$, that is,

$$V_h = \{ v \in C^0(\overline{\Omega}) : v = 0 \text{ on } \partial \Omega, v = p_m \text{ in } \Delta, \Delta \in T_h \},$$

where $p_m$ denotes a polynomial of degree at most $m$. By $P_{\partial \Omega}$ and $P_{\Omega}$ we represent the set of nodes of $T_h$ on $\partial \Omega$ and $\Omega$, respectively. Let $\{ \phi_P, P \in P_{\Omega} \}$ be a basis of $V_h$. The finite element approximation for the solution of the IBVP (1), (5) with homogeneous Dirichlet boundary conditions is $u_h(x, t) = \sum_{P \in P_{\Omega}} \alpha_P(t) \phi_P(x)$ that satisfies the following problem

$$\begin{align*}
\frac{d^2 u_h}{dt^2}(t) + c \frac{du_h}{dt}(t) + (\nabla u_h(t), \nabla w_h) = D_2 \int_0^t K_{er}(t-s)(\nabla u_h(s), \nabla w_h) \, ds + (f(t), w_h), \\
\frac{du_h}{dt}(0) = u_{1,h}, \\
u_h(0) = u_{0,h}.
\end{align*}$$

(13)
In (13) \( u_{1,h} \) and \( u_{0,h} \) are approximations of \( u_1 \) and \( u_0 \) in \( V_h \). To compute \( u_h(t) \) we solve the following second order system of integro-differential equations

\[
\begin{align*}
M_h \alpha''(t) + C_h \alpha'(t) + A_h \alpha(t) &= \int_0^t K_{er}(t - s)B_h \alpha(s) \, ds + F_h(t), \quad t > 0, \\
\alpha'(0) &= U_{1,h}, \\
\alpha(0) &= U_{0,h},
\end{align*}
\]

(14)

where \( \alpha(t) = [(\alpha_P(t))_{P \in \mathcal{P}_\Omega}], \ U_{i,h}, i = 0, 1, \) are the vectors whose components are the coordinates of \( u_{i,h}, i = 0, 1, \) with respect to the basis \( \{\phi_P, P \in \mathcal{P}_\Omega\} \), and

\[
\begin{align*}
M_h &= [(\phi_P, \phi_Q)_{P,Q \in \mathcal{P}_\Omega}], \\
C_h &= [(\nabla \phi_P, \nabla \phi_Q)_{P,Q \in \mathcal{P}_\Omega}], \\
A_h &= [(\nabla \phi_P, \nabla \phi_Q)_{P,Q \in \mathcal{P}_\Omega}], \\
B_h &= [D_2(\nabla \phi_P, \nabla \phi_Q)_{P,Q \in \mathcal{P}_\Omega}], \\
F_h(t) &= [(f(t), \phi_Q)_{Q \in \mathcal{P}_\Omega}].
\end{align*}
\]

Introducing the new variable \( Z(t) = (z_1(t), z_2(t)) \) where \( z_1(t) = \alpha(t), z_2(t) = \alpha'(t) \), then the initial value problem (14) of second order is equivalent to

\[
\begin{align*}
Z'(t) &= A_h Z(t) + \int_0^t K_{er} (t - s) B_h Z(s) \, ds + F_h(t), \quad t > 0, \\
Z(0) &= U_h,
\end{align*}
\]

(15)

where

\[
A_h = \begin{bmatrix}
0 & I \\
-M_h^{-1}A_h & -M_h^{-1}C_h
\end{bmatrix}, \\
B_h = \begin{bmatrix}
M_h^{-1}B_h & 0 \\
0 & 0
\end{bmatrix}, \\
F_h(t) = \begin{bmatrix}
0 \\
M_h^{-1}F_h
\end{bmatrix}, \\
U_h = \begin{bmatrix}
U_{0,h} \\
U_{1,h}
\end{bmatrix}.
\]

As the unique solution of the IVP (15) is smooth enough, then for the unique solution \( u_h(t) \in V_h \) of (13) it can be shown the following semi-discrete version of Theorem 1.

**Theorem 2.** Under the assumptions of Theorem 1 there exists a constant \( C_2 > 0 \), independent of \( t \), such that

\[
E(u_h)(t) + \int_0^t e^{-2\gamma(t-s)} \|\nabla u_h(s)\|^2_0 \, ds \leq C_2 e^{(C_1-2\gamma)t} \left( \|u_{1,h} + \gamma u_{0,h}\|^2_0 + (\gamma^2 - \gamma c_0) \|u_{0,h}\|^2_0 + D_1 \|\nabla u_{0,h}\|^2_0 \right)
\]

(16)

and consequently,

\[
\lim_{t \to \infty} E(u_h)(t) = 0.
\]

(17)
4 Numerical results

In this section we illustrate the qualitative behaviour of the numerical solutions of (14). Given the weak formulation (6), we specify the domain \( \Omega = (-1, 1)^2 \) and the initial data

\[
\begin{align*}
  u(x, y, 0) &= e^{-\frac{x^2+y^2}{2}}, \\
  \frac{du}{dt}(x, y, 0) &= 0, \quad \text{for} \ (x, y) \in \Omega.
\end{align*}
\]

Following the spatial discretisation in (13), we introduce the time step \( \Delta t \) and a uniform partition \( t_j = j\Delta t, j = 0, 1, 2, \ldots, N = \left[ \frac{T}{\Delta t} \right] \). Applying standard centered finite differences schemes in time and the composite trapezoidal rule to the formulation (14), the following method is obtained:

\[
\begin{align*}
  \left( \frac{u_{h}^{n+1} - 2u_{h}^{n} + u_{h}^{n-1}}{\Delta t^2}, v \right) + c \left( \frac{u_{h}^{n+1} - u_{h}^{n-1}}{2\Delta t}, v \right) + D_1 \left( \nabla u_{h}^{n+1}, \nabla v \right) = \\
  = D_2 \Delta t \sum_{j=0}^{n} \left( e^{-\frac{t_{n+1}-t_{j+1}}{\tau}} \nabla u_{h}^{j+1} + e^{-\frac{t_{n+1}-t_{j}}{\tau}} \nabla u_{h}^{j} \right),
\end{align*}
\]

where \( u_{h}^{j} \) is an approximation for \( u(t_j), j = 0, 1, \ldots, N \).

Let \( I_{n+1} = \frac{D_2 \Delta t}{2\tau} \sum_{j=0}^{n} \left( e^{-\frac{t_{n+1}-t_{j+1}}{\tau}} \nabla u_{h}^{j+1} + e^{-\frac{t_{n+1}-t_{j}}{\tau}} \nabla u_{h}^{j} \right) \). Then \( I_n \) satisfies

\[
\begin{align*}
  \begin{cases}
    I_{n+1} &= e^{-\frac{\Delta t}{\tau}} I_n + \frac{D_2 \Delta t}{2\tau} \left( e^{-\frac{\Delta t}{\tau}} \nabla u_0^{n} + \nabla u_0^{n+1} \right), \quad n > 1 \\
    I_1 &= \frac{D_2 \Delta t}{2\tau} \left( e^{-\frac{\Delta t}{\tau}} \nabla u_0^{0} + \nabla u_0^{1} \right).
  \end{cases}
\end{align*}
\]

With this new notation, method (18) can be rewritten as

\[
\begin{align*}
  \left( \frac{1}{\Delta t^2} + \frac{c}{2\Delta t} \right) u_{h}^{n+1}, v \right) + \left( D_1 - \frac{D_2 \Delta t}{2\tau} \right) \left( \nabla u_{h}^{n+1}, \nabla v \right) = \\
  = \left( \frac{2}{\Delta t^2} u_{h}^{n} + \left( \frac{c}{2\Delta t} - \frac{1}{\Delta t^2} \right) u_{h}^{n-1}, v \right) + e^{-\frac{\Delta t}{\tau}} \left( I_n, \nabla v \right).
\end{align*}
\]

Remark 1. The integral term in (6), discretized in (18), should be implemented following (20).

4.1 Solution’s behaviour with damping

Let the fully discretisation of \( E(u)(t) \) (8) be defined by

\[
E_{h,n} = \left\| \frac{u_{h}^{n} - u_{h}^{n-2}}{2\Delta t} \right\|^2_0 + \| u_{h}^{n} \|^2_1 + \| I_n - \nabla u_{h}^{n} \|^2_0, \quad n \geq 2.
\]
The decay of solutions of a wave equation with memory

The behaviour of \(E_{h,n}\) is clearly illustrated in Figure 1, for different values of \(D_2\) and \(\tau\). It can be observed that the larger the damping factor \(c\) is, the faster the energy approximates zero.

![Figure 1: Plot of discrete energy for different damping factors and coefficients \(D_2, \tau\).](image)

A similar result is observed when analysing the numerical solution at the central point \((0,0)\) of the square \([-1,1]^2\). As expected from the previous results, the solution at this point approximates zero. In Figure 2 we plot the numerical solution at this point, for the same profiles as in Figure 1. It is observed that the smaller the value \(D_2\) is, the closer the solutions are, for different values of \(\tau\), to the solution of the limit case \(D_1 = 1\) and \(D_2 = 0\).
Figure 2: Discrete solution at point (0,0) for different damping factors and coefficients $D_2, \tau$.

4.2 Limiting case as $\tau$ tends to zero

For a fixed value of $D_2$, the variation of $\tau$ seems to induce a different time scale on the oscillations of the solutions (for smaller values of $D_2$ such difference is reduced due to the previous conclusions). To further investigate the behaviour of the numerical solution for varying $\tau$, we computed the restriction of the numerical solutions, at time $t = 4$, in the set $[-1, 1] \times \{0\}$. These results are plotted in Figure 3. Information from Figures 2a, 2c and 3 suggests that as $\tau$ approaches zero, the corresponding numerical solution approximates the numerical solution obtained taking the pure wave equation (with damping effect included) with wave coefficient $D_1 - D_2$. In fact, if we consider the differential term

$$I(t; \tau) := \frac{D_2}{\tau} \int_0^t e^{-\frac{t-s}{\tau}} \Delta u(s) \, ds$$
The decay of solutions of a wave equation with memory

Figure 3: Restriction of discrete solution at $[-1, 1] \times \{0\}$ for $c = 0.25$ and two choices of $D_2$.

it can be shown that for a sufficiently smooth function $u$ and fixed $t > 0$,

$$\lim_{\tau \to 0^+} I(t; \tau) = D_2 \Delta u(t)$$

which sheds some light into the observed behaviour. However, we do not have, at this point, a rigorous proof to analytically support this observation.

5 Conclusions

In this paper we establish bounds for an energy functional associated with the solution of a linear wave equation with a memory term. Under certain conditions, the energy converges to zero as $t \to \infty$. It is also shown that a semi-discrete counterpart of the equation (obtained by discretisation in space with finite elements) inherits the same property.

The numerical waves studied also exhibit the same convergence to zero of a discretized energy functional. It is moreover noticeable that as the coefficient $D_2$ approximates zero, the solutions approximate the solution of a pure wave equation (with no memory). Also, as $\tau$ approximates zero, a similar behaviour is observed.

Acknowledgements

This work was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0324/2013.
References


Non-Fickian tracer transport in porous media

José A. Ferreira¹ and Luís Pinto¹

¹ CMUC, Department of Mathematics, University of Coimbra, Portugal
emails: ferreira@mat.uc.pt, luisp@mat.uc.pt

Abstract

Diffusion processes have traditionally been modeled using the classical diffusion equation. However, as in the case of tracer transport in porous media, significant discrepancies between experimental results and numerical simulations have been reported in the literature. Therefore, in order to describe such anomalous behavior known as non-Fickian diffusion, some authors have replaced the parabolic model by continuous random walk models, which have been shown to be very effective. Integro-differential models have been also proposed to describe non-Fickian diffusion in porous media. The aim of this paper is to compare the ability of these classes of models to capture the dynamics of tracer transport in porous media.

Key words: tracer transport, porous media, non Fickian diffusion, CTRW, numerical simulation.

1 Introduction

Tracer transport in porous media is generally described by a diffusion equation that depends on the fluid velocity, which is linked to the pressure by the Darcy’s law (see [24], [9], [12], [23] and [28]). Such diffusion equation is established considering the mass conservation equation

\[ \frac{\partial u}{\partial t} + \nabla(uv) = \nabla J + q_2, \]

where \( u \) denotes the concentration, \( v \) the fluid velocity, \( J \) the mass flux and \( q_2 \) a reaction term. In (1) \( J \) is given by

\[ J = J_{dif} + J_{dis}, \]

where \( J_{dif} \) denotes the mass flux due to molecular diffusion.
where $D_m$ is the effective molecular diffusion tensor, and $J_{\text{dis}}$ satisfies the so called Fick’s law

$$J_{\text{dis}} = -D_d \nabla u,$$

and represents the dispersive mass flux associated with random deviations of fluid velocities within the porous space from their macroscopic value $v$. In the definition of $J_{\text{dis}}$, $D_d$ denotes the dispersive tensor usually given by

$$D_d = \alpha_l \|v\| I + (\alpha_l - \alpha_t) \frac{1}{\|v\|} vv^T,$$

where $\alpha_l$ and $\alpha_t$ are the longitudinal and transversal dispersivities.

Combining (1) with (2) we obtain the parabolic equation

$$\frac{\partial u}{\partial t} + \nabla.(uv) = \nabla.(D_m I + D_d \nabla u) + q_2,$$

where $I$ is the identity tensor. Equation (4) can be rewritten in the following equivalent form

$$\frac{\partial u}{\partial t} + \nabla.(uv) = \nabla.(D \nabla u) + q_2.$$  

The use of equation (5) to describe diffusion processes in porous media in general and tracer transport in particular have been shown inefficient. In fact the numerical results obtained do not accurately reproduce the laboratorial experiments. These facts are well reported in the literature and without be exhaustive we mention [2], [5], [6], [16], [18], [19], [20], [26], [27] and [30]. Several limitations of the parabolic equation to describe the concentration evolution have been pointed out (see, for instance, [4], [17], [21]): it prescribes an infinite speed of propagation for the concentration; it is based on Fick’s law for the mass flux which establishes a linear relation between the concentration and dispersive mass flux; the mass flux $J$ is independent of the history of dispersion; in the dispersive tensor the dispersive coefficients are medium constant and invariant with time and space (often they increase with the distance and/or with time). To circumvent the pathological behavior of the diffusion equation (5) several approaches were proposed in the literature. A summary of some of them is given in [21].

A widely adopted alternative is based on continuous time random walks (CTRW) (see for instance [2], [5], [6], [8], [15], [18], [19]). Let us consider that each particle in diffusion performs jumps or transitions characterized by waiting times between jumps. The jumps and the waiting time are coupled by a joint probability density function (pdf) $\psi$ that describes at $(x,t)$ the jump at position $x$ ant time $t$. When the jumps occur in $\mathbb{R}^n$, the pdf $\psi$ and $u$ are linked by the following generalized master equations

$$\frac{\partial u}{\partial t}(x,t) = \int_0^t \frac{1}{t_1} \left( \int_{\mathbb{R}^n} M(x-z,t-\sigma) u(z,\sigma) dz - \int_{\mathbb{R}^n} M(z-x,t-\sigma) u(x,\sigma) dz \right) d\sigma.$$  

In (6) $t_1$ represents a median transition time, $M(x,t)$ is defined by its Laplace transform $\tilde{M}(x,s)$ which is given by

$$\tilde{M}(x,s) = \frac{s t_1 \tilde{\psi}(x,s)}{1 - \tilde{\psi}(u)}.$$  

In (7) the following notations were used: \( \tilde{\psi}(x,s) \) represents the Laplace transform of the pdf \( \psi(x,t) \) and \( \tilde{\psi}(s) = \int_{\mathbb{R}^n} \tilde{\psi}(x,s) \, dx \). If we assume that the pdf \( \psi \) is such that its Laplace transform can be decoupled in the following form \( \tilde{\psi}(x,s) = p(x) \tilde{\phi}(s) \) where \( p(x) \) denotes the transition length pdf and \( \phi \) denotes the marginal density of \( \psi \), then \( \tilde{M} \) admits also the decoupling \( \tilde{M}(x,s) = p(x)\tilde{M}(s) \), where

\[
\tilde{M}(s) = \frac{st_1 \tilde{\phi}(s)}{1 - \tilde{\phi}(s)}.
\]

From (6), we have in the Laplace space the following generalized mater equation

\[
s\tilde{u}(x,s) - u_0(x) = \frac{1}{t_1} \tilde{M}(s) \left( \int_{\mathbb{R}^n} p(x-z)\tilde{u}(z,s) \, dz - \int_{\mathbb{R}^n} p(z-x)\tilde{u}(x,s) \, dz \right),
\]

which is equivalent to

\[
s\tilde{u}(x,s) - u_0(x) = \frac{1}{t_1} \tilde{M}(s) \sum_{|\alpha|} (-1)^{|\alpha|} \frac{1}{\alpha!} \frac{\partial^{\alpha} \tilde{u}}{\partial x^\alpha}(x,s),
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_n), \alpha \in \mathbb{N}_0, |\alpha| = \sum_{i=1}^n \alpha_i, \alpha! = \prod_{i=1}^n \alpha_i! \), \( \frac{\partial^{\alpha} \tilde{u}}{\partial x^\alpha} = \frac{\partial^{\alpha} \tilde{u}}{\partial x_{1}^{\alpha_1} \cdots x_{n}^{\alpha_n}} \) and

\[
m_{\alpha,p} = \int_{\mathbb{R}^n} p(z)z^\alpha \, dz.
\]

Neglecting in (8) the terms involving partial derivatives with order greater than two we obtain

\[
s\tilde{u}(x,s) - u_0(x) = -\tilde{M}(s) \left( v.\nabla \tilde{u}(x,s) - \nabla.(D\nabla \tilde{u})(x,s) \right),
\]

where \( v \) is the vector with components \( v_i = \frac{1}{t_1} \int_{\mathbb{R}^n} p(z)z_i \, dz \), and \( D \) is the matrix of order \( n \) defined by \( D_{ii} = \frac{1}{2t_1} \int_{\mathbb{R}^n} p(z)z_i^2 \, dz \) and \( D_{ij} = \frac{1}{t_1} \int_{\mathbb{R}^n} p(z)z_iz_j \, dz, i \neq j \). We remark that equation (9) can be also obtained with convenient modifications if the pdf \( \phi \) has a compact support in \( \mathbb{R}^n \).

We observe that the one dimension version of the equation (9) was used for instance in [5], [6], [8] and [15] to simulate tracer transport in porous media when \( p \) is a Gaussian pdf and \( \phi \) is the truncated power law pdf

\[
\phi(t) = \frac{(1 + t/t_1)^{-1-\beta}}{t_1 r^\beta \Gamma(-\beta, r)} \exp\left(-\frac{t_1 + t}{t_2}\right), \quad r = \frac{t_1}{t_2}, \frac{t_1}{t_2} < t_2, \quad 0 \leq \beta \leq 2,
\]
where $\Gamma$ is the incomplete Gamma function. In these papers, the integral terms of equation (6) were replaced by summations over $z$.

Another alternative that can be considered to avoid the pathological behavior of the classical diffusion equation (4) consist in the introduction of an hyperbolic or non-Fickian correction as been proposed, e.g., in [17], [21]. One possible approach is to assume that the dispersive mass flux $J_{dis}$ satisfies the following differential equation

$$\tau \frac{\partial J_{dis}(x,t)}{\partial t} + J_{dis}(x,t) = -D_d \nabla u(x,t),$$

where $\tau$ is a delay parameter. Note that the left hand side of (10) is a first order approximation of the left hand side of $J_{dis}(x, t + \tau) = -D_d \nabla u(x, t)$, which means that the dispersion mass flux at the point $x$ and time $t + \tau$ depends on the gradient of the concentration at the same point but at a delayed time. Equation (10) leads to

$$J_{dis}(t) = - \int_0^t K_{er}(t - s)D_d \nabla u(s) \, ds,$$

provided that $J_{dis}(0) = 0$. In (11) $K_{er}$ is given by $K_{er}(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$. Combining the partition (2), where $J_{dif}$ and $J_{dis}$ are given by (3) and (11), respectively, with (1) we obtain the integro-differential equation

$$\frac{\partial u}{\partial t} + \nabla.(uv) - \nabla.(D_m \nabla u) = \int_0^t K_{er}(t - s)\nabla.(D_d \nabla u)(s) \, ds$$

which replaces (5). The mathematical and numerical analysis of initial boundary value problems based on integro-differential equations of type (12) were studied for instance in [1], [3], [10], [11], [13], [14] and [25].

The objective of this paper is to illustrate the capacity of the three different classes of models introduced to simulate the experiments presented in [2] and [29]. These experiments are characterized by a non-Fickian behavior, i.e., deviation from the classical diffusion equation (5) which can be consequence of the existence of preferential flow paths that strongly influence both water and tracer transport ([22]). In Section 2 we describe the numerical procedures used in the numerical simulations presented in Section 3. In this last section such numerical simulations are compared with experimental results. In Section 4 we summarize some conclusions. We point out that to the best of our knowledge, this is the first work where the validation of mathematical models based on integro-differential equations of type (12) is considered.

## 2 BTCs

In this section, we test and validate the models by fitting breakthrough curves (BTCs) resulting from laboratory tracer tests. These curves describe the evolution of the tracer
concentration at specific point of the spatial domain where the tracer transport occurs. Experimentally BTCs show early arrival times and long later time tails (see for instance [2], [18] and [29]).

A common assumption in the validation of mathematical models in the context of tracer transport is to consider that the experiments can be described by a one dimension model, that is, if the diffusion process occurs in a cylinder, then the behavior of the tracer in a perpendicular disc to the axis of the cylinder is well described by its behavior at the center of such disc. Therefore, equations (5), (9) and (12) are defined in the domain \( (0, T] \times (0, L) \) and they are completed with the initial condition \( u(x, t) = u_0(x), x \in (0, L) \), and boundary conditions of the following type

\[
    u(0, t) = u_I(t), u(L, t) = 0, t \in (0, T],
\]

where the first condition defines the injected fluid while the last one means that all the fluid that attains the end of the spatial domain is removed. In \( [0, T] \times [0, L] \) we introduce a grid \( \{(x_i, t_n), i = 0, \ldots, N_x, n = 0, \ldots, N_t\} \), where \( x_0 = 0, x_{N_x} = L, x_i - x_{i-1} = h, t_0 = 0, t_{N_t} = T \) and \( t_n - t_{n-1} = \Delta t \).

The numerical curves are obtained fitting the BTCs curves using the the root mean square error (RMSE) defined by

\[
    RMSE = \left( \frac{1}{N_t} \sum_{n=1}^{N_t} (u^n - u_h^n)^2 \right)^{1/2},
\]

where \( N_t \) is the number of measurements at the prescribed point and \( u_h^n \) is the corresponding approximation defined by a discretization of the classical diffusion equation (5) or (9) or (12).

We remark that equation (12) for a particular choice of \( \phi(t) \) leads to equation (5). In fact if \( \phi(t) = \lambda e^{-\lambda t} \) and \( t_1 = \lambda^{-1} \) then \( \hat{\phi}(s) = \frac{1}{1 + \lambda^{-1}s} \) and \( M(s) = 1 \). While the best-fit BTCs based on (5) and (12) are computed using the publicly available CTRW toolbox ([7]), the best-fit BTCs based on (12) are computed using a numerical scheme previous studied by the authors in [1], [13] and [14] which, for one dimension, can be written in the following form

\[
    D_{-t}u_h^{n+1} + vD_c(u_h^{n+1}) - D_mD_2u_h^{n+1} = D_d\Delta t \sum_{\ell=1}^{n} K_{er}(t_{n+1} - \ell t)D_2u_h^\ell,
\]

where \( D_{-t} \) denotes the usual backward finite difference operator in time, \( D_c \) and \( D_2 \) represent the second order centered finite difference operators which approximate the first and the second spatial derivatives, respectively. In this case the minimization of the RMSE is carried out using built-in routines of Matlab (version 7.9.0 (R2009b)). To avoid possible influence on the numerical solution, the outlet boundary condition is imposed far enough from the grid point used for the computation of the BTCs.
3 Validation

3.1 Data set 1

In an already classical experiment, Scheidegger [29] used homogeneous Berea sandstone core columns to investigate the accuracy of the classical diffusion model in simulating tracer transport in porous media. During the experiment, columns of different lengths were first fully saturated with tracer and subsequently flushed with clean liquid. The resulting tracer BTCs at the outflow boundary were measured and compared with the ones predicted by the classical diffusion model.

We test the introduced models using typical data from one of these experiments. In this case, the column was $7.62 \times 10^{-1}$ meters (m) long and $5.08 \times 10^{-2}$ m in diameter, the clean liquid was injected at a rate equal to $1.73$ cubic centimeters per minute (cm$^3$/min), and the porosity of the core was 0.204. This gives an average velocity of $4.18 \times 10^{-3}$ m/min. We consider $u_I = 1$, for $t \in (0, T]$. In Figure 1 (a) we show the experimental data and the best-fit curves obtained with the models (5) and (12), and the best-fit curves for models (12) and (9) are shown in Figure 1 (b).

![Figure 1](image)

Figure 1: Experimental data set 1 and the best-fit BTCs: (a) - models (5) (dash line) and (12)(solid line); (b)-models (12) (solid line) and (9) (dash line).

A quick observation of Figure 1 (a) indicates that the model (12) captures the transport dynamics quite well. However, the model (5) fails to describe the data, especially at later times, since it can not reproduce the long tail, a typical indication of non-Fickian transport.
This result for the classical diffusion model (5) is in line with the findings of Scheidegger. The values of the two models constants and the RMSE are listed in Table 1. As indicated by the given results, the application of the proposed model leads to a reduction of about 79% in the RMSE.

Now we compare the results obtained for the integro-differential model (12) with the results for the CTRW model (9). As shown in Figure 1 (b), there is only a small discrepancy between them. The RMSE values, $3.66 \times 10^{-3}$ and $5.11 \times 10^{-3}$, respectively, suggest that the integro-differential model fits the data slightly better. Here and in the following, we omitted the values of $t_1$ and $t_2$ in the model (9). They are consistently very small, and very large, respectively, when compared to the time scale.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>model (5)</th>
<th>model (9)</th>
<th>model (12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v$ (m/min)</td>
<td>$4.65 \times 10^{-5}$</td>
<td>$7.23 \times 10^{-5}$</td>
<td>$4.27 \times 10^{-5}$</td>
</tr>
<tr>
<td>$D$ (m²/min)</td>
<td>$1.35 \times 10^{-5}$</td>
<td>$7.86 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>$D_m$ (m²/min)</td>
<td>$1.08 \times 10^{-5}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_d$ (m²/min)</td>
<td>$1.76 \times 10^{-5}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$ (min)</td>
<td></td>
<td>$25.52$</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td>$1.58$</td>
</tr>
<tr>
<td>RMSE</td>
<td>$1.72 \times 10^{-2}$</td>
<td>$5.11 \times 10^{-3}$</td>
<td>$3.66 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 1: Fitting parameters for the results of Figure 1 and RMSE values.

### 3.2 Data set 2

The second group of data is the result of tracer displacement experiments through homogeneous sand columns reported in [2]. Next, we briefly describe the setup and we refer to that paper for all other experimental details. The columns were incrementally packed with sand particles of different sizes. The diameter of most of the sand particles lie in the range of $0.1 - 0.71$ millimeters. We consider the results for two columns: Column 1, 11 cm in diameter and 10 cm long; and Column 2, also 11 cm in diameter but 40 cm long. The transport experiment was conducted under initially unsaturated conditions, with the water content of 0.24 for Column 1 and of 0.18 for Column 2. A pulse tracer at the flow rate of $4.20 \times 10^{-2}$ cm/min was applied at the top of both columns within the time period of 140 seconds (s) for the smaller column and of 107 s for the longer one. The respective average velocities were $1.86 \times 10^{-3}$ cm/min and $2.28 \times 10^{-1}$ cm/min. After the pulse, water was injected at the same rate. To simulate this scenario, for Column 1 we set at the inlet boundary $u_I = 4.16 \times 10^{-2}$ for $t \leq t_I = 2.33$ min and $u_I = 0$ for $t > t_I$, and for Column 2 the inlet boundary is defined by $u_I = 5.77 \times 10^{-2}$ for $t \leq t_I = 1.78$ min and $u_I = 0$ for $t > t_I$. The observed and fitted BTCs for the models are plotted in Figure 2 for Column 1 and in Figure 3 for Column 2.

In particular, Figures 2 and 3 show that the agreement for the classical diffusion model
Figure 2: Experimental data set 2 and the best-fit BTCs for Column 1: (a) - model (5) (dash line) and model (12) (solid line); (b) - model (9) (dash line) and model (12) (solid line).

Figure 3: Experimental data set 2 and the best-fit BTCs for Column 2: (a) - model (5) (dash line) and model (12) (solid line); (b) - model (9) (dash line) and model (12) (solid line).
is very poor. This is especially true at later times, where the BTC possesses a heavy tail. On the other hand, the integro-differential model (12) captures the BTC behavior much better. The RMSEs presented in Tables 2 and 3 confirm this conclusion.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>model (5)</th>
<th>model (9)</th>
<th>model (12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v$ (m/min)</td>
<td>$1.88 \times 10^{-1}$</td>
<td>1.10</td>
<td>$1.69 \times 10^{-1}$</td>
</tr>
<tr>
<td>$D$ (m$^2$/min)</td>
<td>$3.69 \times 10^{-2}$</td>
<td>$2.92 \times 10^{-2}$</td>
<td></td>
</tr>
<tr>
<td>$D_m$ (m$^2$/min)</td>
<td></td>
<td>$2.59 \times 10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$D_d$ (m$^2$/min)</td>
<td></td>
<td></td>
<td>$10^{-4}$</td>
</tr>
<tr>
<td>$\tau$ (min)</td>
<td></td>
<td>15.95</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>$2.5 \times 10^{-4}$</td>
<td>$9.77 \times 10^{-4}$</td>
<td>$9.53 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 2: Fitting parameters for the models plotted in Figure 2 and RMSE values.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>model (5)</th>
<th>model (9)</th>
<th>model (12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v$ (m/min)</td>
<td>$1.90 \times 10^{-1}$</td>
<td>2.88</td>
<td>$1.74 \times 10^{-1}$</td>
</tr>
<tr>
<td>$D$ (m$^2$/min)</td>
<td>$2.37 \times 10^{-1}$</td>
<td>$1.88 \times 10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$D_m$ (m$^2$/min)</td>
<td></td>
<td>$1.16 \times 10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$D_d$ (m$^2$/min)</td>
<td></td>
<td></td>
<td>$4.25 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\tau$ (min)</td>
<td></td>
<td></td>
<td>42.14</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>$9.64 \times 10^{-1}$</td>
<td>$3.43 \times 10^{-1}$</td>
<td>$1.91 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 3: Fitting parameters for the models plotted in Figure 3 and RMSE values.

4 Conclusions

In this paper three different classes of diffusion models are compared considering tracer transport experimental data: the classical diffusion model (5), the CTRW model (9) and the integro-differential model (12). The classical diffusion model is established assuming Fick’s law for the mass flux while the CTRW model is established considering that the particles perform jumps and between consecutive jumps the particles wait for a random time. The randomness of the transport is described by a joint pdf for the waiting time and for the length of the jump. The integro-differential model is constructed by modifying Fick’s law considering that the mass flux at a certain time depends on the gradient of the concentration at a delayed time.

The numerical simulations show that the classical diffusion model does not capture the long tails of the BTCs at later times while the CTWR model and the integro-differential model are able to reproduce this non-Fickian behavior leading the last model to a greater reduction of RMSE.
Acknowledgements

This work was partially supported by the Centro de Matematica da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT under the project PEst-C/MAT/UI0324/2013 and project UTAustin/MAT/0066/2008.

References


Modeling and Simulating
Colonic Cell Renewal Disruption

Isabel N. Figueiredo¹, Carlos Leal¹ and Giuseppe Romanazzi¹

¹ CMUC, Dep. of Mathematics, University of Coimbra, Portugal
emails: isabelf@mat.uc.pt, carlosl@mat.uc.pt, roman@mat.uc.pt

Abstract

Colorectal cancer is believed to be initiated as a consequence of several genetic mutations in colonic cells. Colon epithelium is densely perforated by small holes (called crypts) wherein a strong cellular activity occurs. Aberrant Crypt Foci (ACF) are one of the first manifestations of the carcinogenesis process. In this paper we summarize different models, that have been recently proposed, for simulating and predicting the evolution of ACF. In all these models, it is assumed that the dynamics of colonic cells is characterized by a transport/diffusion equation coupled with an elliptic type equation, and involving some physiological parameters, as the birth and death rates of proliferative cells. First, a model for the cell dynamics in a single crypt is described. This model yields crypt fission in which a crypt follows a process of splitting due to an increase of the birth rate of proliferative cells. This reproduces a morphological pattern similar to that observed in medical images. The second model, here presented, is used to estimate the birth rate of proliferative cells along the colonic crypt wall. It is an inverse problem formulated as a PDE-constrained optimization model. Finally, by appropriately extending, to the entire colon, the above mentioned cell dynamics model, a periodic two-dimensional model is obtained for simulating the ACF appearance. To solve this last model two different approaches are applied and compared: one relies on heterogeneous multi-scale methods and the other on homogenization techniques.

Key words: Convection-diffusion, Cell Dynamics, Colonic Crypt, Fixed-Point Theorem, Homogenization, Multi-Scale Methods
MSC 2000: 76R99, 35J15, 35B27, 35R37, 47H10, 65M06, 65M50, 65M60

Colorectal cancer is one of the most common types of cancer that has a high rate of death, in the Western World [1]. It is generally accepted that colorectal cancer has its origin in
genetic mutations in colonic cells. Due to the large period of time elapsed between genetic mutations in colonic cells and the appearance of carcinoma (20-40 years according to [2]), it is possible to prevent the formation of carcinomas through an early detection of adenomas. The carcinogenesis process can be in fact stopped by the identification and removal of adenomas that otherwise could develop further into carcinomas. One of the first manifestations of colorectal cancer, detectable by conventional colonoscopy, is the appearance of the so-called Aberrant Crypt Foci (ACF)[3]. These are clusters of abnormal crypts (crypts are small cavities, in the colonic epithelium, containing the colonic cells) that present an abnormal morphology. ACF can be detected in colonoscopy, by the instillation of a dye. In fact, it is known that a few minutes after the instillation ACF become darker than normal crypts. There is no scientific agreement about the ACF morphogenesis. In the top-down theory, the appearance of abnormal cells occurs in the mucosa surface and afterwards they spread laterally and inside the crypts [4]. Instead, in the bottom-up theory, abnormal cells appear first in the bottom of the crypt, where they are prone to accumulate genetic alterations due to high birth rate, and after they migrate to the crypt orifice [5]. For a review of ACF and colorectal cancer medical analysis the reader can refer to [6, 7].

This paper presents a brief summary of our recent results, concerning the modeling and simulation of ACF morphogenesis (for a detailed explanation we refer to [8, 9, 10, 11]).

In [8] it is proposed a hybrid convection-diffusion-shape model, that simulates and predicts aberrant colonic crypt morphogenesis, similar to that observed in medical images. [12, 13, 5]. This model shows crypt fission, where a single crypt follows a subdivision process that starts at the bottom of the crypt, when there is an higher proliferative birth rate than that observed in normal crypts. The overall model couples the cell movement and proliferation equations with the crypt geometry. It relies on classical continuum transport/mass conservation laws and the changes in the crypt shape are driven by the pressure exerted by the cells on the crypt wall. This pressure is related to the cell velocity by a Darcy-type law. Numerical simulations are performed and comparisons with medical results are discussed.

In [9] a cell crypt based parameter estimation model is proposed. It is assumed that the cellular kinetics, occurring inside a single crypt, is governed by the convection-diffusion-shape model introduced in [8]. This latter involves some important physiologic parameters, as the birth and death rate of proliferative cells, for which only qualitative information is available in the literature. These parameters have a crucial role in ACF dynamics and evolution. By resolving an inverse problem (that minimizes the misfit between two proliferative cell densities: one observed and the other predicted by the model), we estimate the birth rate of proliferative cells (considered as a parameter field), along the colonic crypt wall. The location where the birth rate deviates from normal qualitative values can then be determined, and used for further clinical research and better understanding of the abnormal process leading to the increase or decrease of proliferative cells, and the emergence of ACF.

In [10] a multiscale model for aberrant crypt foci is presented. Starting from a three-
dimensional single crypt, we perform its projection into a plane and then build a model in which the colon is a two-dimensional structure with crypts periodically distributed therein. Inside each crypt, the dynamics of the abnormal cells is governed by a convective-diffusive model, whose unknowns are the cell density of abnormal cells and a pressure, as described in [8]. Outside the crypts, in the inter-cryptal region, a proliferative-diffusive model is assumed for the dynamics of abnormal cells. For the numerical implementation of the multiscale model, it is used a technique based on heterogeneous multiscale methods (see for example [14]). Two scales are employed: a macro-scale and a micro-scale. The macro-scale corresponds to the region of the colon where the evolution of ACF is taking place, whilst the micro-scale is related to the region occupied by each crypt and its local inter-cryptal region. Pressure and cell density are computed at the macro-scale level using the micro-scale structure in a neighborhood of the quadrature macro-scale points. This strategy reduces the computational cost of the simulations.

In [11] an homogenization model is described and used to represent the cellular dynamics in the colon epithelium, with the goal of simulating and predicting, *in silico*, the spread and evolution of ACF, as it can be observed in colonoscopy. By assuming that the colon is an heterogeneous media, exhibiting a periodic distribution of very small cavities (the crypts), the periodic model, introduced in [10], is adopted for describing the ACF cell-dynamics in a two-dimensional setting. Then, to this periodic model, homogenization techniques are applied, to find a simpler model, whose solution symbolizes the averaged behavior of ACF. Some theoretical results concerning the existence of solution of the homogenized model are proven, applying a fixed point theorem. Numerical results showing the convergence of the periodic model to the homogenized one are also presented.

**Acknowledgments**

This work was partially supported by the project PTDC/MATNAN/0593/2012 and also by CMUC and FCT (Portugal) through European program COMPETE/FEDER and project PEst-C/MAT/UI0324/2011.

**References**


Cooperative solution of cryptarithmetic problems

José F. Fontanari¹

¹ Instituto de Física de São Carlos, Universidade de São Paulo
emails: fontanari@ifsc.usp.br

Abstract

The presumption that a group of cooperating agents can solve a problem more efficiently than the same group of agents working in isolation is widespread, despite the little quantitative groundwork to support it. Here we use extensive agent-based simulations to investigate the performance of a group of cooperative agents in solving a cryptarithmetic problem. The agents perform individual trial-and-test operations to explore the search space and collaborate with each other in the sense that the best solution at a given trial is exposed in a black-board to which all agents have access in order to pick cues. We find a trade-off between the number of trial-and-test operations and the number of accesses to the black-board: too many accesses to the black-board result in a performance which is poorer than that exhibited by noncooperative agents. For the optimal balance between trial-and-test operations and black-board accesses we find a tenfold speedup of the mean time to find the correct solution with respect to the time taken by the noncooperative group.

Key words: constraint satisfaction, cryptarithmetic, social intelligence

1 Introduction

The advantages of cooperative work were perceived and explored by nature well before the advent of the human species as attested by the collective structures built by slime molds and social insects. These structures may be thought of as the organisms’ solutions to the problems that endanger their existence (see, e.g., [1]) and have inspired the proposal of optimization heuristics based on social interaction, such as the popular particle swarm optimization algorithm [2] and the lesser-known adaptive culture heuristic [3, 4]. Despite the success of the social interaction heuristics in producing near optimal solutions to a variety of combinatorial optimization problems, we know little about the mechanisms that
make cooperation efficient. (Efficiency here means that the effectiveness to solve a problem scales superlinearly with the number of agents or processors employed in the task.) The reason is probably that those social interaction heuristics as well as the problems they are set to solve are too complex for a first-principle analysis.

In this contribution we follow the research strategy set forth by Huberman in the 1990s that circumvents the above-mentioned difficulties by tackling easy combinatorial problems, by endowing the agents with simple random trial-and-test search tools, and by defining plain protocols of collaboration [5, 6]. In particular, we solve a constraint satisfaction cryptarithmetic problem by using a population of \( N \) agents which, in addition to the ability to perform individual trial-and-test searches, can learn from (or imitate) a model agent – the best performing agent at the trial – in accord with the principles of social learning [7]. Cooperation as well as the implicit existence of a central control are parts of our agent-based simulation because at each trial a model agent is chosen and its solution exposed in a black-board, which can be accessed by all agents who then can copy relevant pieces (cues) of the public solution [8].

We find that there is a trade-off between the number of trial-and-test operations and the number of accesses to the black-board: too few accesses lead to a performance slightly better than that of noncooperative agents with the effectiveness to solve the cryptarithmetic problem increasing linearly with the number of agents, whereas too many accesses to cues limit the regions of the search space that can be explored leading eventually to the failure to find the correct solution. Using \( N = 20 \) agents and the optimal balance between trial-and-test operations and black-board accesses results in a tenfold speedup of the mean time to find the correct solution. In addition, we find that the time to find the correct solution is distributed according to an exponential distribution of probability.

The rest of the paper is organized as follows. In section 2 we present the particular cryptarithmetic problem the agents must solve, explain how the solutions are encoded in strings and introduce the cost function associated to each solution. In section 3 we study the performance of a system of noncooperative agents and in section 4 the performance of the cooperative system. Finally, section 5 is reserved to our concluding remarks.

### 2 Cost landscape of cryptarithmetic problems

Cryptarithmetic problems such as

\[
DONALD + GERALD = ROBERT
\]

are constraint satisfaction problems in which the task is to find unique digit assignments to each of the letters so that the numbers represented by the words add up correctly [9]. In this particular cryptarithmetic problem, there are 10! different digit-to-letter assignments, of which only one is the solution to the problem, namely, \( A = 4, B = 3, D = 5, E = 9, G = \ldots \)
Cryptarithmetic problems in which letters form meaningful words are also termed alphametics [10] and were popularized in the 1930s by the Sphinx, a Belgian journal of recreational mathematics [9]. Of course, from the perspective of evaluating the performance of search heuristics on solving cryptarithmetic problems, the meaningfulness of the words are irrelevant, but in this contribution we will focus on the alphametic problem (1) only.

To devise efficient search heuristics for cryptarithmetic problems, it is necessary to assign a cost or energy to each possible digit-to-letter assignment [11]. For the alphametic problem (1) we encode a digit-to-letter assignment by the string $i = (i_1, i_2, \ldots, i_{10})$ where $i_n = 0, \ldots, 9$ represent the 10 digits and the subscripts $n = 1, \ldots, 10$ label the letters according to the convention $1 \rightarrow A$, $2 \rightarrow B$, $3 \rightarrow D$, $4 \rightarrow E$, $5 \rightarrow G$, $6 \rightarrow L$, $7 \rightarrow N$, $8 \rightarrow O$, $9 \rightarrow R$ and $10 \rightarrow T$. For example, the string $(0, 2, 9, 4, 8, 1, 7, 6, 3, 5)$ corresponds to the digit-to-letter assignment $A = 0, B = 2, D = 9, E = 4, G = 8, L = 1, N = 7, O = 6, R = 3, T = 5$. A somewhat natural way to associate a cost to a string $i$ is through the expression

$$C(i) = |R - (F + S)|$$

where $R$ is the result of the operation ($ROBERT$), $F$ is the first operand ($DONALD$) and $S$ is the second operand ($GERALD$). In our example we have $R = 362435$, $F = 967019$ and $S = 843019$ so that the cost associated to string $(0, 2, 9, 4, 8, 1, 7, 6, 3, 5)$ is $C = 1447603$. If the cost of a string is $C = 0$ then the digit-to-letter assignment coded by that string is the correct solution to the cryptarithmetic problem. We must note that the cost function defined in eq. (2) applies to all strings except those for which $i_3 = 0$ corresponding to the assignment $D = 0$, $i_5 = 0$ corresponding to the assignment $G = 0$ and $i_9 = 0$ corresponding to the assignment $R = 0$. Those are in principle invalid strings because they violate the rule of the cryptarithmetic puzzles that an integer number should not have the digit 0 in its leftmost position. Hence for those strings we assign an arbitrary large cost value, namely, $C = 10^7$, so that they now become valid strings but with the highest cost among all strings.

In addition to the assignment of the cost function (2) to each of the $10!$ strings that code the possible digit-to-letter mappings for the alphametic problem (1), the definition of a cost landscape for this problem requires the introduction of an elementary move between strings so that any two strings connected by that move can be considered neighbors. We define the elementary move as follows. Starting from a particular digit-to-letter mapping, say $(0, 2, 9, 4, 8, 1, 7, 6, 3, 5)$, first we select at random a digit $i$ from 0 to 9 and a letter label $n$ from 1 to 10. For the sake of concreteness, say we pick digit 0 and letter label 5, which corresponds to the assignment $G = 0$, according to the convention described before. To obtain a valid digit-to-letter mapping we need to reassign digit 8 to the letter label which was previously assigned to digit 0 so that the resulting mapping after the move is $(8, 2, 9, 4, 0, 1, 7, 6, 3, 5)$. Of course, we could obtain the same result by choosing two letter labels at random and then interchanging the digits assigned to them, but the interpretation
Cooperative solution of cryptarithmetic problems

Figure 1: (Color online) Right panel: Probability $\zeta$ that a collective search involving $N$ noncooperative agents terminates at trial $t$ as function of the ratio between the trial number $t$ and the size of the search space $10!$ for $N = 1(\circ), 5(\triangle), 10(\triangledown)$ and $20(\square)$. Left panel: Collapse of the data for different $N$ when the trial number $t$ and the probability $\zeta$ are further rescaled by $N$. The solid straight line is the exponential distribution $\zeta = a \exp(-at)$ with $a = 0.9N/10!$.

of imposing a particular assignment ($G = 0$ in our example) to an arbitrary string is more suitable for implementing the copying mechanism of the cooperative search strategy described in section 4. Clearly, the repeated application of the elementary move defined above is capable of producing any of the $10!$ strings starting from an arbitrary string of 10 different digits.

3 Noncooperative search strategy

We assume a system composed of $N$ agents or processors. Each agent operates successive elementary moves in an initial random string that encodes a digit-to-letter mapping. There is a different initial string for each agent and the agents operate in parallel. The collective search ends when one of the agents finds the zero-cost solution and we denote by $t$ the number of elementary moves or trials made by the agent that found the correct solution. Of course, $t$ stands also for the number of trials made by any one of the $N$ agents, since they operate in parallel and the search halts simultaneously for all agents. In the following we analyze the results of $10^5$ independent collective searches.

The main results of the noncooperative search strategy are summarized in figure 1, which plots the probability $\zeta$ that the collective search terminates at trial $t$. Since we expect that the typical number of trials necessary to find the correct solution scales with the size of the search space (i.e., $10!$), we use the ratio $t/10!$ for the abscissa and so the
ordinate has to be properly rescaled as well. As expected [5, 6], the results exhibited in
the left panel of figure 1 show a neat exponential distribution with the expected number of
trials (i.e., the inverse of the slope of the straight lines that fit the scattered data) decreasing
linearly with the number of agents $N$. This point can be proved by the collapse of the data
for different $N$ when the coordinate axes are properly rescaled as done in the right panel
of figure 1. In particular, for large $t$ the data is very well fitted by the exponential distribution
$\zeta = a \exp (-at)$ with $a = 0.9N/10!$ which is shown by the solid straight line in the right
panel.

There are two technical points regarding the results exhibited in figure 1 that deserve
some consideration. First, at least for $N = 1$ one would expect that the probability dis-
bution of halting times (i.e., number of trials) is given by the exponential distribution
$\zeta = (1/10!) \exp (-t/10!)$ for large $t$. (It is a geometric distribution for general $t$ [6]). In
fact, if we replace our elementary move by a more global move in which the entire string is
generated randomly at each trial then we find the expected exponential distribution. The
reason that our elementary move is slightly less efficient than the global move in exploring
the search space (i.e., for $N = 1$ the mean number of trials to success is $10!/0.9$ rather
than 10!) is because there is a probability of 2/90 that one reverses the change in the next
trial, whereas this probability is 1/10! for the global move. Second, one would expect the
improvement on the mean number of trials to success to be sublinear on the number of
agents $N$ employed on the search simply because agents duplicate each other's work. The
reason we see a neat linear dependence on $N$ is probably because the search space size 10!
is so large compared to $N$ that the work duplication effect is negligible. However, even in
this case one expects to see this effect in the region $t \gg 10!$ but the statistics in this region
is too poor to produce a reliable conclusion.

4 Cooperative search strategy

The setting here is similar to that of the noncooperative search strategy, except that at each
trial $t$ an agent has a certain probability $p \in [0, 1)$ of copying a digit-to-letter assignment
from a model string in the population rather than performing the elementary move. We
choose the model string as the lowest cost string in the population at a given trial $t$.
To illustrate the copying process let’s assume for the sake of concreteness that the string
to be updated is our already familiar example string (0, 2, 9, 4, 8, 1, 7, 6, 3, 5), whose cost
is $C = 1447603$, and that the model string is (5, 3, 9, 4, 8, 1, 6, 2, 7, 0) whose cost is $C = 
1050568$. The agent selects at random one of the distinct digit-to-letter assignments in
the model string. In our example the distinct assignments occur at the letter labels $n =
1, 2, 7, 8, 9, 10$. Let’s say the letter label $n = 1$, which corresponds to the assignment $A = 5$, is
chosen. Then we impose that this particular assignment be incorporated into the example
string, in a similar way as done for the elementary move. The resulting string becomes
Cooperative solution of cryptarithmetic problems

Figure 2: (Color online) Right panel: Probability $\zeta$ that a collective search involving $N$ cooperating agents terminates at trial $t$ as function of the ratio between the trial number $t$ and the size of the search space $10!$ for $N = 20$ and $p = 0(\bigcirc), 0.5(\triangle), 0.7(\blacktriangle)$ and $0.9(\square)$. Left panel: Mean number of trials per agent to find the correct solution as function of the probability $p$ that an agent will copy a digit-to-letter assignment from the model string (i.e., access the black-board) for $N = 20(\bullet), N = 5(\▲), N = 3(\blacktriangle)$ and $N = 2(\blacksquare)$. (5, 2, 9, 4, 8, 1, 7, 6, 3, 0), whose cost is $C = 1448608$. Every time a string is updated its new cost must be compared with the cost of the model string in the population and, in the case it is lower than that cost, the model string is replaced by the string with the lowest cost. Note that for $p = 0$ the noncooperative search strategy is recovered. The copying procedure proposed here was inspired by the mechanism used to model the influence of an external media in an agent-based model of culture dissemination [12, 13] as well as by Bandura’s social learning hypothesis [7]. In addition, we may interpret the process of learning (or copying) from a model string as a black-board cooperation system where a central control exhibits hints in a public space [8].

Figure 2 exhibits the effect of varying the probability $p$ that an agent will copy a digit-to-letter assignment from the model string (i.e., access the black-board to pick cues). The left panel shows that the distribution of the number of trials to success is exponential as in the case of noncooperative agents ($p = 0$). The dependence on $p$ is rather complex and can be best visualized by plotting the mean number of trials per agent to find the correct solution as done in the right panel. For fixed $p$, this quantity is essentially the reciprocal of the slope of the straight line that fits the scattered data shown in the left panel for $N = 20$. For $p = 0.7$ we observe a tenfold decrease of the mean time per agent needed to find the correct solution. This is a remarkable evidence of the power of collaboration in speeding up the search on the cost function landscape. Note that as $p \to 1$ one expects $\langle t \rangle$ to diverge since the search space cannot be fully explored because the option for the elementary move
is never made in this limit. For $p = 0$ we have the result for the noncooperative search, $\langle t \rangle N/10! \approx 1/0.9 \approx 1.1$. Most interestingly, the cooperative search with $N = 2$ is less efficient than the noncooperative one.

The effect of increasing the number of agents $N$ for a fixed value of the parameter $p$ is summarized in figure 3. For the range of values of $N$ considered (i.e., $2 \leq N \leq 20$) the probability distribution of the number of trials to find the correct solution is an exponential distribution (see left panel) and so, as before, we can derive the mean number of trials to success by estimating the slope of the straight lines that fit the data for $\zeta$ in the semilogarithmic scale of the left panel. The results for the properly scaled mean time to success per agent $\langle t \rangle N/10!$ are shown in the right panel of figure 3. The per agent efficiency of the cooperative system seems to decrease monotonically with increasing $N$. However, we must note that for $N > 20$ we find a considerable deviation from the exponential distribution with a large weight of the distribution being assigned to the region of $t \ll 10!$ (data not shown). In addition, we find that for $N > 20$ the per agent efficiency actually increases with increasing $N$ (data not shown). The explanation to this counterintuitive finding is the subject of ongoing research.

5 Conclusion

It is not the purpose of this paper to offer a novel method to solve cryptarithmetic or, more generally, constraint satisfaction problems; rather our aim here is to study quantitatively

Figure 3: (Color online) Right panel: Probability $\zeta$ that a collective search involving $N$ cooperating agents terminates at trial $t$ as function of the ratio between the trial number $t$ and the size of the search space $10!$ for $p = 0.7$ and $N = 2(\bigcirc), 5(\bigtriangleup), 10(\bigtriangledown)$ and $20(\square)$. Left panel: Mean number of trials per agent to find the correct solution as function of the number of agents $N$ for $p = 0.7(\bullet)$ and $0.4(\triangle)$. 
the advantages of cooperation or, more specifically, learning by imitation in the resolution of complex tasks. In particular, here we have focused on the speedup in the time to find the correct solution with respect to the number of agents when they are allowed to learn from the best performing agent – the model – following a model imitation mechanism inspired in Bandura’s social learning theory [7]. In that line, the results of figure 3 show a tenfold speedup in the search time per agent when about $N = 8$ or more agents are employed in the cooperative search. Most significantly, the per agent rate of improvement becomes negligible as more and more agents are employed in the search. Finally, we note that within the limited range of the number of agents considered in our study ($N \leq 20$) we have not observed a change from the exponential to the lognormal distribution in the effectiveness of an individual agent’s problem solving as predicted by a general theory of cooperative processes [5, 6].

Acknowledgements
This work has been partially supported by grant # 2013/17131-0, São Paulo Research Foundation (FAPESP) and by grant # 303979/2013-5, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References


Filling holes with volume constraints

M. A. Fortes\textsuperscript{1}, P. González\textsuperscript{1}, A. Palomares\textsuperscript{1} and M. Pasadas\textsuperscript{1}

\textsuperscript{1} Departamento de Matemática Aplicada, Universidad de Granada
emails: mafortes@ugr.es, prodelas@ugr.es, anpalom@ugr.es, mpasadas@ugr.es

Abstract

Key words: Filling, Approximation, Finite element, Minimal energy, Volume.

1 Introduction and preliminaries

Let $D \subset \mathbb{R}^2$ be a polygonal domain (an open polygonal connected set) and let us consider the Sobolev space $H^2(D)$, whose elements are (classes of) functions $u$ defined on $D$ such that their partial derivatives (in the distribution sense) $\partial^\beta u$ belong to $L^2(D)$, with $\beta := (\beta_1, \beta_2) \in \mathbb{N}^2$ and $|\beta| := \beta_1 + \beta_2 \leq 2$. For any open subset $X$ of $D$ we will consider the usual inner semi-products

$$(u, v)_{m,X} := \sum_{|\beta|=m} \int_X \partial^\beta u \cdot \partial^\beta v, \quad m = 0, 1, 2;$$

the seminorms

$$|u|_{m,X} := (u, u)_{m,X}^{1/2} = \left( \sum_{|\beta|=m} \int_X (\partial^\beta u)^2 \right)^{1/2}, \quad m = 0, 1, 2;$$

and the norm

$$\|u\|_X = \left( \sum_{m=0}^2 |u|_{m,X}^2 \right)^{1/2} = \left( \sum_{|\beta|\leq2} \int_X (\partial^\beta u)^2 \right)^{1/2}.$$ We will denote $< \cdot, \cdot>_n$ the usual Euclidean norm and $< \cdot, \cdot>_n$ the Euclidean inner product in $\mathbb{R}^n$. 

\textcopyright{CMMSE}  Page 568 of 1485  ISBN: 978-84-616-9216-3
Given $\alpha \geq 1$, let $T$ be an $\alpha$-triangulation of $D$, i.e., a triangulation that satisfies the condition

$$1 \leq \frac{R_T}{2r_T} \leq \alpha$$

for all closed triangles $T \in T$, $R_T$ and $r_T$ being the radii of the circumscribed and inscribed circles of $T$, respectively (see e.g. [5]), and let $V_T$ be the set of all the knots of $T$. We will consider the associated Powell-Sabin triangulation $T_6$ of $T$ (see e.g. [3]), which is obtained by joining the centre $\Omega_T$ of the inscribed circle of each interior triangle $T \in T$ to the vertices of $T$ and to the centres $\Omega_{T'}$ of the inscribed circles of the neighbouring triangles $T' \in T$. When $T$ has a side lying on the boundary of $D$, the point $\Omega_T$ is joined to the mid-point of this side, to the vertices of $T$ and to the centres $\Omega_{T'}$ of the inscribed circles of the neighbouring triangles $T' \in T$. Hence, all the micro-triangles inside any $T \in T$ have the incenter of $T$ as a common vertex.

It is well known ([4]) that given the values of a function $f$ (defined on $D$) and all its first partial derivatives at all the points of $V_T$, there exists a unique $S$ in

$$S^1_2(D, T_6) = \{ S \in C^1(D) : S|_{T'} \in \mathbb{P}_2(T') \hspace{1em} \forall T' \in T_6 \},$$

where $\mathbb{P}_2(T')$ stands for space of polynomials of total degree at most two over $T'$, such that the values of $S$ and all its first partial derivatives coincide with those of $f$ at all the points of $V_T$.

Let $H$ (the hole) be a connected and nonempty subset of $D$. If $H$ were not connected, the technique developed in the work to fill one connected hole would be applied to each of the connected components of $H$. Let

$$T_H = \{ T \in T : T \cap H \neq \emptyset \} \hspace{1em} \text{and} \hspace{1em} H^* = \bigcup_{T \in T_H} T.$$

\section{Volume constraints}

In this work we propose to fill the hole $H$ by fulfilling some volume constraints. More precisely, we will consider two problems:

\subsection{Interpolation volume constraint}

Let $P = \{ p_i \}_{i=1}^q$ be a $P_1$-unisolvent subset of points in $D - H^*$ and $z = (z_1, \ldots, z_q) \in \mathbb{R}^q$ a values vector. Let us define the functional

$$J_i : S^1_2(D, T_6) \to \mathbb{R}$$

$$v \mapsto J_i(v) = \rho(v) - z^T + \tau_1|v|_{1,D}^2 + \tau_2|v|_{2,D}^2,$$
where $\tau_1 \geq 0$, $\tau_2 > 0$ and $\rho$ is the evaluation operator $\rho(v) = (v(p_1), \ldots, v(p_q))$.

Given $V \in \mathbb{R}$, let us consider the set

$$W_i = \{v \in S^1_2(D; T_0) : \int_{H^*} v = V\}.$$

The minimization problem that we want to solve is:

**Problem $\mathcal{V}_i$.** Find $\sigma \in W_i$ such that $\mathcal{J}_i(\sigma) \leq \mathcal{J}_i(v)$ for all $v \in W_i$.

In this work, we show:

**Theorem 1.** Problem $\mathcal{V}_i$ has a unique solution, which is characterized as the unique solution to the next variational problem:

$$\text{Find } \sigma \in W_i \text{ such that } < \rho(\sigma), \rho(v) >_q + \tau_1(\sigma, v)_{1,D} + \tau_2(\sigma, v)_{2,D} = 0 \quad (1)$$

for all $v \in W^0_i = \{v \in S^1_2(D, T_0) : \int_{H^*} v = 0\}$.

**Theorem 2.** There exists a unique $(\sigma, \lambda) \in W \times \mathbb{R}$ such that

$$< \rho(\sigma), \rho(v) >_q + \tau_1(\sigma, v)_{1,D} + \tau_2(\sigma, v)_{2,D} + \lambda \int_{H^*} v = 0 \quad (2)$$

for all $v \in S^1_2(D, T_0)$, where $\sigma$ is the unique solution of Problem $\mathcal{V}_i$.

**Computation.**

Let $\mathcal{B} = \{\omega_1, \ldots, \omega_p\}$ be the Hermite basis of $S^1_2(D, T_0)$ and let

$$\sigma = \sum_{i=1}^{p} \gamma_i \omega_i$$

be the solution of Problem $\mathcal{V}_i$. Then, by applying (2), we know that $\{\gamma_1, \ldots, \gamma_p\}$ are obtained as the solution of the system

$$
\begin{pmatrix}
D & b^t \\
\hline
b & 0
\end{pmatrix}
\begin{pmatrix}
\gamma_1 \\
\vdots \\
\gamma_p \\
\gamma_p / \lambda
\end{pmatrix} =
\begin{pmatrix}
0 \\
\vdots \\
0 \\
V
\end{pmatrix},
$$

where

$$D = AA^t + \tau_1 R_1 + \tau_2 R_2,$$
being

\[ \mathcal{A} = (\omega_i(p_k))_{i=1,\ldots,p} , \]
\[ \mathcal{R}_k = ((\omega_i,\omega_j)_{k,D})_{1 \leq i,j \leq p} , \quad \text{for} \ k = 1, 2, \]
\[ b = \left( \int \int_{H^*} \omega_i \right)_{1 \leq i \leq p} . \]

### 2.2 Approximation volume constraint

Let \( f : D \to \mathbb{R} \) be a function and let us define the inner product

\[ < f, v >_{Vol} = \int \int_{H^*} f \cdot v , \quad \text{for} \ v \in \mathcal{S}_2^1(D, T_6) . \]

Let us consider the functional

\[ \mathcal{J}_a : \mathcal{S}_2^1(D, T_6) \to \mathbb{R} \]
\[ v \mapsto \mathcal{J}_a(v) = < \rho(v - f) >^2_q + \tau_0 < v - f, v - f >_{Vol} + \sum_{k=1}^{2} \tau_k |v|^{2}_{k,D-H^*} , \]

where \( \tau_1 \geq 0 \) and \( \tau_0, \tau_2 > 0 \).

The minimization problem that we want to solve is:

**Problem \( \mathcal{V}_a \).** Find \( \sigma \in \mathcal{S}_2^1(D, T_6) \) such that

\[ \mathcal{J}_a(\sigma) \leq \mathcal{J}_a(v) \quad \text{for all} \ v \in \mathcal{S}_2^1(D, T_6) . \]

In this work, we show:

**Theorem 2.** Problem \( \mathcal{V}_a \) has a unique solution, which is characterized as the unique solution to the next variational problem:

Find \( \sigma \in \mathcal{S}_2^1(D, T_6) \) such that

\[ < \rho(\sigma), \rho(v) >_q + \tau_0 < \sigma, v >_{Vol} + \sum_{k=1}^{2} \tau_k (\sigma, v)_{k,D-H^*} = < f, \rho(v) >_q + \tau_0 < f, v >_{Vol} \quad (3) \]

Let

\[ \sigma = \sum_{i=1}^{p} \gamma_i \omega_i \]

be the solution of Problem \( \mathcal{V}_a \). Then, by applying (3), we know that \( \{ \gamma_1, \ldots, \gamma_p \} \) are obtained as the solution of the system \( CX = B \), where

\[
C = \left( < \rho(\omega_i), \rho(\omega_j) >_q + \tau_0 < \omega_i, \omega_j >_{Vol} + \sum_{k=1}^{2} \tau_k (\omega_i, \omega_j)_{k,D-H^*} \right)_{i,j=1}^{p},
\]

and

\[
B = \left( < f, \rho(\omega_i) >_q + \tau_0 < f, \omega_i >_{Vol} \right)_{i=1}^{p}.\]
References


Inverse-free recursive multiresolution algorithms for a data approximation problem

M. A. Fortes\textsuperscript{1}, M. Raydan\textsuperscript{2} and A. M. Sajo-Castelli\textsuperscript{2}

\textsuperscript{1} Departamento de Matemática Aplicada, Universidad de Granada
\textsuperscript{2} Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar

emails: mafortes@ugr.es, mraydan@usb.ve, asajo@usb.ve

Abstract

We develop inverse-free recursive multiresolution algorithms for a problem posed in the field of data approximation. We consider several preconditioning strategies which take advantage of the very special structure of the involved matrices for solving the associated linear systems. As an application, we consider some examples regarding the noise reduction and the localization of regions in which the energy of a given surface is mostly concentrated.

Key words: Conjugate gradient method, preconditioning, Powell-Sabin element, minimal energy surfaces, data approximation, noise reduction.

1 Introduction

In [1] and references therein a method to obtain a $C^{1}$- spline surface approximating a given data set by minimizing a certain “energy functional” is considered. In [1] it is shown that, under certain conditions, such approximating surface exists and is unique. It is also shown a convergence result and some numerical and graphical examples are given. Later, in [3] a multiresolution scheme associated to the problem stated in [1] is considered. More precisely, the data approximation problem is considered in different resolution levels which are defined thorough a sequence of nested triangulations of the domain over which the surface is defined. The main contribution of [3] is the development of multiresolution algorithms (decomposition and reconstruction) in such a way that the minimal energy approximating splines obtained in different levels may be related. As an application, two examples regarding the noise reduction and the localization of regions in which the energy of a given surface
is mostly concentrated are considered. Nevertheless, the algorithms developed in [3] need to compute the inverse of large and very ill-conditioned matrices, impeding to carry the considered applications beyond resolution levels three or four.

In the present work we intend to overcome this problem by developing an inverse-free recursive equivalent formulation of the multiresolution algorithms. This new formulation allow us to consider several preconditioning strategies which take advantage of the very special structure of the involved matrices for solving the associated linear systems.

2 Preliminaries and basic concepts

Let $D \subset \mathbb{R}^2$ be a polygonal domain (an open, non-empty connected set) in such a way that $\overline{D}$ admits a $\Delta^1$-type triangulation (see e.g. [2]). Let us consider the Sobolev space $\mathcal{H}^2(D)$, whose elements are (classes of) functions $u$ defined on $D$ such that their partial derivatives (in the distribution sense) $\partial^\beta u$ belong to $\mathcal{L}^2(D)$, with $\beta = (\beta_1, \beta_2) \in \mathbb{N}^2$ and $1 \leq |\beta| \leq 2$, where $|\beta| = \beta_1 + \beta_2$. In this space we consider the usual inner semi-products

$$
(u, v)_t = \sum_{|\beta|=t} \int_D \partial^\beta u(x) \partial^\beta v(x) dx \quad \text{for} \quad t = 0, 1, 2,
$$

the corresponding semi-norms

$$
|u|_t = (u, u)_t^{1/2} = \left( \sum_{|\beta|=t} \int_D \partial^\beta u(x)^2 dx \right)^{1/2} \quad \text{for} \quad t = 0, 1, 2,
$$

and the norm

$$
\|u\| = \left( \sum_{t=0}^2 |u|_t^2 \right)^{1/2}.
$$

We will consider the Powell-Sabin subtriangulation $T_6$ associated to $T$ (see [4] or [5]), which is obtained by subdividing each triangle $T \in T$ into six subtriangles by connecting each vertex of $T$ to the midpoint of the opposite side. Hence, all these micro triangles have the barycenter of $T$ as a common vertex. Nevertheless, Powell-Sabin subtriangulation can be also obtained by using the incenter of $T$ instead of the barycenter in the split procedure for a larger class of triangulations (see [6]).

Let us consider the set

$$
S^1_2(D; T_6) = \{ v \in C^1(D) : v|_{T'} \in \mathbb{P}_2(T') \quad \forall T' \in T_6 \},
$$

where $\mathbb{P}_2(T')$ stands for the space of bivariate polynomials of total degree at most two over $T'$. In [4] it is shown that given a function $f \in C^m(\overline{D})$, $m \geq 1$, there exists a unique function
v \in S_2^1(D, T_0) such that the values of v and all its first partial derivatives coincide with those of f at all the vertices of T.

Let us now consider a finite subset \( \mathcal{P} = \{p_1, \ldots, p_q\} \) of points in \( \overline{D} \) and a given vector of real values \( Z = (z_i)_{i=1}^q \in \mathbb{R}^q \). From the continuous injection of \( \mathcal{H}^2(D) \) into \( C^0(\overline{D}) \), we can define the evaluation operator

\[
\rho : \mathcal{H}^2(D) \rightarrow \mathbb{R}^q \\
v \mapsto \rho(v) = (v(p_i))_{i=1}^q.
\]

We are looking for a \( C^1 \)-surface that approximates the points \( \{(p_i, z_i)\}_{i=1}^q \subset \mathbb{R}^3 \) by minimizing the functional energy

\[
J(v) = \langle \rho(v) - Z \rangle^2_q + \tau_1|v|_1^2 + \tau_2|v|_2^2,
\]

where \( \langle \cdot, \cdot \rangle_q \) represents the usual Euclidean norm in \( \mathbb{R}^q \), \( \tau_1 > 0 \) and \( \tau_2 \geq 0 \). Observe that the first term of \( J \) measures (in the least squares sense) how well \( v \) approximates the values in \( Z \), while the second one represents the “minimal energy condition” over the semi-norms \( \cdot \rvert_1 \) and \( \cdot \rvert_2 \) weighted by the parameters \( \tau_1 \) and \( \tau_2 \), respectively. More precisely the minimization problem under consideration is:

Problem 1. Find \( \sigma \in S_2^1(D, T_0) \) such that \( J(\sigma) \leq J(v) \) for all \( v \in S_2^1(D, T_0) \).

In [1] it is shown that, under the condition \( \text{Ker}(\rho) \cap P_1(D) = \{0\} \), Problem 1 has a unique solution. Moreover, if \( N = \dim(S_2^1(D, T_0)) \), we consider a basis \( \{v_1, \ldots, v_N\} \) of the space \( S_2^1(D, T_0) \), and we denote the unique solution of Problem 1 as \( \sigma = \sum_{i=1}^N \alpha_i v_i \), then the coefficients vector \( (\alpha_i)_{i=1}^N \) is obtained as the unique solution of the linear system

\[
CX = T,
\]

where

\[
C = \left( \langle \rho(v_i), \rho(v_j) \rangle_q + \sum_{t=1}^2 \tau_t(v_i, v_j)_t \right)_{i,j=1}^N \quad \text{and} \quad T = (\langle Z, \rho(v_i) \rangle_q)_{i=1}^N.
\]

In addition, in [1] it is also shown that the coefficient matrix \( C \) is symmetric and positive definite.

3 Multiresolution algorithms

Given two sets of functions \( \mathcal{F} = \{f_i\}_{i=1}^{d_f} \) and \( \mathcal{G} = \{g_j\}_{j=1}^{d_g} \) of the space \( S_2^1(D, T_0) \) we will denote

\[
C_{\mathcal{F}, \mathcal{G}} = \left( \langle \rho(f_i), \rho(g_j) \rangle_q + \sum_{t=1}^2 \tau_t(f_i, g_j)_t \right)_{i=1, \ldots, d_f; j=1, \ldots, d_g}
\]
Inverse-free recursive multiresolution algorithms for a data approximation problem

and

\[ T_F = \left( (< Z, \rho(f_i) > q_i)_{i=1}^t \right). \]

Let \( \{ T_j \}_{j \geq 0} \) be a sequence of refined uniform \( \Delta^1 \)-triangulations of \( \Omega \) in such a way that the sequence of splines spaces \( V_j = S^1_0(D, (T_j)_0) \) is nested and let us denote \( n_j = \dim(V_j) \). Let \( B_0 = \{ \varphi^0_k \}_{k=1}^{n_0} \) be the usual Hermite basis of \( V_0 \) and, for all \( j \geq 0 \), let us consider \( \Psi_j = \{ \psi^k_j \}_{k=1}^{n_{j+1}-n_j} \) the functions of the Hermite basis of \( V_{j+1} \) associated to the knots belonging to \( T_{j+1} \) that do not belong to \( T_j \). Then, it is clear that if we define

\[
\begin{cases}
\varphi^k_{j+1} = \varphi^k_j & \text{for } k = 1, \ldots, n_j \\
\varphi^k_{j+1} = \psi^k_{j-n_j} & \text{for } k = n_j + 1, \ldots, n_{j+1},
\end{cases}
\]

we have that \( B_{j+1} = \{ \psi^k_{j+1} \}_{k=1}^{n_{j+1}} = B_j \cup \Psi_j \) is a basis with local support of the space \( V_{j+1} \) for \( j \geq 0 \).

Let \( \sigma_j \) be the unique solution of Problem 1 for \( T_j = T_j \). Then, there exists a sequence \( a_j = \{ a^k_j \}_{k=1}^{n_j} \) such that

\[
\sigma_j = \sum_{k=1}^{n_j} a^k_j \varphi^k_j,
\]

(2)

and, from (1), it follows that, for all \( j \geq 0 \),

\[
C_{B_j, B_j} a_j = T_{B_j}.
\]

(3)

Moreover, for all \( j \geq 1 \) we have that

\[
C_{B_j, B_j} = \left( \frac{C_{B_{j-1}, B_{j-1}}}{C_{\Psi_{j-1}, B_{j-1}}} \right) \left( \frac{C_{B_{j-1}, \Psi_{j-1}}}{C_{\Psi_{j-1}, \Psi_{j-1}}} \right)
\]

(4)

and

\[
T_{B_j} = \left( \frac{T_{B_{j-1}}}{T_{\Psi_{j-1}}} \right).
\]

(5)

Hence, if we denote the element \( \sigma_j = \sum_{k=1}^{n_j} a^k_j \varphi^k_j \) as

\[
\sum_{k=1}^{n_{j-1}} \tilde{a}^k_{j-1} \varphi^k_{j-1} + \sum_{k=1}^{n_{j-1}-n_j} a^k_{j-1} \psi^k_{j-1},
\]

that is,

\[
\{ a^1_j, a^2_j, \ldots, a^{n_j}_j \} = \{ \tilde{a}^1_{j-1}, \tilde{a}^2_{j-1}, \ldots, \tilde{a}^{n_{j-1}}_{j-1}, d^1_{j-1}, d^2_{j-1}, \ldots, d^{n_{j-1}-n_j-1}_{j-1} \},
\]

(6)

then from (3), (4), (5) and (6), it holds
\begin{align}
C_{B_{j-1}, B_{j-1}} \tilde{a}_{j-1} + C_{B_{j-1}, \Psi_{j-1}} d_{j-1} &= T_{B_{j-1}}, \quad (7) \\
C_{\Psi_{j-1}, B_{j-1}} \tilde{a}_{j-1} + C_{\Psi_{j-1}, \Psi_{j-1}} d_{j-1} &= T_{\Psi_{j-1}}, \quad (8)
\end{align}

where \( \tilde{a}_{j-1} = \left( (a^k_j)_{k=n_{j-1}+1}^{n_j} \right)^t \) and \( d_{j-1} = \left( (a^k_j)_{k=1}^{n_j-n_{j-1}} \right)^t \).

By replacing \( T_{B_{j-1}} \) by \( C_{B_{j-1}, B_{j-1}} a_{j-1} \) in (7) we obtain

\[
a_{j-1} - \tilde{a}_{j-1} = (C_{B_{j-1}, B_{j-1}})^{-1} C_{B_{j-1}, \Psi_{j-1}} d_{j-1}. \quad (9)
\]

From this last equation and by the use of the notation introduced in (2), the multiresolution algorithms can be formulated as follows:

\textit{Decomposition algorithm}:

\[
\begin{cases}
    d_{j-1} = \left((a^k_j)_{k=n_{j-1}+1}^{n_j}\right)^t \\
a_{j-1} = \tilde{a}_{j-1} + (C_{B_{j-1}, B_{j-1}})^{-1} C_{B_{j-1}, \Psi_{j-1}} d_{j-1}
\end{cases} \quad (10)
\]

\textit{Reconstruction algorithm}:

\[
\begin{cases}
    \left((a^k_j)_{k=n_{j-1}+1}^{n_j}\right)^t = d_{j-1} \\
    \left((a^k_j)_{k=1}^{n_j-n_{j-1}}\right)^t = \tilde{a}_{j-1} = a_{j-1} - (C_{B_{j-1}, B_{j-1}})^{-1} C_{B_{j-1}, \Psi_{j-1}} d_{j-1}
\end{cases} \quad (11)
\]

4 Numerical study

Algorithms (10) and (11) allow us to transfer information between different resolution levels, in such a way that the minimal energy approximating splines between different levels may be related. This is an important matter since several applications can be derived from this fact: for example, in [3] these algorithms are applied in order to reduce the noise in a given dataset and also to localize the regions where the energy of a given function is mostly concentrated. Nevertheless, the algorithms in their present forms have a numerical drawback: it is necessary to compute the inverse of the matrices \( C_{B_j, B_j} \), which, as \( j \) grows, become large and very ill-conditioned. This is an important criticism to carry the applications further, since it is not possible to consider resolution levels beyond \( j = 3 \) or \( j = 4 \).

In this work we develop a recursive equivalent formulation of the algorithms (10) and (11) with the objective of avoiding the computation of any inverse matrix. Besides, this new formulation allows us to consider several preconditioning strategies which take advantage of the very special structure of the matrices \( C_{B_j, B_j} \). This numerical treatment of the multiresolution algorithms allow us to carry the applications to higher resolution levels.
References


Computation of highly accurate binding energies of large clusters

Joachim Friedrich and Tony Anacker

1 Institut für Chemie, Technische Universität Chemnitz

emails: joachim.friedrich@chemie.tu-chemnitz.de, tony.anacker@chemie.tu-chemnitz.de

Abstract

In this work, we use MP2, CCSD(T) and their explicitly correlated (F12) counterparts to compute the interaction energies of water icosamers with high accuracy [1]. Since these systems are far too big for a standard CCSD(T) treatment in an appropriate basis set, we apply the incremental scheme to compute benchmark energies at the CCSD(T)/CBS(45) and CCSD(T)(F12*)/cc-pVQZ-F12 level of theory. The four structures, dodecahedron, edge sharing, face sharing and fused cubes, are part of the WATER27 test set and therefore, highly accurate interaction energies are required. To obtain these energies, we carefully converge the interaction energies with respect to the basis set. Furthermore, analyze the basis set superposition error as well as the core-valence correlation. Our state of the art interaction energies are: dodecahedron -198.6 kcal/mol, edge sharing -209.7 kcal/mol, face sharing -208.0 kcal/mol and fused cubes -208.0 kcal/mol.

Key words: incremental scheme, high accuracy, coupled cluster calculations

1 Introduction

In a recent work, Goerigk and Grimme presented a quantum chemistry benchmark database for general main group thermo chemistry, kinetics and non-covalent interactions (GMTKN24) consisting of 24 different, chemically relevant subsets [2]. The purpose of such databases is to evaluate the general accuracy of different density functionals and to analyze which functional is optimal for different application classes [3]. The reference values in the GMTKN24 are based on theoretical or experimental reference values. A general accepted benchmark method is coupled-cluster with single, double and perturbative triple excitations (CCSD(T))
in combination with an appropriate one-particle basis set and an extrapolation to the complete basis set (CBS). For water clusters it was found, that the computationally cheap MP2 method in combination with a large basis set is suitable to compute sufficiently accurate interaction energies. Considering the WATER27 subset, Goerigk and Grimme found very big errors in the interaction energies of water icosamers for recently developed, accurate density functionals, when comparing to the MP2/CBS benchmark values from Xantheas [4]. For water clusters the MP2/CBS values should be suitable benchmarks [5, 6]. However, in this work we show that the old benchmarks have rather large errors.

The use of a finite basis set in quantum chemical calculations introduces the basis set superposition error (BSSE). For clusters one can correct for the BSSE using the site-site function counterpoise (SSFC) approach of Wells and Wilson [7]. The SSFC approach is a generalization of the CP-correction of Boys and Bernadi [8, 7, 9, 10]. For a convenient computation of BSSE-effects we use our fully automatic code [11, 12].

2 Computational procedure

2.1 Incremental scheme

The incremental scheme is used to obtain the CCSD(T) correlation energy \( E_{\text{corr}} \) for the large water clusters studied in this work. Within this framework, the total system is divided into small one-site domains consisting of localized orbitals [13, 14, 15, 16]. Based on these domains the correlation energy is computed for all domains, domain pairs and triples of domains. The total correlation energy is computed with the incremental expansion [13, 17]:

\[
E_{\text{corr}} = \sum_{X \in \mathcal{P}(\mathbb{D}) \land |X| < O} \Delta \varepsilon_X
\]

where \( \mathcal{P}(\mathbb{D}) \) is the power set of the set of domains \( \mathbb{D} \). The expansion is truncated by the cardinality of the sets \( X \). The increment itself is defined as

\[
\Delta \varepsilon_X = \varepsilon_X - \sum_{Y \in \mathcal{P}(X) \land |Y| < |X|} \Delta \varepsilon_Y
\]

\( \varepsilon_X \) is the correlation energy of the domain \( X \).

In an incremental calculation the virtual space of a domain is reduced by using the domain-specific basis set approach [18]. With this approximation we account for the fact, that the virtual orbitals far from the correlated orbitals of the domain are not needed in the computation of the correlation energy and can therefore be neglected.

The incremental scheme is very efficient and applicable to large systems, when the incremental expansion is truncated at low order. The error introduced in the domain-specific
basis set approximation can be estimated with the computationally cheap MP2 method:

\[ E_{\text{corr}}(\text{CCSD(T)}|\text{MP2}) = E_{\text{corr}}^{\text{inc}}(\text{CCSD(T)}) - (E_{\text{corr}}^{\text{inc}}(\text{MP2}) - E_{\text{corr}}(\text{MP2})) \]  

We demonstrated in our previous work, that the MP2 based error correction works well [12, 19].

### 2.2 CBS extrapolation

The dominant source of error in most calculations is the truncation of the one-particle basis set. Therefore, we extrapolate the energy to the complete basis set limit (CBS). The Hartree-Fock (HF) energy is extrapolated with the approach of Peterson and coworkers [20] in combination with the augmented correlation-consistent polarized valence \( X \)-tuple basis sets:

\[ E_{\text{SCF}}^{XY} = \frac{e^{-a\sqrt{Y}}E_X - e^{-a\sqrt{X}}E_Y}{e^{-a\sqrt{Y}} - e^{-a\sqrt{X}}} \]  

Where \( X \) and \( Y \) are the \( \zeta \)-level and \( X < Y \). As recommended the value for the universal constant \( a \) is set to 6.30 [20].

For the extrapolation of the correlation energy we use the expression of Helgaker and coworkers [21]:

\[ E_{\text{corr}}^{XY} = \frac{X^3E_X^{\text{corr}} - Y^3E_Y^{\text{corr}}}{X^3 - Y^3} \]  

### 2.3 The interaction energy

The interaction energy \( E_{\text{int}} \) of a water cluster with \( n \) monomers is computed by:

\[ E_{\text{int}} = E_n - nE_{\text{monomer}} \]  

This quantity is usually very difficult to compute for larger clusters, since approximate methods are not accurate enough and highly accurate wave function based methods are very time-consuming, due to their unfavorable scaling behavior with the system-size and their large basis set-requirements.

### 2.4 SSFC

The counterpoise (CP) method of Boys and Bernadi [8] is exclusively applied for two-body interactions. A generalization of this scheme for many-body interactions is the site-site function counterpoise (SSFC) method [7]. In this scheme the total BSSE of a \( n \)-body cluster is approximated as

\[ \text{SSFC} = \sum_i^n \left( E_i^{ijkl...n} - E_i^i \right) \]
where $E_i$ is the energy of the monomer $i$ in the monomer basis and $E_{ijkl\ldots n}$ the monomer energy in the full basis of the $n$-body cluster. Note that at the CBS limit the total BSSE of the SSFC calculation is zero [22].

3 Results

Comparing the MP2/CBS(45) with the MP2-F12/cc-pVQZ-F12 interaction energies, we see that they differ by ca. 1 kcal/mol. The agreement of these energies is a hint that the exact energy is close to these energies. Also the incremental CCSD(T)|MP2 interaction energies are about 1 kcal/mol smaller than the incremental CCSD(T)(F12*)|MP2-F12 energies.

To take account for the core-valence contribution, the difference between the MP2/CBS(45) energies with all electrons correlated using core-valence basis sets and frozen core is added to the CCSD(T)/CBS(45) and CCSD(T)(F12*)/cc-pVQZ-F12 energies.

Finally we use the incremental CCSD(T)|MP2+CP/CBS(45)+ΔMP2 interaction energies as benchmarks, because we expect them to be the most accurate interaction energies of this work. These energies can used to benchmark density functionals to evaluate their accuracy.

Acknowledgements

We thank the Fonds der Chemischen Industrie for financial support due to the Material Cost Allowances program and the DAAD for a travel grant.

References


Table 1: Interaction energies $E_{\text{int}}$ [kcal/mol] of the four structures calculated on different levels of theory. The CBS(45) extrapolation is given for the non-explicitly correlated methods and cc-pVQZ-F12 basis set for the F12 methods. iCC is the shortcut for the incremental CCSD(T)|MP2, iCC(F12*) for the incremental CCSD(T)(F12*)|MP2-F12 method. fc refers to the frozen-core approximation and $\Delta MP2_{cv}$ is the correction of the core-valence contribution.

<table>
<thead>
<tr>
<th>Method</th>
<th>dodecahedron</th>
<th>edge sharing</th>
<th>face sharing</th>
<th>fused cubes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP2$_{fc}$</td>
<td>-198.9</td>
<td>-208.7</td>
<td>-206.7</td>
<td>-206.3</td>
</tr>
<tr>
<td>MP2$_{fc}$+CP</td>
<td>-198.7</td>
<td>-208.4</td>
<td>-206.5</td>
<td>-206.1</td>
</tr>
<tr>
<td>MP2$<em>{fc}$+$\Delta MP2</em>{cv}$</td>
<td>-200.1</td>
<td>-209.9</td>
<td>-207.9</td>
<td>-207.5</td>
</tr>
<tr>
<td>iCC</td>
<td>-197.7</td>
<td>-208.7</td>
<td>-207.1</td>
<td>-207.1</td>
</tr>
<tr>
<td>iCC+CP</td>
<td>-197.4</td>
<td>-208.4</td>
<td>-206.8</td>
<td>-206.8</td>
</tr>
<tr>
<td>iCC+CP+$\Delta MP2_{cv}$</td>
<td><strong>-198.6</strong></td>
<td><strong>-209.7</strong></td>
<td><strong>-208.0</strong></td>
<td><strong>-208.0</strong></td>
</tr>
<tr>
<td>MP2-F12</td>
<td>-199.8</td>
<td>-209.6</td>
<td>-207.7</td>
<td>-207.3</td>
</tr>
<tr>
<td>iCC(F12*)</td>
<td>-198.6</td>
<td>-209.6</td>
<td>-207.9</td>
<td>-207.9</td>
</tr>
<tr>
<td>iCC(F12*)+$\Delta MP2_{cv}$</td>
<td><strong>-199.7</strong></td>
<td><strong>-210.8</strong></td>
<td><strong>-209.1</strong></td>
<td><strong>-209.1</strong></td>
</tr>
</tbody>
</table>
The Holling-Tanner predation model with a special weak Allee effect on prey

Lina María Gallego-Berrío¹ and Eduardo González-Olivares²

¹ Facultad de Ciencias y Tecnologías, Maestra en Biomatemáticas, Universidad del Quindío, Colombia
² Grupo de Ecología Matemática, Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Chile
emails: linag@uniquindio.edu.co, ejgonzal@ucv.cl

Abstract

In this work a modified Holling-Tanner predator-prey model is analyzed, considering two important aspects describing the interaction: the predator growth rate is of logistic type and the prey population is affected by a weak Allee effect.

Making a time rescaling, a polynomial differential equations system is obtained, topologically equivalent to original one.

We prove that for certain subset of parameters, the model exhibits a bistability phenomenon, since, the equilibrium point $(0, 0)$ is an attractor for all parameter values, which coexists with a stable limit cycle or a positive equilibrium point.

Also we show the existence of separatrix curves on the phase plane that divides the behavior of the trajectories, and two nearby solutions can have different $\omega$-limit.

Key words: Predator-prey model, Allee effect, stability, limit cycle, bifurcation, separatrix curve.

1 Introduction

In this work, a deterministic continuous predator-prey model is analyzed, in which are consider two important aspects describe the interaction:

i) the equation for predator is a logistic type growth function and

ii) the prey population is affected by the Allee effect.

The first aspect characterize to Leslie type predator-prey models [19, 25] or Leslie-Gower model [1, 3, 4], where the conventional environmental carrying capacity $K_y$ is a function
Holling-Tanner model with weak Allee effect

of prey population size \( x = x(t) \), that is, dependent of the available resources [19]. Here we consider that \( K_y = n x \), proportional to prey abundance as in the May-Holling-Tanner model [2, 17, 22], assuming implicitly that predators are specialist.

We also assume the prey population is affected by the Allee effect [23], which is a situation at low population sizes where the per-individual growth rate is an increasing function of population density [12]. More precisely, any ecological mechanism that can lead to a positive relationship between a component of individual fitness and either the number or density of conspecifics can be termed a mechanism of the Allee effect [10, 24].

This is a common phenomenon and populations may exhibit Allee effect dynamics due to a wide range of biological phenomena (see Table 1 in [5] or Table 2.1 in [11]). Recent ecological research suggests that two or more Allee effects can lead to these mechanisms acting simultaneously on a single population (Table 2 in [5]); the combined influence of some of these phenomena is known as multiple (double) Allee effect [5].

The Allee effect is also called negative competition effect [26] in Population Dynamic or depensation [8, 12, 18] in Fisheries sciences, have been observed on different organisms, such as vertebrates, invertebrates and plants (see, for instance [5]). Previous studies demonstrate that Allee effects play an important role on the stability analysis of equilibrium points of a population model [9, 11] and usually is assumed that there is a critical threshold below which the population experiences extinction [8, 25].

Many mathematical forms have been proposed to model the Allee effect [5] and most of them are topologically equivalents [14]. Here, we consider the most usual form to describe this phenomenon that we called the multiplicative Allee effect [1], showing its strong influence on the Holling-Tanner model [22], since it can appears until two positive equilibrium points, which can coincide obtaining a cusp point [7] for certain parameter constraints, i.e, it has a Bogdanov-Takens bifurcation [7].

2 The model

The modified May-Holling-Tanner (or Leslie-Gower type model) [4, 22] is expressed by the following autonomous bidimensional differential equations system:

\[
X_\mu : \begin{cases}
\frac{dx}{dt} = x \left( r \left( 1 - \frac{x}{K} \right) (x - m) - \frac{q y}{x + a} \right) \\
\frac{dy}{dt} = s y \left( 1 - \frac{y}{n x} \right)
\end{cases}
\]

(1)

where \( x = x(t) \) and \( y = y(t) \) indicate the prey and predator population size of respectively (measured in biomass, density or number) for \( t \geq 0 \), with

\[ \mu \in \Delta = \{(r, K, q, a, s, n, m) \in \mathbb{R}^7_+ / 0 < a < K; -K < m < K \}. \]

The parameters have the following meaning: \( r \) and \( s \) are the intrinsic growth rates of the prey and predators, respectively; \( K \) is the prey environment carrying capacity; \( q \) is the predator maximum consumption rate per capita; \( m \) is the parameter of Allee effect; \( a \) is the
amount of prey for which the predation effect is maximum; \( n \) is a measure of the quality of the food that provides the prey and it is converted into predators birth.

If \( m > 0 \), it has a threshold level of population below which the population growth rate is negative, having a strong Allee effect; if \( m \leq 0 \), it has a weak Allee effect [23]. In Fisheries Sciences are named critical depensation and noncritical depensation [8], respectively.

Many functions had been employed to describe this effect; in [12, 23] is derived a function as a model of mating probability [12], modifying the logistic growth function.

System (1) is not defined upon the \( y - axis \), in particular at \((0, 0)\); it is a Kolmogorov type system [13] definite in
\[
\Omega = \{(x, y) \in \mathbb{R}^2 \mid x > 0, y \geq 0\}.
\]
The equilibrium point of system (1) or singularities of vector field \( X_\mu \) are \( P_K = (K, 0) \) and the intersection points between the isoclines
\[
r (1 - \frac{x}{K}) (x - m) - \frac{K}{a + x} y = 0 \quad \text{and} \quad y = n x.
\]

In order to simplify the calculations, we follow the methodology used in [16, 20], making a change of variables and a time rescaling.

Let \( \hat{\Omega} = \{(u, v) \in \mathbb{R}^2/ u \geq 0, v \geq 0\} \) and the diffeomorphism [7] given defined by \( \varphi : \hat{\Omega} \times \mathbb{R} \rightarrow \Omega \times \mathbb{R} \) such that,
\[
\varphi (u, v, \tau) = \left( Ku, Knv, \frac{r K}{u(Ku+a)} \tau \right) = (x, y, t)
\]
Since \( \det D\varphi (u, v, \tau) = \frac{K^2nr}{u(Ku+a)} > 0 \), \( \varphi \) is a diffeomorphism preserving the orientation of time; thus, the vector field \( X_\mu \), is topologically equivalent to the vector field \( Y_\eta = \varphi \circ X_\mu \) with \( Y_\eta = P(u, v) \frac{\partial}{\partial u} + Q(u, v) \frac{\partial}{\partial v} \); the associated differential equation system is given by:
\[
Y_\eta : \left\{ \begin{array}{l}
\frac{du}{d\tau} = ((1 - u)(u - M)(u + A) - Qv) u^2 \\
\frac{dv}{d\tau} = B (u - v)(u + A)v
\end{array} \right.
\]
with \( A = \frac{a}{K}, M = \frac{m}{K}, B = \frac{s}{rK}, Q = \frac{qm}{rK} \), and \( \eta = (M, A, Q, B) \in [-1, 0] \times [0, 1] \times \mathbb{R}^2 \); Moreover \( 0 < A < 1 \) and \(-1 \leq M < 1\).

It is clear that system (2) is definite at \( \hat{\Omega} \); the equilibrium points of system (2) are \((0, 0), (1, 0), (M, 0)\) and those points lying over the isoclinic curves \((1 - u)(u - M)(u + A) - Qv = 0\) and \(u - v = 0\), and whose abscise satisfy the cubic equation:
\[
u^3 - (M - A + 1) u^2 - (A - M - Q + AM) u + AM = 0
\]

According to the Descartes sign rule, the above equation can have two positives real root, or one of multiplicity two or none, but always it has a negative real root.

In the following, due to the difficult to express explicitly these positive roots, we will consider only the particular case in which \( M = 0 \) in system (2) (or \( m = 0 \) in vector field \( Y_\eta \), having now the vector field \( Y_\lambda \) described by the system
\[
Y_\lambda : \left\{ \begin{array}{l}
\frac{du}{d\tau} = ((1 - u)(u + A) - Qv) u^2 \\
\frac{dv}{d\tau} = Bv (u - v)(u + A)
\end{array} \right.
\]
Holling-Tanner model with weak Allee effect

with \( \lambda = (A, B, Q) \in [0, 1] \times \mathbb{R}_+^2 \). The equilibrium points of system (3) are \((0, 0)\), \((1, 0)\), and \((u_e, v_e)\) over the isoclinic curves

\[
v = \frac{1}{Q} (1 - u) u (u + A) \quad \text{and} \quad v = u.
\]

Then, the absciss of positive equilibrium points satisfy the quadratic equation:

\[
u^2 - u (1 - A) + (Q - A) = 0 \tag{4}
\]

where \( 1 - A > 0 \). The Jacobian matrix of system (3) is

\[
DY_\lambda(u, v) = \begin{pmatrix}
Y_{11} & -Qu^2 \\
Bv (2u + A - v) & B (u + A) (u - 2v)
\end{pmatrix}
\]

with \( Y_{11} = -u \left( 5u^3 - 4 (1 - A) u^2 - 3Au + 2Qv \right) \)

3 Main results

For system (3) we have the following results:

Lemma 1 Invariant region and boundness

a) The set \( \tilde{\Gamma} = \{(u, v) \in \tilde{\Omega}/ \ 0 \leq u \leq 1, \ v \geq 0 \} \) is an invariant region.

b) The solutions are bounded.

Proof. a) Clearly the \( u - axis \) and the \( v - axis \) are invariant sets.

If \( u = 1 \), we have that \( \frac{du}{d\tau} = -Qv \), and any sign of

\[
\frac{dv}{d\tau} = Bv (1 + A) (1 - v)
\]

the trajectories enter and remain in the set \( \tilde{\Gamma} \).

b) Using the Poincaré compactification [7] given by the transformation \( X = \frac{u}{v} \) and \( Y = \frac{1}{v^2} \) then \( \frac{dX}{d\tau} = \frac{1}{v^3} \left( v \frac{du}{d\tau} - u \frac{dv}{d\tau} \right) \) and \( \frac{dY}{d\tau} = -\frac{1}{v^2} \frac{dv}{d\tau} \) and using the blowing up method.

Lemma 2 i) There exist two positive equilibrium points \((u_1, u_1)\) and \((u_2, u_2)\) given by

\[
u_1 = \frac{1}{2} \left( 1 - A - \sqrt{(1 + A)^2 - 4Q} \right)
\]

\[
u_2 = \frac{1}{2} \left( 1 - A + \sqrt{(1 + A)^2 - 4Q} \right)
\]

if and only if, \( Q < \frac{(A+1)^2}{4} \) and \( Q > A \).

ii) There exists a unique positive equilibrium point \((u_e, u_e) = (1 - A, 1 - A)\), if and only if, \( Q = A \).

iii) There exists a unique positive equilibrium point

\[
u_2 = \frac{1}{2} \left( 1 - A + \sqrt{(1 + A)^2 - 4Q} \right), \text{ if and only if, } Q < A.
\]

iv) There exists a unique positive equilibrium point of multiplicity two \((E_c, E_c)\) with \( E_c = \frac{1}{2} (1 - A) \), if and only if, \( Q = \frac{(A+1)^2}{4} \).

v) There not exists positive equilibrium point if and only if, \( Q > \frac{(A+1)^2}{4} \).
Proof. The quantity of equilibrium point of system (3) depends of equation (4). By the Descartes sign rule, this equation can have two positives real root, or one of multiplicity two or none. Clearly, if \( Q - A \leq 0 \), there exits a unique positive real root.

Assuming \( Q - A > 0 \), then equation (4)

iii) has two real root, if and only if, \((A + 1)^2 - 4Q > 0\),

iii) has one real root of multiplicity two, if and only if, \((A + 1)^2 - 4Q = 0\),

iv) has not real root, if and only if, \((A + 1)^2 - 4Q < 0\).

Lemma 3 For all \( \lambda = (A, B, Q) \in ]0, 1[ \times \mathbb{R}_2^+ \), the singularity \((1, 0)\) is a saddle point.

Proof. The Jacobian matrix evaluated at \((1, 0)\) is

\[
DY_\lambda (1, 0) = \begin{pmatrix}
- (1 + A) & -Q \\
0 & B(1 + A)
\end{pmatrix}
\]

As \( \det Y_\lambda (1, 0) = -B (1 + A)^2 < 0 \), then \((1, 0)\) is an hyperbolic saddle point.

Lemma 4 The point \((0, 0)\) is a non-hyperbolic singularity, which a stable manifold \(W^s (0, 0)\) that determines a separatrix between obtaining an hyperbolic sector and a parabolic sector at the neighborhood of this equilibrium point [21].

Proof. The Jacobian matrix at \((0, 0)\) is the null matrix, and this point is non-hyperbolic singularity. To desingularize the origin we consider the polar blowing up [7, 21] given by

\[\Phi : S^1 \times \mathbb{R}_0^+ \to \mathbb{R}^2\] such as \(\Phi (r, \theta) = r \cos \theta, r \sin \theta\).

Thus, \(\Phi^*(Y_\lambda) = (D\Phi)^{-1} Y_\lambda \phi = \tilde{X}\),

where \(\tilde{X} = r\tilde{X} : S' \times \mathbb{R}_0^+ \to T \left(S' \times \mathbb{R}_0^+\right)\),

with \(\tilde{X} (r, \theta) = rf (r, \theta) \frac{\partial}{\partial r} + g (r, \theta) \frac{\partial}{\partial \theta}\). Then,

\[
rf (r, \theta) = r(-r^3 \cos^8 \theta + (-\cos^5 \theta) A + \cos^5 \theta) r^2
+ (-B \cos \theta \sin \theta + B \cos^3 \theta \sin \theta - (\cos^3 \theta) Q \sin \theta + (\cos^4 \theta) A
- B \cos^4 \theta + B \cos^2 \theta) r - B (\cos^3 \theta) A - BA \sin \theta + BA \sin \theta \cos^2 \theta
+ B (\cos \theta) A)
\]

and

\[
g (r, \theta) = (\cos \theta \sin \theta) ((\cos^4 \theta) r^3 + ((\cos^3 \theta) A - \cos^3 \theta) r^2
+ (-B \cos \theta \sin \theta + (\cos \theta \sin \theta) Q + B \cos^2 \theta - (\cos^2 \theta) A) r
- BA \sin \theta + B (\cos \theta) A)
\]

For \( r > 0 \), the dynamics at \( S^1 \times \mathbb{R}^+ \) of \(\tilde{X}\) and \(\bar{X}\) are qualitatively equivalent and
Thus, the set of singularities of vector field \( \bar{X} \) at the first quadrant of \( S^1 \) is
\[
sing \bar{X}(r, \theta) = \{(0, 0), (0, \frac{\pi}{4}), (0, \frac{\pi}{2})\}
\]
The respective Jacobian matrix in these points are
\[
D\bar{X}(0, \theta) = \begin{cases}
0 & \text{if } \theta = 0 \\
0 & \text{if } \theta = \frac{\pi}{4} \\
-BA & \text{if } \theta = \frac{\pi}{2}
\end{cases}
\]
(5)
Employing the blowing down, we can determine the behavior of system (3) at the origin, by means the transformation
\[
\Phi^{-1}: \mathbb{R}^2 \rightarrow \mathbb{R}^2
\]
\[
(r \cos \theta, r \sin \theta) \sim (u, v)
\]
obtaining that
\[
\Phi^{-1}: (r, 0) \sim (u, 0),
\]
\[
\Phi^{-1}: (r, \frac{\pi}{2}) \sim (0, v), \text{and}
\]
\[
\Phi^{-1}: (r, \frac{\pi}{4}) \sim \left( \sqrt{u^2 + v^2}, \arctan \frac{v}{u} \right) = (u, u).
\]
Therefore, the inclination angle of the separatrix curve is \( v = \tan \frac{\pi}{4} u \), and this curve divides the trajectories in the phase plane. Thus, there exists an hyperbolic sector and a parabolic sector at the neighborhood of the point \((0, 0)\).

This result implies that the point \((0, 0)\) is an attractor for all the trajectories which initial conditions are above the separatrix curve and for the trajectories with initial conditions below this curve the origin is a non-hyperbolic saddle point.

When there no exists a positive equilibrium point we have the following

**Lemma 5** If \( Q > \frac{(A+1)^2}{4} \), then the point \((0, 0)\) is globally asymptotically stable.

**Proof.** As the point \((1, 0)\) is always a saddle point and there no exists positives equilibrium point we have that \((0, 0)\) is an attractor for all the trajectories.

We recall the positives singularities lie over the curve \( v = u \); then, \( Q = (1 - u) (u + A) \) and we have the Jacobian matrix is:
\[
DY_\lambda(u, u) = \begin{pmatrix}
-u^2 (-A - 2u + 2Au + 3u^2) & -(1 - u) (u + A) u^2 \\
Bu (u + A) & -Bu (u + A)
\end{pmatrix},
\]
\[
\det DY_\lambda(u, u) = Bu^4 (A + u) (A + 2u - 1).
\]
a) If \( u > \frac{1-A}{2} \) imply that \( \det DY_\lambda(u, u) > 0 \), the behavior of singularity depends of sign of the \( \text{tr} DY_\lambda(u, u) \)

b) If \( u < \frac{1-A}{2} \) imply that \( \det DY_\lambda(u, u) < 0 \), the point \((u, u)\) is saddle.

Case 1. When there exist an equilibrium point \((E_c, E_c)\) by collapse of two of them we have that \( Q = \frac{(1+A)^2}{4} \).

**Theorem 6** The \((E_c, E_c)\) with \( E_c = \frac{1-A}{2} \), is a non-hyperbolic equilibrium point and is

i) a saddle-node attractor, if and only if, \( B > \frac{1-A^2}{4} \),

ii) a saddle-node repellor, if and only if, \( B < \frac{1-A^2}{4} \),

iii) a cusp point, if and only if, \( B = \frac{1-A^2}{4} \).

**Proof.** In the equilibrium point \((\frac{1-A}{2}, \frac{1-A}{2})\) the Jacobian matrix is

\[
DY_\lambda(E_c, E_c) = \begin{pmatrix}
\frac{1}{16} (A + 1)^2 (1-A)^2 & -\frac{1}{16} (A + 1)^2 (1-A)^2 \\
\frac{1}{16} (A + 1) (1-A) & -\frac{1}{4} (A + 1) (1-A)
\end{pmatrix}.
\]

Then, \( \det DY_\lambda(E_c, E_c) = 0 \), and

\[
\text{tr} DY_\lambda(E_c, E_c) = \frac{1}{16} (1-A) (A + 1) (1-A) (1-A^2 - 4B)
\]

The sign of \( \text{tr} DY_\lambda(E_c, E_c) \) depends of factor \( f(A, B) = 1 - A^2 - 4B \)

In the particular case that \( B = \frac{1-A^2}{4} \) the Jacobian matrix is

\[
DY_\lambda(E_c, E_c) = \frac{1}{16} (A + 1)^2 (1-A)^2 \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.
\]

The Jordan form matrix is [2]

\[
DY_\lambda(E_c, E_c) = \frac{1}{16} (A + 1)^2 (1-A)^2 \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.
\]

The Bogdanov-Takens bifurcation [7] is obtained, and \((E_c, E_c)\) is a cusp point. ■

Case 2. If \( Q = A \) and \( u_c = 1 - A > 0 \), also we have a unique positive equilibrium point. We note that in the case we have the two-parameter system given by:

\[
Y_\eta : \begin{cases}
\frac{du}{d\tau} = (1-u)u(u+A) - Av u^2 \\
\frac{dv}{d\tau} = Bv (u-v)(u+A)
\end{cases}
\]

**Theorem 7** In system (6), the unique positive equilibrium point \((1-A, 1-A)\) is:

a) a local attractor point, if and only if, \( B > (1-A)\left(A - (1-A)^2\right) \). Moreover, it is

a1) an attractor focus, if and only if, \( P < 0 \)

a2) an attractor node, if and only if, \( P > 0 \)

b) a repeller point, if and only if, \( B < (1-A)\left(A - (1-A)^2\right) \). Moreover, it is:

b1) a repeller focus, if and only if, \( P < 0 \)

b2) a repeller node, if and only if, \( P > 0 \)

c) a weak focus, if and only if, \( B = (1-A)\left(A - (1-A)^2\right) \)
\textbf{Theorem 8} In system $H$, it depends on the Jacobian matrix is

$$DY_{\lambda}(1 - A, 1 - A) = \begin{pmatrix} -(A^2 - 3A + 1)(1 - A)^2 & -A(1 - A)^2 \\ (1 - A)B & -(1 - A)B \end{pmatrix}.$$ 

Then, $\det(DY_{\lambda}(1 - A, 1 - A) = B(1 - A)^5 > 0$ and the nature of $(1 - A, 1 - A)$ depends on

$$\text{tr}(DY_{\lambda}(1 - A, 1 - A) = (1 - A) ((1 - A) (-A^2 + 3A - 1) - B).$$

The sign of $\text{tr}(DY_{\lambda}(1 - A, 1 - A)$ depends on factor

$$T = (1 - A) \left( A - (1 - A)^2 \right) - B.$$ 

Moreover, considering

$$P = (\text{tr}(DY_{\lambda}(1 - A, 1 - A))^2 - 4\det(DY_{\lambda}(1 - A, 1 - A)$$

$$= ((1 - A) ((1 - A) (-A^2 + 3A - 1) - B))^2 - 4B(1 - A)^5,$$

determines if $(1 - A, 1 - A)$ is node and focus. 

\textbf{Case 3.} Supposing that

$$u_1 = \frac{1}{2} \left( 1 - A - \sqrt{(1 + A)^2 - 4Q}\right) < 0,$$

we have a unique equilibrium point given by

$$u_2 = \frac{1}{2} \left( 1 - A + \sqrt{(1 + A)^2 - 4Q}\right) = H.$$ 

Then, assuming that $Q < A$, we obtain $(1 - H) (H + A) < A$, i.e., $H (1 - A - H) < 0$. Clearly $H > \frac{1 - A}{2}$, and system (3) becomes

$$Y_{p} : \begin{cases}
\frac{du}{dt} = ((1 - u)(u + A) u - (1 - H)(H + A) v) u^2 \\
\frac{dv}{dt} = Bv (u - v)(u + A)
\end{cases}$$

(7)

The Jacobian matrix is

$$DY_{p}(H, H) = \begin{pmatrix} -H^2\left( -A - 2H + 2AH + 3H^2 \right) & -(1 - H)(A + H)H^2 \\ (A + H)HB & - (A + H)HB \end{pmatrix},$$

\text{with,}

$$\det(DY_{p}(H, H)) = BH(4 (A + H) (A + 2H - 1))$$

and

$$\text{tr}(DY_{p}(H, H)) = -(A + H)B + H (A + 2H) + H (2 - 3H)$$

Let $P = (\text{tr}(DY_{p}(H, H))^2 - 4\det(DY_{p}(H, H))$


\textbf{Theorem 8} In system (7), the unique equilibrium point $(H, H)$ is:

a) a local attractor, if and only if, $B < \frac{H(A(1 - 2H) + H(2 - 3H))}{(A + H)}$. Moreover, $(H, H)$ is

a1) an attractor node, if and only if, $P > 0$,

a2) an attractor focus, if and only if, $P < 0$,

b) a weak focus, if and only if, $B = \frac{H(A(1 - 2H) + H(2 - 3H))}{(A + H)}$,

c) a repeller, if and only if, $B > \frac{H(A(1 - 2H) + H(2 - 3H))}{(A + H)}$. Moreover, $(H, H)$ is
c1) a repeller node, if and only if, \( P > 0 \); then, \((H, H)\) is surrounded by a limit cycle, 
c2) a repeller focus, if and only if, \( P < 0 \).

**Proof.** Clearly, \( \text{det} DY_\rho(H, H) > 0 \), if and only if, \( H > \frac{1 - A}{2} \).

We have that 
a) \( \text{tr} \text{ } DY_\eta(H, H) < 0 \) and the point \((H, H)\) is a local attractor, if and only if, 
\[ B > H \left( A(1 - 2H) + H(2 - 3H) \right) \frac{A}{A + H} \]

b) Assuming that \( \text{tr} \text{ } DY_\eta(H, H) = 0 \), then \( B = \frac{H(1 - 2H) + H(2 - 3H)}{A + H} \) and the point \((H, H)\) is a weak focus.

c) Supposing that \( \text{tr} \text{ } DY_\eta(H, H) > 0 \), we have that, 
\[ B < H \left( A(1 - 2H) + H(2 - 3H) \right) \frac{A}{A + H} \]

\((H, H)\) is a focus or node depending on the value of \( P \). □

**Case 4.** Now we consider the case in which there exists two singularities, at interior of the first quadrant as establish lemma 3.2. As it was seen in lemma it must fulfill that \( A < Q < (A + 1)^2 / 4 \). The nature of \((u_2, u_2) = (H, H)\) is the same that the above case 3, i.e., can be a local attractor, a repeller or a weak focus.

**Theorem 9** The equilibrium point \((u_1, u_1) = (E, E)\) is an hyperbolic saddle.

**Proof.** The determinant of Jacobian matrix evaluated at \((E, E)\) is 
\[ \text{det} \text{ } DY_\rho(E, E) = BE^4(A + E)(A + 2E - 1) \]

As \( E = \frac{1}{2} \left( 1 - A - \sqrt{(1 + A)^2 - 4Q} \right) < \frac{1 - A}{2} \)
we have \( \text{det} \text{ } DY_\rho(E, E) < 0 \), and the singularity is saddle point. □

**Theorem 10** Existence of homoclinic

a) There exist conditions on the parameter values for which an homoclinic curve is determined by the stable and unstable manifold of point \((u_1, u_1)\).

b) There exists an unstable limit cycle that bifurcates of the homoclinic surrounding the point \((u_2, u_2)\).

**Proof.** We denote by \( W^\pm_+(u_1, u_1) \), the under and stable manifold; \( W^\pm_-(u_1, u_1) \), the upper and stable manifold; \( W^\pm_-(u_1, u_1) \), the right and stable manifold and \( W^\pm_-(u_1, u_1) \), the left unstable manifolds of \((u_1, u_1)\), respectively.

As \( \tilde{\Gamma} \) is an invariant region the trajectories can not cross the line \( u = 1 \) towards the right. For the existence and uniqueness theorem, the trajectory determined by \( W^\pm_+(u_1, u_1) \), it can not intersects the trajectory determines by the above stable manifold \( W^\pm_-(u_1, u_1) \), whose \( \alpha – \text{limit} \) must be the point \((u_2, u_2)\) or lie at infinity over \( y – \text{axis} \).

Then, the \( \omega – \text{limit} \) of the trajectory determined by \( W^\pm_+(u_1, u_1) \) must be 
a) the point \((u_2, u_2)\), when this point is an attractor or else a stable (attractor) limit cycle,
b) the point $(0, 0)$, when the point $(u_2, u_2)$ is a repellor and the $\alpha - limit$ of the above stable manifold $W_s^u (u_1, u_1)$.

Then, there exists a set of parameter values for which the trajectories determine by $W^u (u_1, u_1)$ and $W^s (u_1, u_1)$ intersects forming the homoclinic curve.

4 Conclusions

In this work, we analyzed a Holling-Tanner predator-prey model [16, 17], an special case of Leslie-Gower type model, considering that the prey population is affected by a weak Allee effect or non-critical depensation.

By means a diffeomorphism we analyze a topologically equivalent system depending of three parameter to establish the local stability of the equilibrium points.

We show that the dynamics of the model with Allee effect differs of the May Holling-Tanner model [2] studied in [22], for which the Allee effect is not considered since this model can have two equilibrium points at interior of the first quadrant, one of which always is saddle point.

Also, we have that the point $(0, 0)$ is nonhyperbolic singularity that determines a separatrix curve that divides the behavior of trajectories, and the point $(1, 0)$ is always a saddle point for all parameter values.

The existence of the separatrix implies that the trajectories are highly sensitive to initial conditions some of which having to $(0, 0)$ as its $\omega-limit$ and other a stable positive equilibrium point or a stable limit cycle.

If the ratio prey-predator is high (many prey and little predator) then the populations can coexist for a wide set of parameter values, but this ratio is low then there exists a great possibility of two populations can go to extinction.

We conjecture of this model has two limit cycles for a certain parameter values and that the model considering strong Allee effect has similar dynamics since the number of positive equilibrium points are the same of the model here analyzed.

Acknowledgements

The authors wish to thank to the members of the Mathematical Ecology Group of the Instituto de Matemáticas at the Pontificia Universidad Católica de Valparaíso, Chile, for their valuable comments and suggestions. This work was partially financed by Fondecyt 1120218 and DIEA PUCV 124.730/2012 projects.
References


Holling-Tanner model with weak Allee effect


Solving the HP Protein Folding Problem by an Evolutionary Algorithm

J.M. García-Martínez¹, E.M. Garzón¹, J.M. Cecilia², H. Pérez-Sánchez² and P.M. Ortigosa¹

¹ Department of Informatics, ceiA3, University of Almería, Spain
² Bioinformatics and High Performance Computing Research Group (BIO-HPC), Computer Science Department, Universidad Católica San Antonio de Murcia (UCAM), Spain
emails: chema.garc@gmail.com, gmartin@ual.es, jmcecilia@ucam.edu, hperez@ucam.edu, ortigosa@ual.es

Abstract

This work applies an evolutionary algorithm, UEGO, to solve the protein structure optimization problem based on the HP model. To develop the UEGO solution of HP Protein Folding Problem it is necessary to identify the UEGO parameters and to design effective local optimizers in the context of the protein model. This work develops and analyzes the UEGO solution of HP Protein Folding.

Key words: Protein Folding, HP model, Global Optimization, Hybrid Evolutionary Algorithm

1 Introduction

Protein structure optimization is a well-known problem in bioinformatics. One of the most widely studied models of protein folding is the hydrophobic-hydrophilic (HP) model introduced by Dill [5]. In the HP model, chains of amino acids are configured as self-avoiding walks on the 3D cubic lattice (ie., adjacent amino acids of each chain lie on adjacent lattice sites, and no site is occupied by more than one amino acid). Based on the assumption that the hydrophobic reactions make an important contribution to the free energy of the embedding, a protein is modeled as a specific sequence of hydrophobic (H for nonpolar) or hydrophilic (P for polar) monomers. An optimal conformation maximizes the number
of adjacencies between Hs. So, this model is translated in a NP complete optimization problem, as it is shown in [3].

Due to the relevance of protein structure optimization and the effectiveness of HP model an intensive research work in this line has been recently developed. Therefore, several approaches based on the application of different optimization methods are described in the literature including Monte Carlo methods [16], evolutionary algorithm [15], ant colony optimization algorithm [6, 11] and particle swarm optimization [8].

However, nowadays the design effective and efficient solutions of the HP Protein Folding problem continues being an interesting challenge. In this work we explore the use of Universal Evolutionary Global Optimizer (UEGO), an evolutionary algorithm [9], for solving the HP protein folding problem. UEGO has shown its effectiveness for solving different specific optimization problems [14, 10, 12, 13]. Besides, it can be accelerated on several kinds of High Performance Platforms [7, 1] and its structure allows to include specific local search procedures in order to improve the search in particular scenarios. However, it is fair to underline that for solving a particular real problem by UEGO, such as HP Protein Folding, a twofold effort is necessary: (1) to identify the UEGO parameters in the context of the particular problem and (2) to design an effective local optimizer. This work is focussed on the definition of UEGO solution of HP Protein Folding.

2 HP Protein Folding problem

HP Protein Folding model defined by Dill [5] has been widely used for for predicting protein structures [4, 2]. The HP model represents every protein sequence as a string \( A = a_1a_2a_3\ldots a_n \), where \( a_i \in \{H, P\} \) and \( 1 \leq i \leq n \). A conformation of \( A \) is defined by a sequence of fold directions starting from the lattice site occupied by the first amino acid residue \( a_1 \). The protein conformations are restricted to self-avoiding paths on \( D \)-dimensional sequence lattice, where \( D = 2 \) or 3. The protein structure prediction assumes that the native state of the protein is defined by lowest Gibbs free energy. In the HP model, the energy of a conformation is defined as a number of topological contacts between hydrophobic amino acid that are not neighbors in the given sequence.

Therefore, the protein structure prediction is translated to an optimization problem as follows: Given an amino acid sequence \( A = a_1a_2a_3\ldots a_n \) find an energy minimizing conformation.

3 UEGO for solving HP Protein Folding problem

UEGO [9] is a multimodal algorithm which is able both to solve multimodal optimization problems where the objective function has multiple local optima and to discover the structure of these optima as well as the global optimum (see Algorithm 1 for a global description...
Algorithm 1: The UEGO algorithm

1: Init_species_list
2: Optimize_species($n_1$)
3: for $i = 2$ to levels do
4: Determine $r_i$, $new_i$, $n_i$
5: Create_species($new_i/S_i$)
6: Fuse_species($r_i$)
7: Shorten_species_list($M$)
8: Local_Optimize_species($n_i/M$)
9: Fuse_species($r_i$)
10: Shorten_species_list($M$)
11: end for

of UEGO).

A species $s$ can be thought of as a window on the whole search space. This window is defined by its center $C$ and a radius $r_i$ that is associated to the level or iteration $i$ in which the species has been created. The center is a solution, and the radius indicates its attraction area which covers a region of the search space and hence, multiple solutions. The role of this window is to ‘focus’ the optimizer that is always called by a species so that it can only ‘see’ its own window, so every new sample is taken from there.

During the optimization process, a list of species is kept by UEGO, which is in fact a method for managing this species-list. ‘Species-management’ consists of procedures for creating, fusing and eliminating species during the whole optimization process (procedures Create_species, Fuse_species, Shorten_species_list in Algorithm 1, respectively). The maximum length of the species list is given by the input parameter $M$ and the number of existing species at level $i$ is denoted by $S_i$.

The maximum number of function evaluations for the whole optimization process is given by the input parameter $N$. In addition, each level $i$ has two restrictions on the number of function evaluations (f.e.), namely $new_i$ (maximum f.e. allowed when creating new species) and $n_i$ (maximum f.e. allowed when optimizing individual species). These parameters are related by: $N = \sum_{i=1}^{levels} (n_i + new_i)$. For further details about the mathematical expressions of $n_i$ and $new_i$ see [9].

In UEGO every species is intended to occupy a local maximizer of the fitness function, without knowing the total number of local maximizers in the fitness landscape. This means that when the algorithm starts it does not know how many species there will be at the end.

Additionally, UEGO is a hybrid algorithm that introduces a local optimizer into the evolution process (procedure Local_Optimize_species in Algorithm 1). In this way, at every generation, UEGO performs a local optimizer operation on each species, and these locally optimal solutions replace the caller species. UEGO is abstract in the sense that the ‘species-
management’ and the cooling mechanism have been logically separated from the actual local optimization algorithm. Therefore, it is possible to implement any kind of optimizer to work inside a species.

This work is focussed on the UEGO solution of the HP protein folding problem. As above mentioned, the distance between two elements into search space is a key of UEGO. So, in the set of conformations of a protein sequence, $A$, we define the distance function between two conformations, $u$ and $v$, as

$$d(u, v) = \frac{1}{n} \sum_{i=1}^{n} \|u_i - v_i\|^2$$

where $u_i$ and $v_i$ represent the position vectors of the $i$ amino acid in both conformations.

UEGO is used for solving the folding problem according to this definition of distance. The optimum configurations computed by UEGO are analyzed against previous results obtained by Peña et al. [11].

**Acknowledgements**

This work has been funded by grants from the Spanish Ministry of Science and Innovation (TIN2012-37483-C03-03), Junta de Andalucía (P10-TIC-6002, P12-TIC-301) in part financed by the European Regional Development Fund (ERDF). Moreover, it has been developed in the framework of the network High Performance Computing on Heterogeneous Parallel Architectures (CAPAP-H4), supported by the Spanish Ministry of Science and Innovation (TIN2011-15734-E).

**References**


Extraneous attractors for Chebyshev’s method

M. García-Olivo$^1$, J. M. Gutiérrez$^2$ and Á. A. Magreñán$^3$

$^1$ Department of Mathematics, Instituto Tecnológico de Santo Domingo, Dominican Republic
$^2$ Department of Mathematics and Computer Sciences, University of La Rioja, Spain
$^3$ Department of Mathematics, International University of La Rioja, Spain
emails: martin1matdr@hotmail.com, jmguti@unirioja.es, alberto.magrenan@gmail.com

Abstract

The aim of this paper is to show the existence of extraneous fixed points for Chebyshev’s method applied to complex polynomials. This fact is a distinguishing feature in the dynamical study of Chebyshev’s method compared with other known iterative methods as Newton’s or Halley’s methods. In addition, in this work we consider other dynamical aspects of the method as, for instance, the universal Julia set, the dynamical plane or the space of parameters.

Key words: Chebyshev’s method, nonlinear equations, iterative methods, complex dynamics.
MSC 2000: AMS codes (optional)

1 Introduction

Chebyshev’s method is a well-known iterative method for solving nonlinear equations. Although it can be used for solving equations in different spaces (real line, $n$-dimensional Euclidean space, Banach spaces, etc.) in this work we consider it as a method for solving equations in the complex plane $f(z) = 0$, with $f : \mathbb{C} \to \mathbb{C}$. Starting from a given point $z_0 \in \mathbb{C}$, the method generates a sequence $\{z_n\}$, $n \geq 0$, defined by the following iteration function:

$$z_{n+1} = C_f(z_n), \quad C_f(z) = z - \left(1 + \frac{1}{2} \frac{f(z)f''(z)}{f'(z)^2}\right) \frac{f(z)}{f'(z)}.$$  (1)
The out-coming sequence can be seen as the orbit \( \{ z_0, C_f(z_0), C_f^2(z_0), \ldots, C_f^n(z_0), \ldots \} \) of the initial seed \( z_0 \). Therefore, the method can be seen as a discrete dynamical system and not only as a root-finding method. In this paper we take into account this point of view and we are interested in analyzing the dynamical properties of the method (1) and not only the conditions that give rise to convergence to one of the roots of the considered equation.

Historically speaking there is a certain controversy on the authorship of the method (1). Some authors attribute it to Euler but some others to Chebyshev (see [6], [9] for instance). In [7] we can find a possible solution for this academic polemics: the method, in the sense given by (1), was firstly used by Chebyshev. But the method also appears as a particular case of a family of root-finding method used by Euler to solve some polynomial equations. Chebyshev’s method is a cubically convergent algorithm constructed upon the idea of higher order inverse Taylor series. It also has a geometrical derivation using osculating parabolas.

As in the historical case, the dynamical study of Chebyshev’s method has been considered as a part of wider studies involving families of iterative processes (see [7], [3] or [4] for instance). In addition, in [5] Chebyshev’s method (under the name of super-Newton’s method) is analyzed both from a numerical and dynamical point of view, with especial emphasis in quadratic polynomials. One of the main conclusions of the latter paper is that Chebyshev’s method has general convergence for quadratic polynomials, that is for almost every initial point and for almost every polynomial of the given degree, the iterative method (1) converges to a root. However Chebyshev’s method is not generally convergent for cubic or higher order polynomials. This feature is shared with other classical methods (Newton’s or Halley’s methods), but the reasons for this lost of general convergence are not exactly the same. Whereas in the case of Newton’s and Halley’s methods the general convergence is lost because of the appearance of attractive cycles, in the case of Chebyshev’s method there are also extraneous fixed points, that is fixed points of the iteration function defined in (1) that are not roots of the equation \( f(z) = 0 \).

In this work we suppose the reader has a certain knowledge of many of the basic dynamical concepts we are going to use (fixed points, cycles and its character, critical points, basins of attraction, conjugacy, etc.). Otherwise we recommend the unfamiliar reader to look up the aforementioned references or the reference book of Beardon [2].

For our convenience in the rest of the paper we can re-write the iteration function (1) and its derivative in the following form

\[
C_f(z) = z - \left( 1 + \frac{1}{2} L_f(z) \right) \frac{f(z)}{f'(z)}, \quad C_f'(z) = \frac{L_f(z)^2}{2} (3 - L_f(z)) ,
\]

(2)

where

\[
L_f(z) = \frac{f(z)f''(z)}{f'(z)^2}, \quad L_f'(z) = \frac{f'(z)f'''(z)}{f''(z)^2}.
\]

(3)

As we refer them along this paper, we finish this section with the iteration functions of Newton’s \( N_f(z) \) and Halley’s \( H_f(z) \) method (quadratically and cubically convergent
respectively)

\[ N_f(z) = z - \frac{f(z)}{f'(z)}, \quad H_f(z) = z - \left( \frac{2}{2 - L_f(z)} \right) \frac{f(z)}{f'(z)}. \]  

(4)

2 Chebyshev’s method on quadratics

With the aim of understanding the dynamics of Chebyshev’s method in the complex plane, we start by analyzing its behavior for quadratic polynomials with two distinct roots

\[ f(z) = (z - a)(z - b), \quad a \neq b. \]  

(5)

We follow the steps given by Cayley in 1879 (see [5] for instance) in the study of Newton’s method. Cayley showed that the Julia set of the rational map \( N_f(z) \) arising from the application of Newton’s method to (5) is a straight line corresponding to the locus of points equidistant from the two roots. In particular, he proved that \( N_f(z) \) is conjugated to \( N(z) = z^2 \) via the Möbius transform

\[ M(z) = \frac{z - a}{z - b}, \]  

(6)

that is \( N(z) = M \circ N_f \circ M^{-1}(z) \). It is easy to check that the locus of points equidistant from the two roots maps onto the unit circle under (6).

In the same way, [5], Halley’s method on quadratics is conjugated to the map \( H(z) = z^3 \). The Julia set of the corresponding rational map \( H_f(z) \) is again the locus of points equidistant from the two roots.

Let us see what happens for Chebyshev’s method. Firstly we obtain that the rational map for Chebyshev’s method (1) applied to (5) is

\[ C_f(z) = \frac{-ab(a^2 + 3ab + b^2) + 6ab(a + b)z - 6abz^2 - 2(a + b)z^3 + 3z^4}{(-a - b + 2z)^3}. \]  

(7)

The rational map defined in (7) has the following properties:

i. It is a fourth degree rational map depending on two parameters: the roots of the polynomial, \( a \) and \( b \).

ii. It has 5 fixed points: \( a, b, \infty, p_1, p_2 \) where

\[ p_1 = \frac{(5 + \sqrt{5})a + (5 - \sqrt{5})b}{10}, \quad p_2 = \frac{(5 - \sqrt{5})a + (5 + \sqrt{5})b}{10}, \]

with the characters given in Table 1. We highlight the presence of 3 extraneous fixed points, all of them repelling.
Table 1: Classification of the fixed points of the iteration map $C_f(z)$ defined in (7).

<table>
<thead>
<tr>
<th>Fixed point</th>
<th>Character</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>Superattracting</td>
<td>0</td>
</tr>
<tr>
<td>$b$</td>
<td>Superattracting</td>
<td>0</td>
</tr>
<tr>
<td>$\infty$</td>
<td>Repelling</td>
<td>8/3</td>
</tr>
<tr>
<td>$p_1$</td>
<td>Repelling</td>
<td>6</td>
</tr>
<tr>
<td>$p_2$</td>
<td>Repelling</td>
<td>6</td>
</tr>
</tbody>
</table>

iii. Taking into account (2)) for quadratic polynomials, we deduce that

$$C'_f(z) = \frac{3}{2} L_f(z)^2 = 6 \frac{f(z)^2}{f'(z)^4}.$$  

Then the roots of $f(z)$ are critical points of $C_f(z)$, (with multiplicity 3). In addition, $(a + b)/2$ is another critical point with multiplicity 3.

Instead of studying directly the iteration map (7), we consider the rational map obtained under conjugation with (6):

$$S(z) = M \circ C_f \circ M^{-1}(z) = \frac{z^3(z + 2)}{2z + 1}. \tag{8}$$

The Möbius transform $M$ introduced in (6) maps the root $a$ into 0 and the root $b$ into infinity. In addition, there are other 3 extraneous fixed points, as shown in Table 2. The main properties of function $S(z)$ are:

i. It is a forth degree rational map that does not depend on any parameter.

ii. It has five fixed points: $z = 0$, $z = \infty$ and the three extraneous fixed points $z = 1$, $z = (-3 - \sqrt{5})/2$ and $z = (-3 + \sqrt{5})/2$. Their characters are shown in Table 2.

iii. As

$$S'(z) = \frac{6z^2(z + 1)^2}{(2z + 1)^2},$$

we have that 0 and $-1$ are critical points of $S(z)$ with multiplicity 3. In addition, $\infty$ is a critical point of $S(z)$ with multiplicity 3.

The following result extends Proposition 3.3.3 given by Kneisl in [5]. It gives, in an analytical way, some of the properties of the iteration map $S(z)$ defined in (8) and, by extension, some dynamical properties of Chebyshev’s method on quadratics. This result can be graphically interpreted by means of Figure 1.
Table 2: Classification of the fixed points of the iteration map $S(z)$ defined in (8).

<table>
<thead>
<tr>
<th>Fixed point</th>
<th>Character</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Superattracting</td>
<td>0</td>
</tr>
<tr>
<td>$\infty$</td>
<td>Superattracting</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>Repelling</td>
<td>8/3</td>
</tr>
<tr>
<td>$(-3 - \sqrt{5})/2$</td>
<td>Repelling</td>
<td>6</td>
</tr>
<tr>
<td>$(-3 + \sqrt{5})/2$</td>
<td>Repelling</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 1: On the left, basins of attraction of the rational function $S(z)$ defined in (8). On the right, basins of attraction of Chebyshev’s method applied to $p(z) = z^2 - 1$. The border between the basins in the first graphic is the Universal Julia set for Chebyshev’s method on quadratics. The second graphic is conformally equivalent to the first one. It allow us to appreciate one property of Chebyshev’s method: there are points closer to one of the roots whose orbits converge to the other root. This property is not shared with Newton’s or Halley’s method.

Theorem 1 The Fatou set, $F(S)$, related with the map $S(z)$ defined in (8) is formed by two basins of attraction, corresponding to the two superattracting fixed points 0 and $\infty$. In addition, $F(S)$, has got infinite components. The unit circle is contained in the Julia set of $S(z)$, $J(S_2)$ and its Lebesgue measure is zero.

Proof. Most of these properties are shown in in [5]. We focus our attention in showing the existence of an infinite number of components in $F(S)$.

Let us denote $F_0$ the immediate basin of attraction of 0. Then $F_0 \subseteq F(S)$, $F_0$ is
contained into the unit disc $D^1 = \{ z \in \mathbb{C}, |z| < 1 \}$ and $F_0 \cap S^1 = \emptyset$, where $S^1 = \{ z \in \mathbb{C}, |z| = 1 \}$ is the unit circle.

As $S(-2) = 0$, there exists another component of $F(S)$, $F_1$, that contains to $-2$. Note that $S(F_1) = F_0$ and then the orbits of $z \in F_1$ converge to 0. In addition $F_1 \cap S^1 = \emptyset$ and $F_1 \cap F_0 = \emptyset$.

Let us denote $F_\infty$ the immediate basin of attraction of $\infty$. As $\infty$ is a superattracting fixed point of $S$, if $z \in F_\infty$, then $S^n(z) \to \infty$ when $n \to \infty$.

Consequently the Fatou set $F(S)$ has at least three disjoint components: $F_0$, $F_1$ y $F_\infty$. As it is known that the number of components of the Fatou set of a rational map is 0, 1, 2 or infinity (see [2, Thm. 5.6.2]), we deduce that $F(S)$ has an infinite number of components.

3 Chebyshev’s method on cubics

As in the case of Newton’s method, the application of Chebyshev’s method to cubic polynomials deserves interesting dynamical questions. Setting aside the case of cubic polynomials with a triple root (in general, Chebyshev’s method applied to polynomials in the form $p(z) = (z - a)^n$ with $a \in \mathbb{C}$, $n \in \mathbb{N}$, is conformally equivalent to the function $(2n^2 - 3n + 1)/(2n^2)z$, whose dynamics are trivial), we focus our attention in cubic polynomials with a double root:

$$p(z) = (z - a)^2(z - b), \quad a, b \in \mathbb{C}, \quad a \neq b.$$  (9)

In this case we can follow the steps given in the previous section. Firstly, the related rational function is

$$C_p(z) = \frac{15z^4 + A_3z^3 + A_2z^2 + A_1z + A_0}{(-a - 2b + 3z)^3},$$  (10)

where we have denoted

$$A_0 = -ab(a^2 + 6ab + 5b^2),$$
$$A_1 = 3ab(3a + 5b) - 3b^3,$$
$$A_2 = -3b(2a - 5b),$$
$$A_3 = -7a - 26b.$$

As in the quadratic case, instead of studying the previous function (10) depending on the roots $a$ and $b$, we consider the rational function obtained after conjugation with the M"obius transform defined in (6):

$$T(z) = M \circ C_p \circ M^{-1}(z) = \frac{z(z^3 + 5z^2 + 6z + 3)}{7z + 8}.$$  (11)

The M"obius transform $M$ maps the double root $a$ into 0 and the simple root $b$ into $\infty$. The iteration function (11) has the following properties:
Table 3: Classification of the fixed points of the map $T(z)$ defined in (11).

<table>
<thead>
<tr>
<th>Fixed point</th>
<th>Character</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Attracting</td>
<td>$3/8$</td>
</tr>
<tr>
<td>$\infty$</td>
<td>Superattracting</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>Repelling</td>
<td>$9/5$</td>
</tr>
<tr>
<td>$-1$</td>
<td>Repelling</td>
<td>9</td>
</tr>
<tr>
<td>$-5$</td>
<td>Repelling</td>
<td>$49/9$</td>
</tr>
</tbody>
</table>

i. It is fourth order rational map free on parameters.

ii. It has 5 fixed points with the characters shown in Table 3. Note that the 3 extraneous fixed points are repelling.

iii. The critical points are $-2$ (with multiplicity 2), $\infty$ (with multiplicity 3), $(-3 - i\sqrt{5})/7$ and $(-3 + i\sqrt{5})/7$. Note that

$$T'(z) = \frac{3(z + 2)^2 (7z^2 + 6z + 2)}{(7z + 8)^2}, \quad 1/T(1/z) = \frac{z^3(8z + 7)}{3z^3 + 6z^2 + 5z + 1}.$$ 

The Fatou and Julia sets related to function (11) are shown in Figure 2. $\infty$ and 0, as superattracting and attracting fixed points respectively, have their own basins of attraction (in black and white respectively). The behavior of the rest of critical points is the following: the orbits of $(-3 \pm i\sqrt{5})/7$ converge to the fixed point $z = 0$; $T(-2) = 1$, that is a repelling fixed point and then belong to the Julia set of $T(z)$. Consequently there is not more basins of attraction. A graphical inspection to the Fatou set of $T(z)$ reveals that it has infinite many components. However, in this case, is not easy to obtain an invariant curve that plays the role of the unit circle $S^1$ in the quadratic case. In the right figure we can see as the basin of attraction of the multiple root (in white) “invades” the basin of the simple root, as it was pointed out in [8]. Then the symmetry respect the imaginary axis shown in the quadratic case is lost.

Now we consider a generic cubic polynomial, that is a polynomial with its three roots different:

$$p(z) = (z - a)(z - b)(z - c).$$

It is known [5] that all the three methods considered in this paper has not general convergence. Then, for each of them, there exists a region $A$ in the complex plane, with positive measure, such that if $z_0 \in A$, the corresponding sequence does not converge to any of the roots of $p(z)$. We call these regions “black holes”.

Our aim now is to find “bad polynomials” for Chebyshev’s method in the following sense: its dynamics contain “black holes”. In addition, we show that the nature of these black
Figure 2: On the left, basins of attraction of the rational function $T(z)$ defined in (8). On the right, basins of attraction of Chebyshev’s method applied to $p(z) = (z - 1)^2(z + 1)$. The border between the basins in the first graphic is the Universal Julia set for Chebyshev’s method on cubic polynomials with a double root. The second graphic is conformally equivalent to the first one and allow us to appreciate the lost of symmetry in the basins.

The presence of black holes is different in Chebyshev’s method than in Newton’s or Halley’s methods. Actually, in the former methods black holes are motivated by the presence of attracting cycles (see [10] for instance). There is not attracting extraneous fixed points in these two methods. With more detail, $\infty$ is the only fixed point for Newton’s method that is different from the roots. But it is always repelling. For Halley’s method it is known ([5]) that the extraneous fixed points are repelling with multiplier $1 + 2/j$ for some $j \in \mathbb{N}$.

**Theorem 2** Let $\alpha \in \mathbb{C}$ be a fixed point of $C_p(z)$ the iteration map of Chebyshev’s method applied to $p(z)$. Let us assume the following conditions:

i. $L_p(\alpha) = -2$,

ii. $|3 - L_p'(\alpha)| < 1/2$.

Then $\alpha$ is a extraneous fixed point of $C_p(z)$. If $L_p'(\alpha) = 3$ $\alpha$ is a superattracting fixed point.

**Proof.** It is easy to verify that the fixed points of $C_p(z)$ are the roots of $p(z)$ and the solutions of $L_p(z) = -2$. In addition, as

$$C_p'(z) = \frac{L_p(z)^2}{2} - (3 - L_p'(z)),$$
and $L_p(\alpha) = -2$, we have

$$C'_p(\alpha) = 2(3 - L_p'(z)).$$

and the result holds. ■

This result allows us to characterize cubic polynomials with a superattracting extraneous fixed point at $z = 0$.

**Corollary 1** Let us assume the coefficients of the polynomial $p(z) = z^3 + az^2 + bz + c$ satisfy:

i. $b^2 + ac = 0$,

ii. $3|2a^2 - b| < a^2$.

Then Chebyshev’s method applied to the polynomial $p(z)$ has an extraneous fixed point at $z = 0$. If $b = 2a^2$, this point is superattracting.

**Proof.**

The result follows directly by applying Theorem 2 for $\alpha = 0$. In this case:

i. $L_p(0) = 2ac/b^2 = -2$,

ii. $|3 - L_p'(\alpha)| = |3(1 - b/2a^2)| < 1/2$. ■

Now we give a couple of examples to illustrate the previous results. Let us consider the polynomial $p_1(z) = z^3 + z^2 + 2z - 4$ and the corresponding iteration function for Chebyshev’s method:

$$C_{p_1}(z) = \frac{z^2(15z^5 + 26z^4 + 27z^3 + 75z^2 + 80z + 120)}{(3z^2 + 2z + 2)^3}.$$  

$z = 0$ is an superattracting extraneous fixed point ($p_1(0) \neq 0$, $C_{p_1}(0) = 0$ and $C'_{p_1}(0) = 0$).

Orbits of points close to the origin (more or less a ball centered in the origin with radius 0.08 as we can see in Figure 3) converge to the extraneous attractor. For instance, these are the first terms of the orbit of $z_0 = 0.08$:

{$0.08, 0.0758219, 0.0712853, 0.0637754, 0.0520671, 0.0357771, ...$}.

There are several uni-parametric families of cubic polynomials that represent (under conformal transforms) the behavior of an iterative method for cubic polynomials. In this paper we consider the one proposed by Vrscay and Gilbert ([12]) amongst other authors:

$$p_\lambda(z) = z^3 + (\lambda - 1)z - \lambda, \quad \lambda \in \mathbb{C}. \quad (12)$$
$p_\lambda(z)$ has the following roots: $z_1 = 1$, $z_2 = (-1-\sqrt{1-4\lambda})/2$ and $z_3 = (-1+\sqrt{1-4\lambda})/2$. Let us denote $C_\lambda(z)$ the iteration function of Chebyshev’s method applied to $p_\lambda(z)$ defined in (12). The solutions of

$$12z^4 + 9(\lambda - 1)z^2 - 3\lambda z + (\lambda - 1)^2 = 0$$

are the extraneous fixed points of $C_\lambda(z)$. In addition,

$$C'_\lambda(z) = \frac{3p_\lambda(z)^2(15z^2 - \lambda + 1)}{(3z^2 + \lambda - 1)^4},$$

and then the free critical points (critical points that differ from the roots of $p_\lambda(z)$) are

$$c_1 = \sqrt{\frac{\lambda - 1}{15}}, \quad c_2 = -\sqrt{\frac{\lambda - 1}{15}}.$$ (13)

Here we are considering the principal branch of the complex squared root.

Figure 3: On the left, parameter plane obtained for $c_1$ and the Chebyshev’s iteration function $C_\lambda(z)$, related with the cubic family of polynomials $p_\lambda(z)$. On the right a detail showing the appearance of extraneous fixed point (these figures also appear in [12]).

The parameter plane is a powerful graphical tool to analyze the dynamics of functions depending on parameters, as it is the case we are considering. A classical result by Fatou and Julia guarantees that every attracting cycle must attract at least one critical point. Then the orbits of the free critical points give us information about the existence of attractor different from the roots. For each free critical point, we can color a parameter plane by assigning to each $\lambda \in \mathbb{C}$ a different color depending if the orbit of the free critical point converges to one
M. García-Olivo, J. M. Gutiérrez, Á. A. Magreñán

of the roots. Black color is reserved for other situations (extraneous attractors, cycles, not convergence).

For instance, if we impose that $c_1$ is a root $Q_\lambda(z) = 12z^4 + 9(\lambda - 1)z^2 - 3\lambda z + (\lambda - 1)^2$, then the same $c_1$ is a extraneous fixed point.

$$Q_\lambda(c_1) = \frac{124}{75} (\lambda - 1)^2 - \sqrt[5]{\frac{3}{5}} \sqrt[5]{\lambda - 1} = 0.$$ 

Then for $\lambda \neq 1$, $Q_\lambda(c_1) = 0$ if $\lambda = 31/16$. Consequently, for the iteration function $C_{31/16}(z)$ obtained from the polynomial $p_{31/16}(z) = z^3 + 15/16z - 31/16$, has in $c_1 = 1/4$ an extraneous fixed point. Then around the point $\lambda = 31/16 \approx 1.9375$ in the parameter plane there must be a region colored with the color corresponding to a extraneous fixed point, as we see in Figure 3. This technique could be reproduced also for the second free critical point, $c_2$. In this case we obtain new extraneous fixed points at $\lambda = 4 \left(154 - 45\sqrt{3}i\right)/961 \approx 0.641 - 0.324i$ and $\lambda = 4 \left(154 + 45\sqrt{3}i\right)/961 \approx 0.641 + 0.324i$ (see Figure 4).

Figure 4: On the left, parameter plane obtained for $c_2$ and the Chebyshev’s iteration function $C_\lambda(z)$. On the right a magnification showing a Mandelbrot-like set produced by the presence of other extraneous fixed points.

As a conclusion, we have seen that Chebyshev’s method has not extraneous fixed point when it is applied to quadratic polynomials or cubic polynomials with a double root. In addition, we have proved in two constructive ways the existence of extraneous fixed points for Chebyshev’s method applied to generic cubic polynomials. These techniques could be extended for polynomials of higher degrees, but this target will be covered in further works.
Acknowledgements

This work has been partially supported by the research project Ref. MTM2011-28636-C02-01 by the Spanish Ministry of Science and Innovation. All the graphics in this paper has been made by using Mathematica 9 and modifying the instructions given by Varona [11].

References


Melting simulations of gallium clusters: transitions between the low temperature bulk phases

Nicola Gaston\textsuperscript{1} and Krista G. Steenbergen\textsuperscript{2}

\textsuperscript{1} MacDiarmid Institute for Advanced Materials and Nanotechnology, School of Chemical and Physical Sciences, Victoria University of Wellington, New Zealand
\textsuperscript{2} Department of Chemistry, The University of Kansas, U.S.A.
emails: nicola.gaston@vuw.ac.nz, kgsteen@gmail.com

Abstract

Elemental gallium is a molecular metal, a phenomenon which partly explains its low melting temperature. Small clusters of gallium, in contrast, have been shown to melt at much higher temperatures than the bulk metal. We complete an in-depth structural analysis at finite temperatures, based on the results of first-principles Born-Oppenheimer molecular dynamics simulations. Melting is preceded by a transition from either the $\delta$ or $\beta$ phases to the $\gamma$ phase. A structural feature is identified, that increases with the latent heat and appears throughout the observed phase changes of this curious metal. We suggest a first explanation for the greater-than-bulk melting temperatures of the clusters through analysis of the isomers found in the liquid phase, and relate the driving force for the existence of these isomers to quantum size effects. Key words: gallium, melting, polymorphism, phase transitions

1 Introduction

The low melting temperature of elemental gallium is an anomaly of significant interest for our understanding of metallic systems. Called a ‘molecular metal’ due to the coexistence of covalently-bound dimers with the metallicity of the $\alpha$-phase, gallium also adopts a range of low-temperature structures $\beta$, $\gamma$, $\delta$, and $\epsilon$, as well as a range of structures found under pressure. The high-pressure allotropes, in contrast to the low-temperature structures, become denser and more metallic with pressure, as may be expected.

However, none of the extensive bulk phase-diagram of gallium can yet explain the discovery of greater-than-bulk melting in small clusters of gallium, at sizes ranging from
Melting simulations of gallium clusters 17 to 55 atoms[1, 2, 3]. Numerous experimental and theoretical studies have searched for the origin of this phenomenon, which contravenes the understood variation of melting temperature with size, known as melting point depression.

Previous studies have demonstrated that first principles Born-Oppenheimer MD simulations can accurately reproduce the experimentally observed variation of specific heat with temperature, which suggests that the questions of how these clusters melt, and why the melting temperatures are elevated, are able to be answered. However, previous studies of melting of these clusters have compared melting characteristics only to the global minimum structure, without fully addressing structural changes that take place at finite temperature, both below and above the melting temperature.

In our previous work [4, 5, 6] we have demonstrated the ability of density functional theory calculations to reproduce the experimental findings, in particular the strong size-sensitivity of the melting temperatures.

2 Structural transitions

Extensive structural analysis at finite temperature has revealed a measure of local order that closely follows the latent heat of melting in small gallium cluster cations. This order parameter defines regions of local atomic clustering that counterintuitively increase with the latent heat. We describe this structural pattern as “Gangnam,” explained most simply as the structure obtained through the gesture of crossing bent wrists. It contains a central atom having 4 near-neighbor atoms with cross-central bond angles that are nearly, but not quite, planar. The correlation between this structural measure and features in the specific heat curves led us to investigate the structures that exist throughout the finite temperature trajectories.

Using the criterion of existence for >50% of the trajectory at a given temperature in the solid phase, we identify 2-3 stable, pre-melting structural transitions for each cluster size. The global minimum (GM) structures are consistently favoured at the lowest temperatures. However, significant geometric rearrangement at higher temperatures yielded stable configurations which are typically structured around a tetrahedral core and assume a more spherical geometry. At higher temperatures, up to \( T_m \), a tetrahedral-based spherical structure becomes the more favorable geometry. This structure strongly resembles the capped-sphere structure of \( \text{Ga}_{20}^+ \).[6] This capped-sphere structural motif is found for all clusters up to \( \text{Ga}_{35}^+ \) at temperatures close to melting.

The rich potential energy landscape for these small clusters is strongly reminiscent of the polymorphism known in bulk gallium. This raises the immediate question: Do these various cluster isomers correspond to different allotropes of the bulk metal? In order to explore this question, we compare bond signatures between the cluster structures and the low-temperature bulk phases known for gallium.
Using statistically robust methods of comparison[7] between the structures of the finite clusters and those known for the bulk, we have identified the cluster isomer structures as being characteristic of either the $\beta$, $\gamma$, or $\delta$ phases.

**Acknowledgements**

This work has been supported by the Marsden Fund of the Royal Society of New Zealand under contract IRL0801. We thank the New Zealand eScience Infrastructure (NeSI), particularly the BlueFern (University of Canterbury, nesi67) and Pan (University of Auckland, nesi7) supercomputer teams for computational time and support.

**References**


Comparing total and partial connections in a patched population model

Carlotta Gazzola\textsuperscript{1} and Ezio Venturino\textsuperscript{1}

\textsuperscript{1} Dipartimento di Matematica “Giuseppe Peano”, via Carlo Alberto 10, 10123 Torino, Italy, Università di Torino
emails: carlotta.gazzola@hotmail.com, ezio.venturino@unito.it

Abstract

The possible dynamics of an ecosystem with three interconnected patches among which one population can migrate are investigated, paying attention to the consequences that possible disruption in the communicating paths can cause.

Key words: refuge, niches, disease transmission, ecoepidemics
MSC 2000: AMS codes 92D25, 92D40

1 Introduction

Metapopulations are a tool for the theoretical investigation of fragmented landscapes, \cite{14,15}. Whether the fragmentation is due to human activity or natural causes like landslides, fires and so on, it might represent for the animal populations living in the interested ecosystem a possible danger. Basically, this framework consists in formulating models in which to the local population dynamics interpatch migrations are suitably added. Mathematical models are needed even by field ecologists, in view of the difficulty of gathering migration data. Persistence of wild populations in heterogeneous environments is of particular concern for conservationist biologists, \cite{24,25}.

Metapopulation dynamics has been successfully applied to model for instance the mountain sheep (\textit{Ovis canadensis}), \cite{6}, or the spotted owl (\textit{Strix occidentalis}), \cite{11}.

The study of communities living in separate environments that are connected by possible migration paths has led to some counterintuitive results, such as the fact that the global population thrives, while locally in some patches becoming extinct, \cite{5,10,13,26}. In this
setting, also recent works on the role of niches as safety refuges can be accounted for, [4, 7, 8, 9, 17].

More recently, [23], in the framework of modelling heterogenous habitats, also the influence of diseases have been introduced. The effects of diseases on communities are a fact in nature and therefore also demographic models cannot ignore them. In fact, since a quarter of a century now, models accounting for interacting populations where also epidemics spread have been proposed and investigated, see Chapter 7 of [16] and the papers [12, 2, 18, 19, 3, 1, 20, 21, 22].

In general, the mathematical models that are introduced in this context can contain several patches and are usually analysed for the existence of the equilibria and possibly for their stability. In this paper we want to consider a rather simple system, composed of three patches that are joined together by connecting directed paths. Our aim is to investigate how its behavior changes once some of these connections are broken, whether accidentally or, as mentioned above, due to human artifacts that partly or entirely disrupt these communications between these habitats.

The paper is organized as follows. The general model with all possible connections between the three environments is presented and analysed in the next Section, finding its possible equilibria and studying their stability. Section 3 contains the models in which some of the paths become unavailable for the animals interpatch movement. A final discussion of the ecological implications concludes the paper.

2 The general model

We consider here an environment made out of 3 patches that are interconnected for migrations of a population $P$, as depicted in the following diagram.

\[
\begin{align*}
\begin{array}{c}
\begin{array}{c}
\downarrow \downarrow \downarrow \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } \\
\text{ } 
\end{array}
\end{array}
\end{align*}
\]

The size of each subpopulation in each patch $k$ is denoted by $P_k$, $k = 1, 2, 3$. Assuming that migrations from each patch are possible in all directions toward both other patches, the model is

\[
\begin{align*}
\dot{P}_1 &= r_1 P_1 \left( 1 - \frac{P_1}{k_1} \right) + m_{12} P_2 + m_{13} P_3 - m_{21} P_1 - m_{31} P_1 \\
\dot{P}_2 &= r_2 P_2 \left( 1 - \frac{P_2}{k_2} \right) - m_{32} P_2 - m_{12} P_2 + m_{21} P_1 + m_{23} P_3 \\
\dot{P}_3 &= r_3 P_3 \left( 1 - \frac{P_3}{k_3} \right) + m_{32} P_2 - m_{13} P_3 - m_{23} P_3 + m_{31} P_1
\end{align*}
\]
All the parameters are implicitly assumed to be nonnegative. Each equation describes the population dynamics in each patch. Each subpopulation reproduces logistically, with parameters that are environment-dependent, namely net reproduction rate $r_i$ and carrying capacity $k_i, i = 1, 2, 3$. In addition, denoting by $m_{ij}$ the migration rates from patch $j$ into patch $i$, we assume that movements in between different patches depend directly from the population level in the outgoing patch.

For later stability analysis purposes, it is convenient also to consider the Jacobian $J$ of (2),

$$
\begin{bmatrix}
J_{11} & m_{12} & m_{13} \\
m_{21} & J_{22} & m_{23} \\
m_{31} & m_{32} & J_{33}
\end{bmatrix}
$$

with

$$
J_{11} = r_1 - 2\frac{r_1}{k_1} P_1 - m_{21} - m_{31},
J_{22} = r_2 - 2\frac{r_2}{k_2} P_2 - m_{12} - m_{32},
J_{33} = r_3 - 2\frac{r_3}{k_3} P_3 - m_{13} - m_{23}.
$$

### 2.1 Feasible equilibria

There are only two possible equilibria, the origin at which the ecosystem disappears, and possibly the coexistence equilibrium, in which all the patches are populated. We now prove its existence.

Solve for $P_3$ the first two equations of (2) obtaining two surfaces:

$$
P_3^{(1)} = \frac{-1}{m_{13}} \left[ r_1 P_1 \left( 1 - \frac{P_1}{k_1} \right) + m_{12} P_2 - m_{21} P_1 - m_{31} P_1 \right]
$$

$$
P_3^{(2)} = \frac{-1}{m_{23}} \left[ r_2 P_2 \left( 1 - \frac{P_2}{k_2} \right) - m_{32} P_2 - m_{12} P_2 + m_{21} P_1 \right]
$$

The restriction of $P_3^{(1)}$ to the $P_2 - P_3$ coordinate plane is a straight line with negative slope through the origin. Instead, the restriction of $P_3^{(1)}$ to the $P_1 - P_3$ coordinate plane is a convex parabola through the origin. This parabola has another zero at the point $P_3^{(0,1)} = k_1 r_1^{-1} [r_1 - (m_{21} + m_{31})]$. Hence a feasible branch emanates from this latter point, when $P_3^{(0,1)} \geq 0$, or from the origin in the opposite case. A similar result holds for $P_3^{(2)}$, where in this case the root is $P_3^{(0,2)} = k_2 r_2^{-2} [r_2 - (m_{12} + m_{32})]$ Note that $P_3^{(0,1)} \geq 0$ and $P_3^{(0,2)} \geq 0$ hold when respectively the following conditions are satisfied

$$
r_1 \geq m_{21} + m_{31} := M_1, \quad r_2 \geq m_{12} + m_{32} := M_2.
$$
Connections in a patched population model

To better study the problem, however, we consider the intersections of these surfaces with the horizontal planes, $P_3 = h \geq 0$. Again, two parabolae are found,

$$
\sigma_h : P_2 = -\frac{1}{m_{12}} \left[ r_1 P_1 \left( 1 - \frac{P_1}{k_1} \right) - (m_{21} + m_{31}) P_1 + m_{13} h \right]
$$

$$
\rho_h : P_1 = -\frac{1}{m_{21}} \left[ r_2 P_2 \left( 1 - \frac{P_2}{k_2} \right) - (m_{32} + m_{12}) P_2 + m_{23} h \right]
$$

with $\sigma_h$ being a convex function of $P_2$ and $\rho_h$ a convex function of $P_1$. Both have nonpositive values at the origin, so that their two roots are of opposite signs. Hence a feasible branch emanates from the positive root, when $h > 0$, or from the origin where this root degenerates for $h = 0$.

Both parabolae have only one branch that lies in the feasible orthant. Hence, the two curves must meet at exactly one point $Q_h$, with nonnegative coordinates. In particular on the $P_3 = 0$ plane, i.e. for $h = 0$, $Q_0$ at worst could coincide with the origin in case (4) both do not hold. Since $h$ is arbitrary, it follows that the two surfaces $P_3^{(1)}$ and $P_3^{(2)}$ meet along a line $\ell$ in the positive orthant:

$$
\ell = \{ Q_h \equiv \sigma_h \cap \rho_h : \forall h \geq 0 \}.
$$

We now consider the surfaces $\Sigma_{\pm} : P_3^{\pm} \equiv P_3^{\pm}(P_1, P_2)$ originating from the third equation of (2), given explicitly by

$$
P_3^{\pm} = \frac{k_3}{2r_3} \left[ r_3 - m_{13} - m_{23} \pm \sqrt{(r_3 - m_{13} - m_{23})^2 + 4 \frac{r_3}{k_3} (m_{32} P_2 + m_{31} P_1)} \right].
$$

Since the term under the square root exceeds the one outside it, for every possible value of $P_1 \geq 0$ and $P_2 \geq 0$ the surface $\Sigma_+$ is always nonnegative, while $\Sigma_-$ is always nonpositive. Hence the intersection of $\ell$ and $\Sigma_+$ in the first orthant is always guaranteed, which provides the unique feasible coexistence equilibrium.

In summary, we have shown the following result.

**Theorem** The coexistence equilibrium always exists.

### 2.2 Equilibria stability

The stability of the equilibria can be assessed rather simply by using Descartes’ rule of signs on the characteristic equation. The latter is the cubic

$$
\lambda^3 - \text{tr}(J)\lambda^2 + M_J \lambda - \det(J) = 0,
$$

where $M_J$ represents the sum of the principal minors of $J$ of order 2. To have all negative roots, we need the conditions

$$
\text{tr}(J) < 0, \quad M_J > 0, \quad \det(J) < 0.
$$
Let
\[ \Pi_i = r_i \left( 1 - \frac{2}{k_i} P_i \right) \]

Explicitly, they become
\[
\sum_{i=1}^{3} \Pi_i < m_{21} + m_{31} + m_{12} + m_{32} + m_{13} + m_{23} \tag{5}
\]
for the trace,
\[
\Pi_1 \Pi_2 + \Pi_1 \Pi_3 + \Pi_2 \Pi_3 + m_{31} m_{32} + m_{21} m_{32} + m_{13} m_{31} \\
+ m_{12} m_{23} + m_{31} m_{23} + m_{13} m_{21} + m_{21} m_{13} + m_{12} m_{23} + m_{32} m_{13} \\
> \Pi_1 (m_{32} + m_{12} + m_{13} + m_{23}) + \Pi_2 (m_{31} + m_{21} + m_{13} + m_{23}) \\
+ \Pi_3 (m_{21} + m_{31} + m_{12} + m_{32}) \tag{6}
\]
for $M_J$ and finally for the determinant we have
\[
\Pi_1 \Pi_2 \Pi_3 + \Pi_1 (m_{12} m_{13} + m_{12} m_{23} + m_{32} m_{13}) \\
+ \Pi_2 (m_{21} m_{31} + m_{21} m_{23} + m_{31} m_{23}) + \Pi_3 (m_{21} m_{32} + m_{31} m_{12} + m_{31} m_{32}) \\
> \Pi_1 \Pi_2 (m_{13} + m_{23}) + \Pi_1 \Pi_3 (m_{12} + m_{32}) + \Pi_2 \Pi_3 (m_{21} + m_{31}). \tag{7}
\]

For the origin, note the simplification $\Pi_i = r_i$. For the coexistence equilibrium, the above stability conditions are more involved to assess. Numerical simulations however reveal its stability.

### 3 The models with some broken paths

There are several situations that can arise, when due to human artifacts or some natural catastrophic events some of the connecting paths become unavailable for the population migrations. We avoid to consider the situations in which one or all the patches become isolated. In this case indeed the isolated subpopulation would thrive independently of the others, due to the intrinsic resources represented by each logistic model in the formulation of (2), and the remaining configuration is simply given by two possibly connected patches, and therefore it is very easy to analyse. There are thus nine possible situations.

In fact, we can remove one (directed) path between any of the 3 patches in just one way. Combinatorically indeed it does not make any difference among which nodes we decide to break the connection and furthermore also the direction of the removed arc is immaterial, since by relabeling the nodes we would end up with the same situation. The result is represented in the picture below.
Connections in a patched population model

\[ P_1 \overset{\Rightarrow}{\leftrightarrow} P_2 \quad (EX2) = (5) \]

\[ P_3 \]

We can then remove two edges in several ways. From the same nodes, we get the following configuration

\[ P_2 \quad P_3 \]

\[ P_1 \]

\[ P_2 \overset{\Leftarrow}{\leftrightarrow} P_3 \quad (EX3) \]

\[ P_1 \]

If they are removed from different connected nodes, there are three alternatives: either the node that is connected with both the other nodes by just one arc has the two edges one outgoing and one incoming, or both outgoing, or both incoming. The pictures below will better illustrate these 3 situations.

\[ P_3 \overset{\Rightarrow}{\leftrightarrow} P_2 \quad (EX7) \]

\[ P_1 \]

\[ P_3 \overset{\Rightarrow}{\leftrightarrow} P_2 \quad (EX8) \]

\[ P_1 \]

Next, we can remove 3 edges. If we remove one edge from each pair of nodes, the only alternative is the way in which the orientation is considered. We can either remove all the edges in the same direction, but in such case which direction is immaterial, by a suitable relabeling of the nodes, or one edge is removed in one direction and the remaining two
in the opposite one; again due to symmetries this leads to just one configuration. These alternatives are depicted below.

\[
P_3 \leftarrow P_2 \quad (EX1)
\]
\[
\downarrow \quad \uparrow
\]
\[
P_1
\]

\[
P_3 \leftarrow P_2 \quad (EX6)
\]
\[
\downarrow \quad \uparrow
\]
\[
P_1
\]

(13)

(14)

If we remove 2 edges connecting the same nodes and remove another one, apart from symmetries there are only two configurations possible, namely

\[
P_1 \quad P_2 \quad (EX2 \ NEW) = (5)
\]
\[
\downarrow \quad \uparrow
\]
\[
P_3
\]

\[
P_1 \quad P_2 \quad (EX7 \ NEW)
\]
\[
\downarrow \quad \uparrow
\]
\[
P_3
\]

(15)

(16)

Finally, we can remove 4 edges and no more, otherwise at least one node will be disconnected from the other ones. The possible system configurations are as follows:

\[
P_1 \rightarrow P_2 \rightarrow P_3 \quad (17)
\]

\[
P_1 \rightarrow P_2 \leftarrow P_3 \quad (18)
\]

\[
P_1 \leftarrow P_2 \rightarrow P_3 \quad (19)
\]

For the analysis of all these models with broken paths, we must set some of the migration rates to zero. We will then investigate whether new equilibria arise, and, if possible, whether the stability of the origin and of coexistence are altered in the new configurations.
3.1 The models (8), (9), (10), (13)

In general, in these three models, the only equilibria are those of the general model, as no other ones can arise. For (9) no changes are necessary in the proof of the coexistence equilibrium. In the other cases, the proof however requires some attention.

Specifically, consider first (8) where \( m_{23} = 0 \). Note that in this case the approach used to show the existence of the equilibrium fails. We can instead solve the first and third equilibrium equations for \( P_2 \) and intersect the corresponding surfaces \( P_2^{(1)}(P_1, P_3) \) and \( P_2^{(2)}(P_1, P_3) \) with the planes \( P_2 = h \) to get

\[
\alpha_h : P_3 = -\frac{1}{m_{13}} \left[ r_1 P_1 \left( 1 - \frac{P_1}{k_1} \right) - (m_{21} + m_{31}) P_1 + m_{12} h \right],
\]

\[
\beta_h : P_1 = -\frac{1}{m_{31}} \left[ r_3 P_3 \left( 1 - \frac{P_3}{k_3} \right) - m_{13} P_3 + m_{32} h \right].
\]

These parabolae are seen to intersect with each other along a line that itself intersects the remaining surface originating from the second equilibrium equation, as done in the proof of the Theorem.

For (10) instead we have \( m_{31} = m_{12} = 0 \), solving the first and second equilibrium equation as done in the Theorem, for \( P_3^{(1)} \) we obtain a parabolic cylinder. Its intersection with \( P_3 = h \) gives a straight line, and the latter always meets the parabola obtained intersecting \( P_3^{(2)} \) with \( P_3 = h \), as in the proof of the Theorem, and existence follows accordingly.

Finally for (13) in which \( m_{31} = m_{12} = m_{23} = 0 \), solving all equilibrium equations we obtain always parabolic cylinders, with each axis parallel to a different coordinate axis:

\[
\gamma_1 : P_3 = -\frac{1}{m_{13}} \left[ r_1 P_1 \left( 1 - \frac{P_1}{k_1} \right) - m_{21} P_1 \right],
\]

\[
\gamma_2 : P_1 = -\frac{1}{m_{21}} \left[ r_2 P_2 \left( 1 - \frac{P_2}{k_2} \right) - m_{32} P_2 \right],
\]

\[
\gamma_3 : P_2 = -\frac{1}{m_{32}} \left[ r_3 P_3 \left( 1 - \frac{P_3}{k_3} \right) - m_{13} P_3 \right].
\]

Intersecting \( \gamma_1 \) with the plane \( P_3 = h \) gives a straight line \( L_1 \) parallel to the \( P_2 \) axis and \( L_1 \cap \gamma_2 \) gives a point, and therefore we get the line in space parametrized by \( Q_{12}^{(1)}(h, \ell^h, h) \), which itself, as in the Theorem, must intersect the \( \gamma_3 \) surface, thereby providing the coexistence equilibrium.

Furthermore, the stability analysis hinges on the Jacobian (3), in which the above conditions make some simplifications. But even for the origin the stability conditions nevertheless remain quite involved. More specifically, for all these models we find that (5) becomes sharper, as the right hand side will become smaller. The left hand side contains only demographic parameters, and explicitly no migration rates. On the other hand it also
contains the population levels at equilibrium, which in turn depend on the migration rates. Changes in the latter could in principle bring the equilibrium populations up or down and therefore influence stability as well. Similar considerations hold for (6) and (7).

**Remark.** It is thus hard to state whether the stability conditions will be easier or more difficult to be satisfied. These general considerations hold also for all the other models with some broken paths, unless we explicitly present some further remarks.

### 3.2 The model (15)

Here we have $m_{12} = m_{21} = m_{32} = 0$. These simplifications do not harm the proof of the Theorem, so that the coexistence equilibrium is guaranteed to be feasible. In addition, however, they show that the origin is certainly unstable, as one eigenvalue for this equilibrium is $J_{22} = r_2 > 0$.

In addition, this model allows also the equilibrium $X = (0, k_2, 0)$, for which one eigenvalue is explicit, $-r_2 < 0$, and the remaining ones provide the stability conditions

\[
m_{31} + m_{32} + m_{13} > r_3 + r_1, \\
(m_{31} - r_1)(m_{32} + m_{13} - r_3) > m_{13}m_{31}.
\]

### 3.3 The models (11), (16)

For (11) we need to take $m_{21} = m_{23} = 0$. It follows immediately that

\[
P_2^* = \frac{k_2}{r_2} (r_2 - m_{12} - m_{32}).
\]

From the remaining equilibrium equations we discover that the two parabolae

\[
P_3 = \frac{1}{m_{13}} \left[ m_{31}P_1 - r_1P_1 \left( 1 - \frac{P_1}{k_1} \right) - m_{12}P_2^* \right], \\
P_1 = \frac{1}{m_{31}} \left[ m_{13}P_3 - r_3P_3 \left( 1 - \frac{P_3}{k_3} \right) - m_{32}P_2^* \right],
\]

are convex, with negative value at 0, so that they always meet in the first quadrant, thus providing the remaining components of the coexistence equilibrium.

Stability at coexistence comes just from the negativity of one explicit eigenvalue, providing

\[
m_{12} + m_{32} + 2\frac{r_2}{k_2} P_2^* > r_2,
\]
while the Routh-Hurwitz conditions on the remaining minor hold always true,

\[
\frac{r_1}{k_1} P_1^* + m_{12} \frac{P_2^*}{P_1^*} + m_{13} \frac{P_3^*}{P_1^*} + \frac{r_3}{k_3} P_3^* + m_{32} \frac{P_2^*}{P_3^*} + m_{31} \frac{P_1^*}{P_3^*} > 0, \\
\left(\frac{r_1}{k_1} P_1^* + m_{12} \frac{P_2^*}{P_1^*}\right) \left(\frac{r_3}{k_3} P_3^* + m_{32} \frac{P_2^*}{P_3^*}\right) + m_{31} \frac{P_1^*}{P_3^*} \left(\frac{r_1}{k_1} P_1^* + m_{12} \frac{P_2^*}{P_1^*}\right) \left(\frac{r_3}{k_3} P_3^* + m_{32} \frac{P_2^*}{P_3^*}\right) + m_{31} \frac{P_1^*}{P_3^*} \left(\frac{r_1}{k_1} P_1^* + m_{12} \frac{P_2^*}{P_1^*}\right) > 0.
\]

Stability at the origin is obtained by

\[ m_{12} + m_{32} > r_2, \quad m_{13} + m_{31} > r_1 + r_3, \quad r_1 r_3 > r_1 m_{13} + r_3 m_{31}. \] (22)

Further, for the model (16) it is enough to set \( m_{12} = 0 \) and all the above considerations still carry on to this case.

In these models there is one more feasible equilibrium in addition to origin and coexistence, namely \( Q_1 = (P_1^Q, 0, P_3^Q) \). To find it in both cases (11) and (16), we solve the first and third equilibrium equations of (2) to find

\[
P_3 = \frac{P_1}{m_{13}} \left[ m_{31} - r_1 \left( 1 - \frac{P_1}{k_1} \right) \right], \\
P_1 = \frac{P_3}{m_{31}} \left[ m_{13} - r_3 \left( 1 - \frac{P_3}{k_3} \right) \right].
\]

These are two convex parabolae, with roots at the origin and respectively at the points \( P_1^{(a)} = k_1 r_1^{-1} [r_1 - m_{31}] \) and \( P_3^{(a)} = k_3 r_3^{-1} [r_3 - m_{13}] \). These points are nonnegative if the conditions \( r_1 \geq m_{31} \) and \( r_3 \geq m_{13} \) hold. When at least one of these conditions holds sharply, then an intersection between the parabolae is guaranteed in view of their convexity. If instead both are equalities, then the parabolae are both tangent to the axes at the origin, and therefore one intersection is the origin itself and another one exists also in this case. When instead both are not satisfied, we need to compare the parabolae slopes at the origin to determine whether an intersection between their feasible branches exists. We find

\[
P_3'(P_1) = \frac{1}{m_{13}} \left( m_{31} - r_1 + 2 \frac{r_1}{k_1} P_1 \right), \\
P_1'(P_3) = \frac{1}{m_{31}} \left( m_{13} - r_3 + 2 \frac{r_3}{k_3} P_3 \right),
\]

and we must impose that

\[ P_3'(0) > (P_1^{-1})'(0), \]

in order to ensure that the branches meet in the first quadrant. This amounts to requiring

\[ r_1 (r_3 - m_{13}) - r_3 m_{31} > 0, \]
which is impossible, in view of the restrictions holding in this situation

\[ r_1 < m_{31}, \quad r_3 < m_{13}. \]  \hspace{1cm} (25)

Thus in this situation the point \( Q_1 \) is infeasible.

\( Q_1 \) is stable if the eigenvalue that is immediately found is negative, entailing

\[ r_2 < m_{12} + m_{32}, \]  \hspace{1cm} (26)

and for the model (16) this condition simplifies since \( m_{12} = 0 \). In fact, the remaining Routh-Hurwitz conditions stemming from a 2 by 2 reduced Jacobian \( \tilde{J} \) are satisfied,

\[-\text{tr} \tilde{J} = \frac{m_{13}}{P_1^Q} P_3^Q + \frac{r_1}{k_1} P_1^Q + \frac{m_{31}}{P_3^Q} P_1^Q + \frac{r_3}{k_3} P_3^Q > 0,\]

\[ \det \tilde{J} = \frac{m_{13} r_3}{k_3 P_1^Q} (P_3^Q)^2 + \frac{m_{31} r_1}{k_1 P_3^Q} (P_1^Q)^2 + \frac{r_1 r_3}{k_1 k_3} P_1^Q P_3^Q > 0.\]

Thus these models admit the origin, the patch-2-population-free point and coexistence as possible equilibria.

### 3.4 The model (12)

Here we set \( m_{21} = m_{31} = 0 \). For coexistence the approach of the general case still works, it only simplifies giving for \( P_3^{(2)} \) a cylinder with axis parallel to the \( P_1 \) coordinate axis. Stability at coexistence is guaranteed by the explicit eigenvalue,

\[ k_1 < 2P_1^*. \]  \hspace{1cm} (27)

**Remark.** Note that this is an eigenvalue also for the origin, thereby providing its instability, in view of \( k_1 < 0 \).

The remaining Routh-Hurwitz conditions are always satisfied, since they reduce to

\[ \frac{m_{23}}{P_2^*} P_3^* + \frac{r_2}{k_2} P_2^* + \frac{m_{32}}{P_3^*} P_2^* + \frac{r_3}{k_3} P_3^* > 0,\]

\[ \frac{m_{23} r_3}{k_3 P_3^*} (P_3^*)^2 + \frac{m_{32} r_2}{k_2 P_3^*} (P_2^*)^2 + \frac{r_2 r_3}{k_2 k_3} P_2^* P_3^* > 0.\]

In addition to origin and coexistence, here we find also the equilibrium \( M_2 = (k_1, 0, 0) \) which is clearly always feasible. For its stability, the Jacobian simplifies even further. The first eigenvalue is \(-r_1 < 0\), the Routh-Hurwitz conditions on the remaining ones give the stability conditions

\[ r_2 + r_3 < m_{12} + m_{32} + m_{13} + m_{23}, \]  \hspace{1cm} (28)

\[ (r_2 - m_{12})(r_3 - m_{13}) > (r_2 - m_{12}) m_{23} + (r_3 - m_{13}) m_{32}.\]

The former condition when becomes an equality gives rise to a Hopf bifurcation. Explicitly, this occurs for

\[ r_2 = r_2^* = m_{13} + m_{23} + m_{32} + m_{12} - r_3. \]  \hspace{1cm} (29)
3.5 The model (14)

Set $m_{21} = m_{31} = m_{23} = 0$ in (2). Note that the origin in this model is unstable, since one eigenvalue is $J_{11} = r_1 > 0$.

Coexistence can be calculated explicitly, to give

$$P_1^* = \frac{k_1}{2} \left[ 1 + \sqrt{1 + \frac{4}{k_1 r_1} (m_{12} P_2^* + m_{13} P_3^*)} \right],$$

$$P_2^* = \frac{k_2}{r_2} (r_2 - m_{32} - m_{12}),$$

$$P_3^* = \frac{k_3}{2r_3} \left[ r_3 - m_{13} + \sqrt{(r_3 - m_{13})^2 + \frac{4}{k_3} r_3 m_{32} P_2^*} \right].$$

Feasibility of the coexistence equilibrium is ensured just by

$$r_2 > m_{12} + m_{32}. \quad (30)$$

The eigenvalues become

$$J_{11} = -\frac{1}{P_1^*} (m_{12} P_2^* + m_{13} P_3^*) < 0, \quad J_{22} = -\frac{r_2}{k_2} P_2^* < 0, \quad J_{33} = -\frac{r_3}{k_3} P_3^* - \frac{m_{32}}{P_3^*} P_2^* < 0,$$

showing that it is always stable, when feasible.

In this case the two new equilibrium points arise $I_2 = (k_1, 0, 0)$, $I_3 = (\alpha, 0, \beta)$, with

$$\alpha = \frac{k_1}{2} \left( 1 + \sqrt{1 + \frac{4m_{13} k_3 (r_3 - m_{13})}{r_1 r_3 k_1}} \right), \quad \beta = \frac{k_3}{r_3} (r_3 - m_{13}),$$

feasible for

$$r_3 > m_{13}. \quad (31)$$

At equilibrium $I_2$ we find the eigenvalues $J_{11} = -r_1 < 0$, $J_{22} = r_2 - m_{12} - m_{32}$, $J_{33} = r_3 - m_{13}$. Stability conditions are therefore

$$r_2 < m_{12} + m_{32}, \quad r_3 < m_{13}. \quad (32)$$

When $I_2$ is stable, the equilibria $I_3$ and coexistence are infeasible. Thus at $r_2^+ = m_{12} + m_{32}$ and $r_3^+ = m_{13}$ there are two transcritical bifurcations, the first one taking $I_2$ into coexistence, the second one taking it into $I_3$.

At equilibrium $I_3$, the eigenvalues are

$$J_{11} = -\frac{r_1}{k_1} \alpha - m_{13} \frac{\beta}{\alpha} < 0, \quad J_{22} = r_2 - m_{12} - m_{32}, \quad J_{33} = -\frac{r_3}{k_3} \beta < 0.$$

Stability is ensured by the first condition (32). Thus stability of $I_3$ also prevents feasibility of the coexistence equilibrium. We have thus another transcritical bifurcation at $r_2^+ = m_{32} + m_{12}$ taking $I_3$ into coexistence.
3.6 The model (17)

Here $m_{13} = m_{31} = m_{12} = m_{23} = 0$. This implies that at the origin one eigenvalue is $J_{33} = r_3 > 0$ so that this equilibrium is unstable.

Coexistence can be stably attained, in view of the eigenvalues

\[ J_{11} = -\frac{r_1}{k_1}P_1^* < 0, \quad J_{22} = -\frac{r_2}{k_2}P_2^* - \frac{m_{21}}{P_2^*}P_1^* < 0, \quad J_{33} = -\frac{r_3}{k_3}P_3^* - \frac{m_{32}}{P_3^*}P_2^* < 0, \]

at the levels

\[ P_1^* = \frac{k_1}{r_1}(r_1 - m_{21}), \quad P_2^* = \frac{k_2}{2r_2}\left[ r_2 - m_{32} + \sqrt{(r_2 - m_{32})^2 + \frac{4}{k_2}r_2m_{21}P_1^*} \right], \]
\[ P_3^* = \frac{k_3}{2r_3}\left[ r_3 + \sqrt{r_3^2 + \frac{4}{k_3}r_3m_{32}P_2^*} \right]. \]

In this case we find also the equilibrium $W_2 = (0, 0, k_3)$ and $W_3 = (0, P_2^+, P_3^+)$, with

\[ P_2^+ = \frac{k_2}{r_2}(r_2 - m_{32}), \quad P_3^+ = \frac{k_3}{2r_3}\left[ r_3 + \sqrt{r_3^2 + \frac{4}{k_3}r_3m_{32}P_2^*} \right]. \]

At $W_2$ the eigenvalues are $J_{11} = r_1 - m_{21}$, $J_{22} = r_2 - m_{32}$, $J_{33} = -r_3 < 0$, providing stability when

\[ r_1 < m_{21}, \quad r_2 < m_{32}. \quad (33) \]

At $W_3$ the eigenvalues are $J_{11} = r_1 - m_{21}$ and

\[ J_{22} = -\frac{r_2}{k_2}P_2^+ < 0, \quad J_{33} = -\frac{m_{32}}{P_3^+}P_2^+ - \frac{r_3}{k_3}P_3^+ < 0, \]

giving stability when the first condition (33) holds.

3.7 The model (18)

When $m_{13} = m_{31} = m_{12} = m_{32} = 0$, one eigenvalue at the origin is $J_{22} = r_2 > 0$, showing its instability. Coexistence is allowed at the population values

\[ P_1^* = \frac{k_1}{r_1}(r_1 - m_{21}), \quad P_3^* = \frac{k_3}{r_3}(r_3 - m_{23}), \]
\[ P_2^* = \frac{k_2}{2r_2}\left[ r_2 + \sqrt{r_2^2 + \frac{4}{k_2}r_2(m_{21}P_1^* + m_{23}P_3^*)} \right]. \]

This equilibrium is feasible for

\[ r_1 > m_{21}, \quad r_3 > m_{23}. \quad (34) \]
The eigenvalues are always negative, so that it is always stable, when feasible:

\[ J_{11} = -\frac{r_1}{k_1} P_1^* < 0, \quad J_{22} = -\frac{r_2}{k_2} P_2^* - \frac{m_{21}}{P_2^*} P_1^* - \frac{m_{23}}{P_2^*} P_3^* < 0, \quad J_{33} = -\frac{r_3}{k_3} P_3^* < 0. \]

We find also the equilibrium \( X_1 = (0, k_2, 0) \), stable for

\[ r_2 < m_{21}, \quad r_3 < m_{23}, \tag{35} \]

and more equilibria with either patch 1 or patch 3 empty, namely \( X_2 = (P_1^X, P_2^X, 0) \), with \( P_1 = P_1^* \),

\[ P_2^X = \frac{k_2}{2r_2} \left[ r_2 + \sqrt{r_2^2 + \frac{4}{k_2} r_2 m_{21} P_1^X} \right]. \]

and \( Y_3 = (0, P_2^Y, P_3^Y) \), \( P_3^Y = P_3^* \),

\[ P_2^Y = \frac{k_2}{2r_2} \left[ r_2 + \sqrt{r_2^2 + \frac{4}{k_2} r_2 m_{23} P_3^Y} \right]. \]

The eigenvalues at \( X_2 \) are

\[ J_{11} = -\frac{r_1}{k_1} P_1^X < 0, \quad J_{22} = -\frac{r_2}{k_2} P_2^X - \frac{m_{21}}{P_2^X} P_1^X < 0, \quad J_{33} = r_3 - m_{23}, \]

giving stability for

\[ r_3 < m_{23}. \tag{36} \]

At \( Y_3 \) we have instead

\[ J_{11} = r_1 - m_{21}, \quad J_{22} = -\frac{r_2}{k_2} P_2^Y - \frac{m_{23}}{P_2^Y} P_3^Y < 0, \quad J_{33} = -\frac{r_3}{k_3} P_3^Y < 0 \]

and consequently the stability conditions become

\[ r_1 < m_{21}. \tag{37} \]

### 3.8 The model (19)

Finally we consider \( m_{13} = m_{31} = m_{21} = m_{23} = 0 \), with the eigenvalue \( J_{11} = r_1 > 0 \) at the origin, giving instability.

For coexistence we find the population levels

\[ P_2^* = \frac{k_2}{r_2} (r_2 - m_{12} - m_{32}), \]

\[ P_1^* = \frac{k_1}{2r_1} \left[ r_1 + \sqrt{r_1^2 + \frac{4}{k_1} r_1 m_{12} P_2^*} \right], \]

\[ P_3^* = \frac{k_3}{2r_3} \left[ r_3 + \sqrt{r_3^2 + \frac{4}{k_3} r_3 m_{32} P_2^*} \right]. \]
It is feasible for

\[ r_2 > m_{32} + m_{12} \]  

and when feasible it is always stable, since the eigenvalues are

\[
J_{11} = -\frac{r_1}{k_1} P^*_1 - \frac{m_{12}}{P^*_1} P^*_2 < 0, \quad J_{22} = -\frac{r_2}{k_2} P^*_2 < 0, \quad J_{33} = -\frac{r_3}{k_3} P^*_3 - \frac{m_{32}}{P^*_3} P^*_2 < 0.
\]

In addition to the origin and coexistence, we find \( Z_1 = (k_1, 0, 0) \), \( Z_2 = (0, 0, k_3) \) and \( Z_3 = (k_1, 0, k_3) \). The former two are unstable, one eigenvalue is positive, respectively \( J_{11} = r_1 > 0 \) and \( J_{33} = r_3 > 0 \). For the last equilibrium \( Z_3 \), the eigenvalues are \( J_{11} = -r_1 < 0 \), \( J_{22} = r_2 - m_{12} - m_{32} \), \( J_{11} = -r_3 < 0 \) giving stability for

\[ r_2 < m_{12} + m_{32}. \] (39)

\section{Discussion}

\subsection{The original model}

The findings of this paper show that coexistence can be attained always, for the general model and for all the other models in which some interconnecting paths become unpracticable. When it is feasible, and when the local stability analysis can be performed, it appears that whenever feasible, the coexistence equilibrium is also locally asymptotically stable. We conjecture that in such case it is also globally asymptotically stable, in view of most of the other results, including the transcritical bifurcations found in some of the reduced models.

Another good result from the conservationist point of view is that the ecosystem never disappears, in almost all the cases in which the stability conditions for the origin can be evaluated explicitly. In view of the logistic growth assumption for the populations in each patch, we conjecture that this result holds true also for the more interconnected models. A notable exception is given however by models (11) and (16), see conditions (22).

\subsection{The broken paths models}

For the models (8), (10), (13) our results are again good from both the conservationist point of view as well as for the development of human artifacts, because it shows that in these cases some of the migration paths can be removed without harming too much the whole ecosystem behavior. Its equilibria indeed remain the same of the original model (2), namely the origin and coexistence. However some changes occur in the stability conditions of these equilibria. Therefore changes leading to the situations modeled by systems (8), (10), (13) should be treated with care. It is also interesting to note that these results hold for models where either one, two or three arcs are removed, therefore they really depend...
on the configuration of the system, rather than on the number of allowed connecting paths between patches.

Note that the common characteristic of these models, which is not shared by all the other ones, is that there is always the possibility of cycling between all the patches, i.e. starting from patch 1, say, to go to patch 2 and then 3 and finally returning to patch 1. Thus, when this cycle can be performed, no equilibria other than survival in all patches is allowed, except possibly ecosystem disappearance.

In general, in all other models the allowed system stable configurations, apart from origin and coexistence, are those of the patches with incoming paths. For instance in model (19) the population cannot survive only in patch 2, since it contains only outgoing paths.

Models (11) and (12) are also interesting. The two interconnected patches are a possible system configuration in the former but not in the latter. The reason, once more, is the fact that in model (11) the two patches have both incoming paths from the remaining patch, while in (12) they are sources for the paths leading to the remaining patch. This feature is also present in model (15), where the interconnected patches cannot be stably present in the equilibrium configurations because one of the interconnected patches is the origin of an outgoing path. Therefore just the presence of just one such outgoing path is enough to destabilize an interconnection. Conversely, in model (16) the interconnected patches are stable, because the remaining patch is a source of an outgoing flow.

The remaining models share another interesting property. In model (14) there are two additional possible stable configurations. The patch with both incoming paths, which is expected in view of the considerations holding for the previous situations, and the configuration with this patch together with the intermediate node. In other words only the patch from which both paths are outgoing cannot be present in the stable configurations. A similar result holds in model (17), where only the source patch cannot be stable, with either the end sink patch or both the intermediate and the sink patches give rise to stable configurations. Model (18) combines these results, as the sink or either one of the other two patches can be stable. Model (19) again excludes the only patch that is a source for both paths connecting it to the other two patches, while both these two patches can be stable at the same time.

References


A Fuzzy Representation of Vehicle Trajectories using Motion Data from H264/AVC Video

J. Giralt\textsuperscript{1}, J. Moreno-Garcia\textsuperscript{1}, L. Jimenez\textsuperscript{1}, E. del Castillo\textsuperscript{1} and L. Rodriguez-Benitez\textsuperscript{1}

\textsuperscript{1} Information Systems and Technologies Department, University of Castilla-La Mancha (Spain)

emails: juan.giralt@uclm.es, juan.moreno@uclm.es, luis.jimenez@uclm.es, ester.castillo@uclm.es, luis.rodriguez@uclm.es

Abstract

A method to obtain and represent the trajectories of a vehicle in motion is proposed. The approach only uses as input data the motion vectors generated by a H264/AVC video encoder and the experimental videos were recorded from a moving vehicle. First, each frame of the sequence is processed by a statistical method of analysing the angle direction of the motion vectors. This process obtain a numerical value representing the vehicles movement. In a second stage, this value is represented in the fuzzy domain to obtain a linguistic description for the vehicle trajectory in every frame of the video. Finally, these linguistic values are compared between consecutive frames and if they are the same they will be grouped. Three video sequences with different characteristics were used in the experimentation phase. One of them is an on-board video captured from the World Rally Car (WRC) competition while the others are sequences from our own dataset.

Key words: motion vectors, H264/AVC, vehicle trajectories, driver assistance systems.

1 Introduction

Nowadays, there exists a great interest in the development of computer vision methods for driver assistance systems in order to improve vehicles safety [8]. This kind of systems needs to detect the motion objects in the scene like pedestrians, other vehicles, etc. In order to discriminate between the motion objects and the background elements, it is necessary to
estimate the vehicle ego-motion, i.e., the motion characteristics of the vehicle’s movement from where the video is recorded. The ego-motion concept consists of a set of measures, then, it is something more complete than the simple trajectory estimation proposed in this work. Anyway, the authors consider that the proposed approach is conceptually linked to the set of techniques that computes ego-motion. There exist two major approaches to compute ego-motion, those based on monocular cameras [10] and others that use stereo video cameras. We focus our interest in the first type of methods because the video capture system is more economical.

Another classification distinguishes between discrete and differential techniques [3], depending of they are based on a search of characteristics points or on the optical flow [4, 5, 7]. The motion vector field is a sparse and imprecise approximation of the optical flow, then the proposed technique can be considered as a differential one.

In general terms without focusing in driving assistance systems, Raudies et al. [9] reviewed the great variety of methods to compute and evaluate ego-motion. For example, Kang et al. [6] research is related to fuzzy logic. They used the ego-motion estimation for robot browsing and in the segmentation stage they extract objects of interest in the scene using type-2 fuzzy sets.

The rest of the paper is organized as follows. We present the concepts of macroblock and H264 motion vectors in Section 2. In Section 3 we describe a statistical method to obtain a numerical value representing the vehicles movement. In Section 4, the vehicle trajectories are represented using linguistic variables and we detail a comparison process between consecutive frames of these fuzzy representations. Finally, in Sections 5 and 6 we detail experimental results, conclusions and future works.

2 H264/AVC motion vectors

H.264/AVC is a standard for video compression that removes redundant information between consecutive frames using motion compensation of macroblocks. The encoder looks for a motion pattern between frames and represents this pattern as a two dimensional motion vector. The motion information represented by motion vectors is the input data of the method proposed in this work.

Only one motion vector is generated for a macroblock and there are different macroblocks sizes, known as well as decision modes. These sizes are 16x16, 16x8, 8x16 y 8x8 (pixelXpixel) and corresponds with decision modes 1, 2, 3 and 4, respectively. Processing only the motion vectors of macroblocks with a size equal to 8x8 is enough for our purposes. As it is shown in Figure 1, these motion vectors contain the same information of the others, even they provide additional information.

One major advantage of working on compressed domain is that the input data is very small compared to that methods that must process the information of every pixel of every
Table 1: Average number of motion vectors processed per frame.

<table>
<thead>
<tr>
<th>Decision Mode</th>
<th>Exp. 1</th>
<th>Exp. 2</th>
<th>Exp. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>716</td>
<td>376</td>
<td>1185</td>
</tr>
</tbody>
</table>

Formally, the motion vectors in a frame $F$ are represented by the expression:

$$ VM(F) = \{mv_i, i = 0, ..., n - 1\} $$

where $mv_i$ is the 4-tuple $(x, y, I_x, I_y)$, being $(x, y)$ the position of the macroblock in the frame and $(I_x, I_y)$ the Cartesian coordinates of the motion vector for the macroblock.
This representation of motion vectors using Cartesian coordinates is substituted by another representation using polar coordinates: $r$ and $\theta$. Then, now $mv_i$ is the 4-tuple $(x, y, r, \theta)$. For example, the motion vector $mv_i$: $(5, 18, -7, 4)$ is represented in polar coordinates as: $(5, 18, 8, 150^\circ)$. This change of representation is justified because our study extracts information from motion direction, so adding the angle in the motion vector representation will facilitate this process.

3 Motion direction analysis

A first approach to determine the vehicle trajectories was to discriminate between motion vectors representing movements to the left and vectors representing movements to the right in all the frames of the video sequence. Then Equations 1 and 2 are proposed, where the symbol $\#$ is the cardinal of the set, $R(F)$ means number of motion vectors with a direction to the right in the frame $F$ and $L(F)$ means numbers of motion vectors with a direction to the left in $F$.

$$R(F) = \#\{mv_i \in VM(F), 0^\circ \leq \theta \leq 90^\circ \text{ y } 270^\circ \leq \theta \leq 360^\circ\}. \quad (1)$$

$$L(F) = \#\{mv_i \in VM(F), 90^\circ < \theta < 270^\circ\}. \quad (2)$$

First experimental results showed that $R(F)$ increased its value with respect to $L(F)$ when the vehicle was turning to the right. If the vehicle was driving through a bend to the left $L(F)$ increased its value with respect to $R(F)$. For example, Figure 2.a shows a frame containing a bend to the left, the results for Equations 1 and 2 are $(L(F)=476, R(F)=246)$. The values for the bend to the right showed in Figure 2.b are $(L(F)=107, R(F)=950)$.

Figure 2: Turning Left (a) and Turning Right (b)
Unfortunately, we observed how the increases and decreases of $L(F)$ or $R(F)$ depending on the vehicle trajectory did not happen as we expected. Concretely in those frames with the presence of moving or static objects with big size. Figure 3 shows an example of this kind of situation where a truck is present in the frame and the trajectory of the vehicle is straight. $L(F)$ is 141 and $R(F)$ is 213 and using the previous reasoning it was determined that the vehicle was turning to the right. So, Equations 1 and 2 were not useful in these situations. Then, we must compute the vehicle’s movement using techniques that are robust to situations where the number of motion vectors in a frame suddenly increases normalizing the differences between motion vectors of each direction. The proposed measure is called Curvature, $C(F)$, and it is defined in Equation 3.

$$C(F) = (R(F)/L(F)) - (L(F)/R(F)).$$

Figure 3: Truck detection during straight motion.

Figure 4 shows a graphic with the values of $C(F)$ in a subset of frames corresponding to Experiment 1 (Section 5). The greater the value of $C(F)$ is, the sharper is the turn to the right. The lower the value of $C(F)$ is, the sharper is the turn to the left. Mean values of $C(F)$ represent driving straight. For example, $C(F)$ value for Figure 2.a was -1.371 (Turning Left), for Figure 2.b was 14.65 (Turning Right) and for Figure 3 was 4.439 (Straight).

Now, our objective is to obtain automatically a correspondence between $C(F)$ and a vehicle trajectory: Turning Left (TL), Straight (S) and Turning Right (TR). We compute two values to split the domain of the variable $C(F)$ in three intervals that correspond to TL, S and TR. These values are called $T_{min}$ and $T_{max}$ and were empirically determined observing
the maximum and minimum of $C(F)$ in a sequence of frames where the vehicle’s movement is straight in a straight road segment. For example, in the graph of Figure 4 the values for the thresholds $T_{\text{min}}$ and $T_{\text{max}}$ are -0.8 and 2.8, respectively and they are represented in the graph by two horizontal lines. Now, considering these thresholds, Equation 4 obtains a description in natural language of the trajectory of the vehicle in a frame $F$. It is called Trajectory ($T(F)$).

$$T(F) \leftarrow \begin{cases} 
\text{Turning Left, if } C(F) < T_{\text{min}} \\
\text{Straight, if } T_{\text{min}} \leq C(F) \leq T_{\text{max}} \\
\text{Turning Right, if } C(F) > T_{\text{max}} 
\end{cases}$$

(4)

### 4 Processing data with fuzzy techniques

From the first experimental results obtained from the method proposed in Section 3, it was observed that a set of common errors appear frequently and recurrently. These errors could be produced because when we established how to compute $C(F)$ by means of Equation 4 some factors might have not been taken into account. More concretely, specific aspects of processing data in compressed domain and managing data related to the motion of a vehicle. As the input signal is from a physical data source it must satisfy some constraints of continuity, for example, due to inertia the motion of the vehicle cannot change instantaneously. But the individual analysis of each frame made possible the appearance of results like this: $(frame_i, TL), (frame_{i+1}, TR)$, that did not satisfy this continuity assumption. Also, as the videos selected for the experiments were captured from a camera in motion, static and moving objects produced motion vectors. $T(F)$ was very sensible to the appearance of new objects in the scene because in these situations the value of $C(F)$ could change slightly. Then, when $C(F)$ is near to the threshold values ($T_{\text{min}}, T_{\text{max}}$) the vehicle motion could be characterized incorrectly.
For that reasons, two new processes are presented. First one represents in the fuzzy domain the data obtaining a linguistic description for the vehicle trajectory in a frame and the second one process simultaneously several frames obtaining an unique linguistic description of the vehicle’s movement in this set of frames.

4.1 Definition of the linguistic variable Fuzzy Trajectory (FT)

Now, we define a linguistic variable [11] called Fuzzy Trajectory (FT) represented by three trapezoidal fuzzy sets, each one corresponding to a value of \( T(F) \) (Equation 4). The universe of discourse is taken from the motion information stored in each frame, although to simplify we assume that the physical domain is \( \mathbb{R} \). Table 2 shows the values \((a, b, c, d)\) that define each trapezoid where \( k \) is empirically determined from the minimum and maximum values of \( C(F) \) in each video.

<table>
<thead>
<tr>
<th>Label</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL</td>
<td>(-\infty)</td>
<td>(-\infty)</td>
<td>(U_{min} - k)</td>
<td>(U_{min} + k)</td>
</tr>
<tr>
<td>S</td>
<td>(U_{min} - k)</td>
<td>(U_{min} + k)</td>
<td>(U_{max} - k)</td>
<td>(U_{max} + k)</td>
</tr>
<tr>
<td>TR</td>
<td>(U_{max} - k)</td>
<td>(U_{max} + k)</td>
<td>(\infty)</td>
<td>(\infty)</td>
</tr>
</tbody>
</table>

4.2 Fuzzification of data and clustering

Now, we describe the process to fuzzify \( C(F) \) into linguistic concepts using \( FT \). First, for every frame, \( C(F) \) is computed and fuzzified and the results within the number of frame are stored in a tuple called \( FT(F) \):

\[
FT(F) = (F, \mu_{TL}(C(F)), \mu_{S}(C(F)), \mu_{TR}(C(F)))
\]

For example, if the frame is the number 31 and the membership values to \( FT \) are \( \mu_{TL}(C(31)) = 0, \mu_{S}(C(31)) = 0.524 \) and \( \mu_{TR}(C(31)) = 0.476 \) then \( FT(31) \) is \((31, 0, 0.524, 0.476)\). The video is represented as an ordered list of tuples \( FT(F) \).

Now, we propose to process \( m \) consecutive frames instead of individual frames. \( m \) were empirically determined observing the average module of the motion vectors, that is, \( m \) is related to the velocity of the vehicle where the camera is installed. That is because with greater velocities \( m \) should be smaller because the changes in the trajectory of the vehicle occurs in a smaller number of frames. Processing every consecutive \( m \) frames, \( m \) consecutive tuples \( FT(F) \), we obtain an unique value that will represent the trajectory of the vehicle during these \( m \) frames. This value is computed by means of a membership histogram and
Equations 5 to 7 show how it is computed as a average membership degree for a set of frames in the ordered list Vehicle Trajectories, where the first frame is the number \( j \) and the last one in the set is the frame number \( j + m - 1 \).

\[
Av(TL_j) \leftarrow \frac{\sum_{i=j,j+m-1} \mu_{TL}(C(i))}{m} \\
Av(S_j) \leftarrow \frac{\sum_{i=j,j+m-1} \mu_{S}(C(i))}{m} \\
Av(TR_j) \leftarrow \frac{\sum_{i=j,j+m-1} \mu_{TR}(C(i))}{m}
\] (5)  (6)  (7)

Now, the linguistic label with higher average value is selected to represent the vehicle trajectory in the \( m \) frames. The grouped information is stored in an ordered list called Vehicle Trajectories Video (VTV) containing tuples called MaxFT where:

\[
MaxFT = (frame_{\text{ini}}, \ frame_{\text{end}}, \ Label)
\]

For example, the tuple representing a straight trajectory between the frame number 31 and the frame number 60 is \( MaxFT = (31, 60, S) \) being \( m \) for this example 30. Now, we group the consecutive tuples MaxFT in Vehicle Trajectories Video with the same value for the element Label. Next example shows a VTV obtained from a video with 2420 frames and it can be observed how there is a straight motion from frame 0 to 109 and the vehicle is turning to the left between frames 110 and 289, etc.

\[
VTV = \{(0, 109, S), (110, 289, TL),...,(2300, 2419, S)\}
\]

5 Experiments

Three video sequences with very different characteristics were processed. In the Experiment 1, the vehicle drives in a completely straight road, the traffic density is very low and there are barely another different objects than cars, like trees, houses, etc.. The video of the Experiment 2 was captured from an onboard camera of a rally car competition. In this case, there are sudden changes in direction at high speeds and there a lot of vegetation in the ditch and in the road edge because the car drives along a rural minor road. Finally, in the Experiment 3 the car drives along a road with slight bends to the left and to the right. Table 3 shows the number of frames processed in each video and the average number of motion vectors processed per frame. Table 4 shows the parameter values for the three experiments.

Tables 5 and 6 present the list Vehicle Trajectories Video (VTV) obtained for Experiments 1 and 2, respectively. As it can be observed, the tables contain only a part of the
Table 3: Number of frames and average of motion vectors.

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th>Exp. 2</th>
<th>Exp. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frames</td>
<td>2479</td>
<td>2479</td>
<td>5284</td>
</tr>
<tr>
<td>Vectors</td>
<td>716</td>
<td>376</td>
<td>1185</td>
</tr>
</tbody>
</table>

Table 4: Parameter values.

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th>Exp. 2</th>
<th>Exp. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{min}$</td>
<td>-0.8</td>
<td>0</td>
<td>-0.3</td>
</tr>
<tr>
<td>$T_{max}$</td>
<td>2.8</td>
<td>4.8</td>
<td>1.8</td>
</tr>
<tr>
<td>$k$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$m$</td>
<td>10</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 5: $VTV$ obtained from Experiment 1

<table>
<thead>
<tr>
<th>$frame_{ini}$</th>
<th>$frame_{end}$</th>
<th>Result Label</th>
<th>Expected Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>110</td>
<td>283</td>
<td>TL</td>
<td>TL</td>
</tr>
<tr>
<td>284</td>
<td>327</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>328</td>
<td>458</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>459</td>
<td>479</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>481</td>
<td>490</td>
<td>TL</td>
<td>TL</td>
</tr>
<tr>
<td>491</td>
<td>534</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>535</td>
<td>545</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>546</td>
<td>687</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>688</td>
<td>698</td>
<td>TL</td>
<td>S</td>
</tr>
<tr>
<td>699</td>
<td>709</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>710</td>
<td>752</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>753</td>
<td>796</td>
<td>TL</td>
<td>TL</td>
</tr>
<tr>
<td>797</td>
<td>807</td>
<td>S</td>
<td>TR</td>
</tr>
<tr>
<td>808</td>
<td>839</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>841</td>
<td>999</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

list and there is a column that shows the expected result. This desired result was manually labelled. Analysing in detail these tables, it can be observed in Table 5 an error in the interval (688, 698). The error is due to the appearance of lines all road width that generates a motion vector field modifying the expected results from equations. It also appears in the frames 752 and 753 a change between the labels $TR$ and $TL$. Really, this appears
because the car is driving through a roundabout and this change cannot be considered an error. In Experiment 2 (Tabla 6), the main errors are that vehicles’ straight movements are labelled as “Turning Right”. This is due the fact the car continuously drives from one lane to another.

<table>
<thead>
<tr>
<th>frame_init</th>
<th>frame_end</th>
<th>Result Label</th>
<th>Expected Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>17</td>
<td>32</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>33</td>
<td>65</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>66</td>
<td>70</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>71</td>
<td>130</td>
<td>S</td>
<td>TL</td>
</tr>
<tr>
<td>131</td>
<td>136</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>137</td>
<td>147</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>148</td>
<td>169</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>170</td>
<td>179</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>180</td>
<td>196</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>197</td>
<td>207</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>208</td>
<td>234</td>
<td>TL</td>
<td>TL</td>
</tr>
<tr>
<td>235</td>
<td>239</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>240</td>
<td>305</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>306</td>
<td>316</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>317</td>
<td>327</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>328</td>
<td>359</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>360</td>
<td>370</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>371</td>
<td>387</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>388</td>
<td>398</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>399</td>
<td>447</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>448</td>
<td>458</td>
<td>TR</td>
<td>TR</td>
</tr>
<tr>
<td>459</td>
<td>572</td>
<td>S</td>
<td>TR</td>
</tr>
<tr>
<td>573</td>
<td>583</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>584</td>
<td>659</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>660</td>
<td>665</td>
<td>TL</td>
<td>TL</td>
</tr>
<tr>
<td>666</td>
<td>698</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>699</td>
<td>703</td>
<td>TL</td>
<td>S</td>
</tr>
<tr>
<td>704</td>
<td>752</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>753</td>
<td>763</td>
<td>TR</td>
<td>S</td>
</tr>
<tr>
<td>764</td>
<td>769</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Finally, Table 7 shows percentage of hits and errors for every experiment. The difficulty of Experiment 2 makes the hits rate is the lowest of all experiments but encouraging results have been obtained in Experiments 1 and 3.
Table 7: Final results

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th>Exp. 2</th>
<th>Exp. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>93.7</td>
<td>77.6</td>
<td>98.1</td>
</tr>
<tr>
<td>Errors</td>
<td>6.3</td>
<td>22.4</td>
<td>1.9</td>
</tr>
</tbody>
</table>

6 Conclusions and Future Works

In this work we have presented a new technique to represent in a linguistic way the trajectories of a vehicle in a video sequence using as input data a very little amount of data compared to methods than needs to use all the pixel information. The use of fuzzy logic allows to work with linguistic representations and to process information from several frames simultaneously.

As future work, the operation parameters should be generated in an automatic way. Another research line could be to combine our results with another information like the road geometry to prevent risk situations when there is no correspondence between the vehicle’s movements and the road shape.

Acknowledgements

This work has been supported by Information Systems and Technologies Department, University of Castilla-La Mancha.

References


Characteristic times for multiscale diffusion of active ingredients in coated textiles

T. Goessens\textsuperscript{1} and D. Constales\textsuperscript{1}

\textsuperscript{1} Department of Mathematical Analysis, Research Group NaM\textsuperscript{2}, Faculty of Engineering and Architecture, Ghent University, Krijgslaan 281 - S22, Ghent, Belgium

emails: tineke.goessens@ugent.be, denis.constales@ugent.be

Abstract

A three-scale approach for textile models was given in [1]: a one-dimensional fiber model and a room model, with a meso-level in between, which is the yarn scale. To analyse and simplify the model, its characteristic times are investigated here. At these times the fiber and yarn model, and the yarn and room model, respectively, tend to reach an equilibrium concentration. The identification of these characteristic times is key in reducing the model to its variously scaled components when simplifying it.

\textbf{Key words:} diffusion; textile modeling; multiscale modeling; controlled release; characteristic time

1 Introduction

We focus on the diffusion of a substance to the outer boundary of textiles. The fibers used to construct this fabric are coated with a polymer solution of an active ingredient (AI), e.g. an insect repellent, a perfume or a healing substance. This substance can easily be replaced by other volatiles. The goal is to investigate how much of the AI has to be present on the textile fiber and which polymer substance to use to coat the fiber so that the concentration at the outer boundary of the textile stays high enough for as long as required to be effective (e.g. repel or even kill mosquitoes, spread a noticeable odor for humans, have a healing effect ...).

The application in mind has the purpose to track the diffusion of an active component released by the fibers of an open textile structure, like a woven scrim, e.g. a gauze bandage. Models and algorithms for this application were discussed in [1, 2, 3] where a meso-level
Characteristic times for multiscale diffusion

model that describes the release of the active component in the yarn cross-section is included in between the standard fiber model and the room model. Upscaling from one level to another is done by volume averaging or overlapping domain decomposition.

2 Characteristic times for the three-level diffusion

The governing system of equations of the complete three-level model is

\[
\begin{align*}
\frac{\partial C_f(r,t)}{\partial t} &= \frac{1}{r} \frac{\partial}{\partial r} \left( r D_f \frac{\partial C_f(r,t)}{\partial r} \right), \quad r \in [R_f, 2R_f] \\
\frac{\partial C_y(\tilde{r},t)}{\partial t} &= \frac{1}{\tilde{r}} \frac{\partial}{\partial \tilde{r}} \left( \epsilon \tilde{r} D_y \frac{\partial C_y(\tilde{r},t)}{\partial \tilde{r}} \right) + \Gamma_{\text{in}}(\Omega_f^y, t) - \Gamma_{\text{out}}(\Omega_o^y, t), \quad \tilde{r} \in [0, 2R_y] \\
\frac{\partial C_{\text{room}}(x,t)}{\partial x} &= \frac{\partial}{\partial x} \left( D \frac{\partial C_{\text{room}}(x,t)}{\partial x} \right) + \Gamma_s(x,t), \quad x \in [R_y, L]
\end{align*}
\]

with evaporation flux at the right boundaries for the fiber and yarn model, and a homogeneous Neumann BC at their left boundaries and for the room model at both ends.

At certain points in time equilibrium is essentially reached between the three models. Plotting the logarithmic concentration against the logarithmic time scale shows that, for standard parameters, after a rather short time (approximately 0.05 s) the fiber and yarn concentrations coincide and after approximately 0.1 s those concentrations coincide with the concentration in the middle of the room.

As an upscaling method volume averaging is used, where the averaged outcome of one model serves as boundary conditions for the other.

These moments in time where equilibria are reached correspond with the systems characteristic times. These are the time scales \( \tau \) for a particle to travel over a distance \( x \) and on average these are given by \( \tau^d \approx \frac{x^2}{D} \) for diffusion and \( \tau^e = \frac{x}{v} \) for evaporation.

As a first estimation of these times one may calculate them by this rule of thumb for each of the levels as

\[
\begin{align*}
\tau^d_f &= \frac{(\Delta r)^2}{D_f}, \quad \tau^e_f = \frac{\Delta r}{v_{\text{fiya}}}, \\
\tau^d_y &= \frac{(\Delta \tilde{r})^2}{D_y}, \quad \tau^e_y = \frac{\Delta \tilde{r}}{v_{\text{yaro}}},
\end{align*}
\]

where \( v_{\text{fiya}} \) is the evaporation speed for the AI from the fiber surface to the yarn gaps, \( v_{\text{yaro}} \) is the evaporation speed for the AI from the yarn surface to the room, \( \Delta r \) and \( \Delta \tilde{r} \) are the width of the space discretization cells, and \( D_f \) and \( D_y \) are the respective diffusion coefficients of the first two levels.

A more precise way to calculate these characteristic times uses the Laplace transform of the flux. At interesting points of the system we interpret the diffusive flux \( F(x) \) as the
probability distribution function of the times $T$ when a particle passes by position $x$. The moment-generating function is then related to the Laplace transform of the flux:

$$M_T(-s) = E_T(e^{-sT}) = \int_0^{+\infty} e^{-st} F(s) ds = \mathcal{L}[F(x)](s),$$

and the cumulant-generating function $g(-s)$ is the logarithm of the Laplace transform of the flux, where $s$ is in the Laplace domain.

Doing so, we are particularly interested in the first and second cumulants. The first cumulant is

$$c_1 = -\frac{\partial}{\partial s} g(s)\bigg|_{s=0} = -\frac{\partial}{\partial s} \left[ \log(\mathcal{L}[F(x)](s)) \right]_{s=0},$$

which is the mean of the probability distribution, i.e. the residence time of the diffusion equation or the average time it takes a particle to pass a certain point.

Also the second cumulant or the second derivative of the logarithm of the Laplace transform of the flux in $s = 0$, i.e. the variance of the flux, is useful for interpreting the system.

All of these values can be exactly calculated in function of the parameters in the above equations and will help to understand the diffusion in open textile structures.

3 Conclusion and future work

A three-scale model consisting of a micro-, meso- and macrolevel was implemented in C language. Upscaling was done by volume averaging for the concentration calculated from the smaller level to serve as a source term and BC for the larger level. This can be adapted to the previously used overlapping domain decomposition method. Characteristic times were confirmed using the Laplace transform of the flux.

References

