

Numerical methods for deterministic and stochastic differential equations with multiple scales and high contrasts

Présentée le 18 septembre 2020

à la Faculté des sciences de base
Chaire d'analyse numérique et mathématiques computationnelles
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Giacomo ROSILHO DE SOUZA

Acceptée sur proposition du jury

Prof. D. Kressner, président du jury
Prof. A. Abdulle, directeur de thèse
Prof. M. J. Gander, rapporteur
Prof. S. Descombes, rapporteur
Prof. M. Picasso, rapporteur

A Mamma, Papà e Zum.

Acknowledgments

After a very long and diverse journey, where at times I got lost in a wild Brazilian rainforest and in others I found my way on a smooth and clear Swiss path, I finally reached the end of my PhD and I would like to thank all the people who has been at my side during this period.

First of all I would like to thank my thesis advisor Assyr for accepting me in his group and teaching me how to seriously carry out a research. Thank you very much for letting me be very independent, to believe in me even when I was getting lost and finding my way was hard and time consuming; but also for being around whenever I needed your directions. Thank you for teaching me to be more precise, clear, persevering, and for sharing your expertise and great ideas. Thank you for transmitting your passion and enthusiasm about the fascinating Runge–Kutta methods and the art of studying their stability. You have taught me to walk, and this we learn just once.

Furthermore, I would like to thank the jury members Marco Picasso, Martin J. Gander and Stéphane Descombes for taking their time to read through my thesis and for the rich and thoughtful discussions in the thesis defense. In addition, I thank the jury president Daniel Kressner for accepting to conduct the defense.

Thanks to the ANMC crew Andrea, Andrea Z., Doghonay, Edoardo, Giacomo G., Orane, Simon and Timothée for the helpful mathematical discussions, the pleasant environment and the funny and mind distracting moments. And I thank very much Virginie for taking charge very promptly of all the administrative tasks and for the funny jokes. A special thanks goes to the outsider Riccardo for his friendship, sincerity and the countless hilarious moments.

I could not forget to mention all my friends from Lausanne: Alessandro, Edone, Eva, Fabian, Gaya, Gonzalo, Juli, the two Michelle, Nicolas, Raïssa, Ronch, Shery, Stefania, Viola and Xavier, for making my life in Lausanne joyful, fulfilling and partygoing, with our endless aperos, dinners and bike rides. Thanks to you, my years in Lausanne will never be forgotten. And thanks to the FermentaTori brewers for all of ours very, very long and messy brews until 2am and both the successful and objectionable exBEERiments. Thanks also to Gaya and Viola for letting me distract myself during the writing of this thesis by fermenting everything I could find on my way and decorating the house with bizarre jars.

E infine, un immenso grazie va alla mia famiglia per l'amore ed il sostegno incondizionati. A Giulio e Jorge, per la vostra vitalità e sregolatezza rigeneranti. Ma in particolare ai miei genitori: Mamma, Papà e Zum, per gli stimoli che mi avete sempre dato, per aver sempre appoggiato le mie scelte e per avermi dato la possibilità di cominciare gli studi all'EPFL. Grazie ♡.

Vezio, 11 June 2020
Giacomo Rosilho de Souza

Abstract

Mathematical models involving multiple scales are essential for the description of physical systems. In particular, these models are important for the simulation of time-dependent phenomena, such as the heat flow, where the Laplacian contains mixed and indistinguishable fast and slow modes. Stationary problems can also exhibit a multiscale nature. For example, elliptic equations governed by a diffusion coefficient with strong discontinuities have solutions characterized by regions with a high gradient. Simulating such models is very demanding, as the computational cost of standard numerical methods is usually ruled by the fastest dynamics or the smallest scale.

In the first part of this thesis, we develop multirate integration methods for deterministic and stochastic time-dependent problems with disparate time-scales. The cost of traditional schemes for such problems is prohibitive due to step size restrictions in the explicit case or solutions to large nonlinear systems in the implicit case. Existing multirate methods are either implicit or make use of interpolations, which trigger instabilities, or are based on a scale separation assumption, which is not satisfied by parabolic problems. Here we introduce a new framework based on modified equations which allows for the development of a whole new class of interpolation-free explicit multirate numerical methods, which do not need any scale separation, are stable and accurate. For deterministic problems, our methodology is based on the replacement of the original right-hand side by an averaged force, whose stiffness is reduced due to a fast but cheap auxiliary problem. Integrating the modified equation and the auxiliary problems by explicit schemes is generally cheaper than integrating the original problem. We thus introduce a multirate method based on stabilized explicit schemes and prove its efficiency, stability and accuracy. Numerical experiments show that standard schemes and our multirate approach provide essentially the same solutions; hence, the bottleneck caused by the stiffness of a few degrees of freedom is overcome without sacrificing accuracy. We also generalize the same framework to stochastic differential equations, where we need to introduce a damped diffusion term for which the resulting modified equation inherits the mean-square stability properties of the original problem. An interpolation-free stabilized multirate method for stochastic equations is then derived.

In the second part of this thesis, we consider elliptic problems with high gradients and develop a local adaptive discontinuous Galerkin scheme. Local methods for such problems already exist in literature; however, they are usually based on iterations and have several downsides. In particular, their a priori error analysis is based on rather strong and nonphysical assumptions and they lack a rigorous a posteriori error analysis. The scheme that we propose is based on a coarse solution on the full domain which is subsequently improved by solving local elliptic problems only once on subdomains with artificial boundary conditions. The a priori error analysis is performed under minimal regularity assumptions due to the gradient discretization framework. Furthermore, we derive a posteriori error estimators based on conforming fluxes and potential reconstructions which can be used to identify the local subdomains on the fly, are free of undetermined constants and robust in singularly perturbed regimes.

Key words. multirate methods, Chebyshev methods, explicit time integrators, stiff equations, stochastic differential equations, multiscale problems, local methods, discontinuous Galerkin, gradient discretization, a posteriori error estimators

Sommario

Modelli matematici che coinvolgono più scale sono imprensindibili per la descrizione di svariati sistemi fisici. Per esempio, questi modelli sono importanti per la simulazione di fenomeni tempo-dipendenti, come il flusso di calore, dove il Laplaciano contiene frequenze miste e indistinguibili, alte e basse. Anche problemi stazionari possono essere di natura multiscala: equazioni ellittiche definite da un coefficiente di diffusione fortemente discontinuo hanno soluzioni caratterizzate da regioni ad alto gradiente. La simulazione di tali modelli è molto costosa, poiché il tempo di calcolo dei metodi numerici classici è solitamente determinato dalla dinamica più veloce o dalla scala più piccola.

Nella prima parte di questa tesi, sviluppiamo metodi di integrazione multiscala per problemi deterministici e stocastici tempo-dipendenti. Il costo di schemi tradizionali per tali problemi è insostenibile a causa delle restrizioni sul passo di tempo nel caso esplicito o della risoluzione di sistemi non lineari nel caso implicito. I metodi multiscala esistenti sono impliciti o ricorrono a interpolazioni, che innescano instabilità. Oppure, si basano su di un'ipotesi di separazione di scala, non soddisfatta dai problemi parabolici. Qui presentiamo una nuova metodologia basata su equazioni modificate che permette lo sviluppo di un nuovo tipo di metodi numerici multiscala esplicativi, esenti da interpolazioni, che non necessitano di alcuna separazione di scala, stabili e accurati. Per problemi deterministici, il metodo si basa sulla sostituzione della forzante originale con una sua media, la cui rigidità è ridotta grazie ad un problema ausiliario veloce ma economico. Integrare l'equazione modificata ed i problemi ausiliari con schemi esplicativi è generalmente più economico che integrare il problema originale. Introduciamo quindi un metodo multiscala basato su schemi stabilizzati esplicativi e ne dimostriamo l'efficienza, la stabilità e l'accuratezza. Dopodiché, generalizziamo la stessa metodologia alle equazioni differenziali stocastiche, dove introduciamo un termine di diffusione smorzato per il quale l'equazione modificata che ne risulta eredita le proprietà di stabilità del problema originale. Da quest'equazione modificata sviluppiamo un metodo multiscala stabilizzato esplicito per le equazioni differenziali stocastiche.

Nella seconda parte della tesi, consideriamo problemi ellittici con gradienti elevati e sviluppiamo uno schema locale-adattivo discontinuo di Galerkin. Metodi locali per tali problemi esistono in letteratura; tuttavia, essi sono solitamente iterativi ed hanno diversi lati negativi. Per esempio, la loro analisi degli errori a priori si basa su ipotesi piuttosto forti e inverosimili, per di più mancano di una rigorosa analisi degli errori a posteriori. Lo schema che proponiamo si basa su di una soluzione grossolana sull'intero dominio che viene successivamente raffinata risolvendo dei problemi ellittici locali definiti in sottodomini con condizioni al bordo artificiali. L'analisi dell'errore a priori viene eseguita con ipotesi di regolarità minime grazie al metodo di discretizzazione del gradiente. Inoltre, deriviamo stimatori di errore a posteriori basati su flussi conformi e ricostruzioni del potenziale che possono essere utilizzati per identificare i sottodomini locali, sono privi di costanti indeterminate e affidabili in regimi singolarmente perturbati.

Parole chiave. metodi multiscala, metodi di Chebyshev, equazioni rigide, integratori temporali esplicativi, equazioni differenziali stocastiche, problemi multiscala, metodi locali, metodo di Galerkin discontinuo, discretizzazione del gradiente, analisi degli errori a posteriori

Contents

Acknowledgements	i
Abstract (english/italiano)	iii
Notation	xi
Introduction	1
I Stabilized explicit multirate methods for stiff deterministic and stochastic differential equations	11
1 Stabilized explicit Runge–Kutta methods	15
1.1 Optimal stability polynomials for first-order methods	15
1.2 First-order stabilized explicit methods	18
1.2.1 Stability condition and computational advantages	18
1.2.2 Factorization methods	19
1.2.3 Diagonal methods	20
1.2.4 First-order Runge–Kutta–Chebyshev (RKC) methods	20
1.3 Higher order methods	23
1.3.1 An overview of higher order stabilized methods	23
1.3.2 Second-order Runge–Kutta–Chebyshev methods	23
1.4 Stabilized explicit methods for stochastic differential equations	26
1.4.1 The second kind orthogonal Runge–Kutta–Chebyshev (SK-ROCK) method	26
1.4.2 The second kind τ -leap orthogonal Runge–Kutta–Chebyshev (SK- τ -ROCK) method	29
2 An interpolation based additive Runge–Kutta–Chebyshev method	35
2.1 The additive RKC (aRKC) method	35
2.1.1 Equation splitting	36
2.1.2 The additive RKC algorithm	36
2.2 Instabilities and order reduction	39
2.2.1 Instability on a model problem	39
2.2.2 Order reduction in the second-order aRKC scheme	40
2.3 Numerical experiments	43
2.3.1 Instability on the model problem	43
2.3.2 Order reduction on the heat equation	44
3 Stabilized explicit multirate methods for stiff differential equations	45
3.1 The modified equation	47
3.1.1 The averaged force	47
3.1.2 Multirate test equation and stability	50
3.2 The multirate explicit Euler (mEE) method	52

3.2.1	The mEE algorithm	52
3.2.2	Efficiency of the mEE method	52
3.2.3	Stability and convergence analysis	53
3.3	The semidiscrete multirate RKC method	57
3.3.1	The semidiscrete multirate RKC algorithm	57
3.3.2	The multirate exponential Euler-RKC method	57
3.3.3	Stability and convergence analysis	58
3.4	The multirate RKC (mRKC) method	59
3.4.1	The mRKC algorithm	59
3.4.2	Efficiency of the mRKC method	60
3.4.3	Stability and convergence analysis	63
3.4.4	A step size control strategy	68
3.4.5	The mRKC method for problems with well separated scales	72
3.5	Numerical Experiments	74
3.5.1	Robertson's stiff test problem	74
3.5.2	Heat equation in the unit square	76
3.5.3	Diffusion across a narrow channel	76
3.5.4	Reaction-convection-diffusion problem	78
3.5.5	The Brusselator reaction-diffusion problem	81
3.6	Proofs of technical results	82
3.6.1	Technical results for the mEE method	84
3.6.2	Technical results for the mRKC method	85
4	Stabilized explicit multirate methods for stiff stochastic differential equations	89
4.1	The stochastic modified equation	90
4.1.1	Preliminary motivations	90
4.1.2	The damped diffusion	92
4.2	The multirate Euler–Maruyama (mEM) method	93
4.2.1	The mEM algorithm	93
4.2.2	Efficiency of the mEM method	94
4.2.3	Mean-square stability analysis	95
4.2.4	Convergence analysis	98
4.3	The multirate SK-ROCK (mSK-ROCK) method	99
4.3.1	The mSK-ROCK algorithm	99
4.3.2	Efficiency of the mSK-ROCK method	100
4.3.3	Mean-square stability analysis	101
4.3.4	Convergence analysis	104
4.4	Numerical experiments	106
4.4.1	Nonstiff problem convergence experiment	106
4.4.2	Stiff problem convergence experiment	107
4.4.3	E. Coli bacteria heat shock response	107
4.4.4	Diffusion across a narrow channel with multiplicative space-time noise	108
4.5	Proofs of technical results	112
4.5.1	Technical results for the mEM method	112
4.5.2	Technical results for the mSKROCK method	116
5	Conclusion of Part I	123
II	Local adaptive discontinuous Galerkin schemes for elliptic equations	125
6	The weighted discontinuous Galerkin gradient discretization scheme	129

6.1	The gradient discretization method for homogeneous Dirichlet boundary conditions	130
6.1.1	Definition and approximation properties	130
6.1.2	A priori error analysis for linear and quasilinear elliptic equations	132
6.2	The Weighted Discontinuous Galerkin Gradient Discretization (WDGGD)	133
6.2.1	The gradient discretization	134
6.2.2	Analysis of approximation properties	135
6.2.3	Equivalence to the symmetric weighted interior penalty method	136
7	A local weighted discontinuous Galerkin gradient discretization scheme for linear and quasilinear elliptic equations	139
7.1	Notation and preliminary results	139
7.2	The local WDGGD scheme for linear elliptic problems	143
7.2.1	A priori error analysis for the local solution	144
7.2.2	Improved local estimate	149
7.2.3	A priori error analysis for the global solution	150
7.3	The local WDGGD scheme for quasilinear elliptic problems	153
7.3.1	A priori error analysis for quasilinear problems	153
7.4	Numerical experiments	156
7.4.1	Convergence rates	156
7.4.2	Influence of artificial boundary conditions	158
7.4.3	Non regular problem: discontinuous data	159
7.4.4	Computational efficiency for a linear equation	161
7.4.5	Quasilinear equation	162
8	A posteriori error analysis of a local adaptive discontinuous Galerkin scheme for advection-diffusion-reaction equations	165
8.1	The local adaptive discontinuous Galerkin scheme	166
8.1.1	Preliminary definitions	166
8.1.2	The local adaptive algorithm	167
8.2	A posteriori error estimators via flux and potential reconstructions	169
8.2.1	A posteriori error estimators	170
8.2.2	Main results	171
8.3	Potential and flux reconstructions, proofs of the main results	172
8.3.1	Potential and flux reconstruction via the equilibrated flux method	172
8.3.2	Constant definitions and preliminary results	175
8.3.3	Proof of the theorems	179
8.3.4	Alternative error bounds	179
8.4	Numerical experiments	180
8.4.1	Error estimators rate of convergence	181
8.4.2	Reaction dominated problem	182
8.4.3	Convection dominated problem	184
8.4.4	A smooth problem	185
9	Conclusion of Part II	189
Bibliography		191
Curriculum Vitae		199

Notation

Abbreviations

ODE	ordinary differential equation
SDE	stochastic differential equation
PDE	partial differential equation
RK	Runge–Kutta
EE	explicit Euler
EM	Euler–Maruyama
SE	stabilized explicit
RKC	Runge–Kutta–Chebyshev
SK-ROCK	second kind orthogonal Runge–Kutta–Chebyshev
SK- τ -ROCK	second kind τ -leaping orthogonal Runge–Kutta–Chebyshev
mEE	multirate EE
mEM	multirate EM
mRKC	multirate RKC
mSK-ROCK	multirate SK-ROCK
GD	gradient discretization
GDM	gradient discretization method
GS	gradient scheme
DG	discontinuous Galerkin
SWIP	symmetric weighted interior penalty
WDGGD	weighted discontinuous Galerkin gradient discretization

Standard sets of numbers

\mathbb{N}	set of positive integers $\{0, 1, 2, \dots\}$
\mathbb{N}^*	set of strictly positive integers $\{1, 2, 3, \dots\}$
\mathbb{Z}	set of integers $\{\dots, -1, 0, 1, \dots\}$
\mathbb{R}	set of real numbers
\mathbb{C}	set of complex numbers

Differentials

∂_t	partial differential with respect to the time t
∇	gradient operator
$\nabla \cdot$	divergence operator
Δ	Laplacian operator

Functional spaces

Let D be an open domain of \mathbb{R}^d , $d \in \mathbb{N}^*$, and consider functions $D \rightarrow \mathbb{R}$.

$C^k(D)$	space of k -times continuously differentiable functions
$C_p^k(D)$	functions of $C^k(D)$ having derivatives with at most polynomial growth
$C_b^\infty(D)$	infinitely many times differentiable functions with compact support in D
$\mathbb{P}_k(D)$	space of polynomials in D of total degree k
$L^p(D)$	usual Lebesgue space with $p \in [1, \infty]$
$W^{k,p}(D)$	usual Sobolev space with $k \in \mathbb{N}$ and $p \in [1, \infty]$
$H^k(D)$	Sobolev space $W^{k,2}(D)$
$H_0^1(D)$	closure in $H^1(D)$ of infinitely many times differentiable functions with compact support in D
$H_{\text{div}}(D)$	space of function in $L^2(D)^d$ with divergence in $L^2(D)$

Introduction

Mathematical models for physical, engineering or social sciences are often endowed with a multiscale structure or with high contrasts, which may be caused either by the intrinsic nature of the problems or by their mathematical description. For instance, the flow of water in an unsaturated porous media, as soil, has high contrasts due to different materials' permeabilities and strong nonlinearities in the mathematical model. Another example is given by the modeling of thermal effects in nanodevices, as transistors, which at the level of chip size imposes different scales of resolution. A different class of problems may arise from chemical systems with reactions occurring at disparate rates, as well as population growth models where different species may grow at very different speeds. In this class of time dependent problems, an additional difficulty is introduced when a clear-cut separation of the fast and the slow dynamics is not observable. The most striking example of this phenomenon is probably a spatially discretized heat equation, where the fast dynamics usually emerge only in refined regions and, in contrast, slow dynamics appear everywhere in the computational domain. As a consequence, in refined regions we have mixed and indistinguishable fast and slow dynamics.

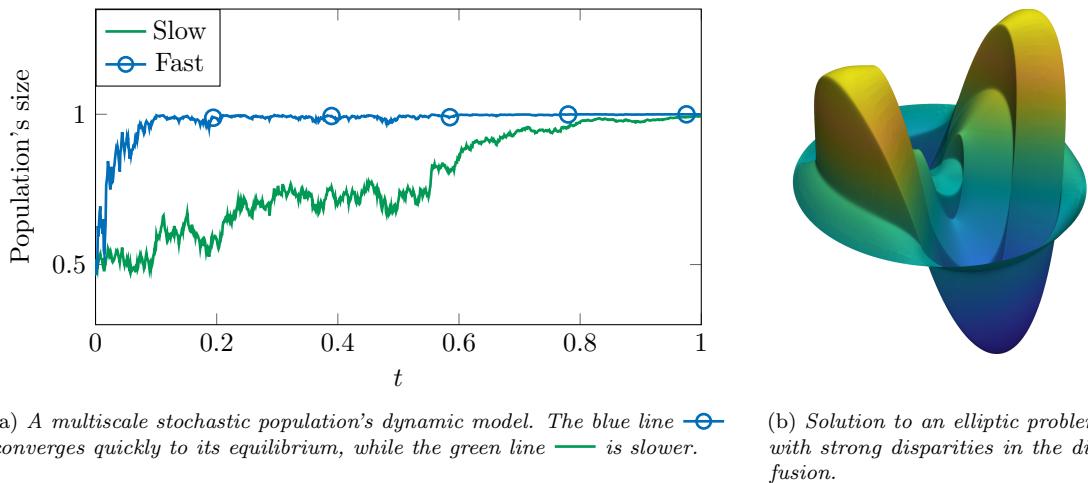


Figure 1. Examples of multiscale and high contrast problems.

When it comes to approximate the solution to such problems, the computational demands of traditional numerical methods are generally ruled by the smallest scales or the fastest dynamics. Hence, classical schemes rapidly become overly expensive or even practically impossible; thereby, multiscale or adaptive numerical methods are paramount for the approximation of such problems. Nowadays, there is a large body of literature concerned with the development of these schemes, where the aim is to take advantage of the problem's characteristics (as localized contrasts or only a few fast components) in order to decrease the methods' computational demands.

In this thesis, we develop numerical schemes for two classes of multiscale problems. First, we consider ordinary or stochastic differential equations with fast and slow terms, as the one in Figure 1(a). We also look at situations where the fast and slow scales are not separated, as in a spatially discretized parabolic equation. Second, we consider elliptic partial differential equations with high gradients in the solution, as in Figure 1(b). Since these two classes of problems are different in nature, this thesis is structured in two parts, described below.

Part I: Stabilized explicit multirate methods for stiff deterministic and stochastic differential equations

In Part I of the thesis, we start considering multirate ordinary differential equations of the type

$$y' = f(y) := f_F(y) + f_S(y), \quad y(0) = y_0, \quad (1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ splits into an expensive but only mildly stiff part, f_S , associated with relatively slow (S) time-scales, and a cheap but severely stiff part, f_F , associated with fast (F) time-scales.

Practical examples of (1) are systems of chemical reactions occurring at disparate time-scales or mixed digital and analogical electric circuits operating in nano- and micro-seconds, respectively. However, our main interest is on systems stemming from the spatial discretization of parabolic (or diffusion dominated) partial differential equations. These are large systems of stiff ordinary differential equations where the eigenvalues of the Jacobian matrix of f lie in a narrow strip along the negative real axis whose extent scales as h^{-2} , where h is the smallest mesh size. Therefore, we deal with problems where we cannot make any assumption of scale separation, that is, to suppose that the spectrum of the Jacobian of f can simply be split into fast and slow modes. However, it is still possible to split f in two terms f_S and f_F representing the Laplacian in coarse and refined regions, respectively. The Jacobian of f_S will have only relatively small eigenvalues while the Jacobian of f_F both small and large eigenvalues.

When applied to (1), standard explicit methods are highly inefficient due to the stringent stability constraint on the step size, which scales as $1/\rho_F$, where ρ_F is the spectral radius of the Jacobian of the stiff term f_F . Hence, the number of function evaluations is proportional to ρ_F and the integration cost (number of function evaluations times the evaluation cost) is proportional to $\rho_F c_S$, where c_S is the cost of evaluating f_S (remember that f_F is cheap to evaluate). For the spatially discretized parabolic problem, for instance, the step size is proportional to h^2 and the number of time steps is proportional to h^{-2} , with h the smallest mesh size; the cost is therefore proportional to c_S/h^2 . Looking at the quantity $\rho_F c_S$, we readily see that explicit methods can improve their efficiency for (1) only if evaluation of f_F and f_S is decoupled. As long as they are evaluated concurrently, very few degrees of freedom inducing a severe stiffness in f_F (large ρ_F) drastically increase the integration cost, even if f_F remains inexpensive.

In contrast, implicit methods, albeit unconditionally stable, require the solution of a large linear (or possibly nonlinear) system of equations at each time step, a high price to pay in terms of computer memory and execution time. Moreover, when sheer size requires the use of iterative methods, the overall performance heavily relies on the availability of efficient preconditioners while the convergence of Newton-like nonlinear solvers is not even guaranteed for larger step sizes.

Stabilized Runge–Kutta methods (also called Chebyshev methods) fall somewhere in between: they are explicit and thus avoid the solution of large systems of equations, while their stability interval on the negative real axis is proportional to s^2 for an s -stage method. Thanks to this remarkable quadratic dependency, the total number of function evaluations is proportional to $\sqrt{\rho_F}$, hence the integration cost is proportional to $\sqrt{\rho_F} c_S$. For spatially discretized parabolic

problems, the integration cost scales linearly with c_S/h , in contrast to the quadratic dependence on h of standard explicit integrators. However, f_F and f_S are still evaluated concurrently and the presence of a few degrees of freedom in f_F inducing severe stiffness (large $\sqrt{\rho_F}$) completely destroys the efficiency of the stabilized explicit schemes. Thus, also for stabilized explicit schemes, evaluation of f_F and f_S must be decoupled.

The same computational challenges arise when solving stiff multirate stochastic differential equations as

$$dX(t) = f_F(X(t)) dt + f_S(X(t)) dt + g(X(t)) dW(t), \quad X(0) = X_0, \quad (2)$$

where $X(t)$ is a stochastic process in \mathbb{R}^n , $f_F, f_S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are as in (1), $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ is the diffusion term and $W(t)$ is an m -dimensional Wiener process. Indeed, due to the stiffness of f_F traditional explicit schemes as Euler–Maruyama face stringent conditions on the step size, and on the other hand implicit and semi-implicit methods require the solution to large nonlinear systems. Stabilized explicit methods for (2) exist, but they however face the same efficiency loss when a few degrees of freedom in f_F induce severe stiffness in the problem.

Due to the inefficiency of standard integration schemes when applied to (1) or (2), one should resort to *multirate methods*: they take advantage of the special structure of $f = f_F + f_S$ in order to reduce the computational costs.

Literature review on multirate methods. In the literature, numerical methods which exploit the special structure of (1) in order to reduce the computational costs are called multirate methods. By using different step sizes or even entirely different schemes for f_S and f_F , multirate methods overcome the stringent step size restriction due to the cheap but stiffer part, f_F , while retaining the efficiency of explicit time integration for f_S . Since the early work of Rice [102], various explicit, implicit or hybrid multirate schemes have been developed based on partitioned Runge–Kutta methods [20, 44, 64, 65, 73, 81, 105, 106, 108]. These methods make no assumption of time-scale separation, but instead take advantage of the special structure of f in (1). In their seminal work [58], Gear and Wells developed a number of multirate strategies for the interlaced time integration of the “slow” and “fast” components using classical multistep schemes. However, all those methods usually require a predictor step and either interpolate or extrapolate between “fast” and “slow” state variables, which is prone to instability: “*The moment the system is not triangular, any form of stability at infinity is lost because of the extrapolation which brings in off-diagonal blocks in the numerical operator which are polynomial in H rather than rational*” [58]. Although some of the implicit-explicit (IMEX) methods are provably stable, they are more cumbersome to implement and rapidly become too expensive as the number of “fast” unknowns increases.

Following the original local adaptive mesh refinement (AMR) strategy for first-order hyperbolic conservation laws by Berger and Oliger [28], various multirate (or local time-stepping in the PDE context) methods were also proposed for parabolic problems, which use different step sizes in different regions of the computational domain. To overcome the inherent stiffness of parabolic problems in the presence of local mesh refinement, Ewing et al. [50, 51] proposed a fully implicit space-time approach which uses implicit Euler steps of different sizes in different parts of the domain. Dawson, Du and Dupont [38] devised a finite difference domain decomposition algorithm, where the heat equation is locally integrated in each subdomain using the implicit Euler method while shared boundary values at interfaces are computed explicitly on a coarser mesh. In [94, 107], various predictor-corrector and domain decomposition methods were combined to iteratively correct the solution or its boundary values at artificial interfaces.

By using static-regridding, Trompert and Verwer [113, 114, 115, 116] developed a number of

multirate time-stepping strategies for local uniform grid refinement (LUGR). Given a fixed step size, a first integration is performed on an underlying coarse grid covering the entire domain. Then the numerical solution is iteratively refined on nested, finer-and-finer, uniform subgrids interpolating needed values between different levels, until an error tolerance comparable to that obtained using a globally fine grid is achieved. See [56] for a review on local time-stepping methods.

More recently a stabilized explicit Runge–Kutta method, called RKC, was combined with the AMR approach [14, 95] to tackle diffusion dominated problems. Here, fine and coarse cells correspond to stiffer and less stiff parts of the problem, respectively, and two RKC schemes adapting the number of stages according to the local mesh size are employed. This leads to asynchronous integration of the fine and coarse regions and whenever ghost cell values are needed a linear interpolation in time between stages is employed. For a certain class of problems, however, linear interpolation of missing values may render the scheme unstable [14] — see also Chapter 2.

Finally, there is a last class of schemes based on a scale separation assumption, i.e. f can be easily split into fast and slow terms; those are the multiscale and projective methods. The heterogeneous multiscale methods (HMM) [42, 121] are based on the derivation of an effective equation for the slow variables, which depends on the invariant measure of the fast dynamics. This limiting equation is integrated with large step sizes and its coefficients are estimated on the fly using micro-simulations of the fast dynamics. Projective integration methods [57] are a different class of schemes based as well on a scale separation assumption. They perform a sequence of short relaxation steps in order to highly damp the fast variables, then extrapolate the last relaxation step over a larger time-step. As the fast variables are damped, the large step is stable. Multiscale and projective methods are strongly based on a scale separation assumption and therefore they cannot be employed, for instance, when (1) stems from the spatial discretization of a parabolic partial differential equation.

For stiff multirate stochastic equations as (2) the literature on multirate methods is much scarcer than for deterministic problems. The only methods capable to exploit the special structure of $f = f_F + f_S$, up to the author’s knowledge, are the stochastic HMM [43, 87, 120] and the stochastic projective methods [62, 77], which are, again, strongly based on a scale separation assumption. As for deterministic problems, the stochastic HMM methods hinge on an effective equation for the slow variables, whose terms depend on the invariant measure of the fast processes; this invariant measure is estimated on the fly taking ensemble-time averages over numerical solutions of the fast processes. In the stochastic framework, projective methods get closer to the HMM schemes and make use of an effective equation too [62], see [121] for a comparison. An extension of HMM methods to stochastic PDEs is found in [10, 11].

Summarizing, we notice that the current multirate methods for (1) are: explicit methods which face stringent step size restrictions or are prone to instabilities, implicit methods which are stable but must solve complicated linear (or nonlinear) systems, multiscale or projective methods which are based on a scale separation assumption and therefore cannot be used for spatially discretized parabolic problems. For (2), only multiscale or projective methods are available.

Part I, main contributions. In this thesis we propose a new class of multirate methods for (1) and (2) which, in contrast to previous schemes, are explicit, do not have any step size restriction, do not need any interpolations nor extrapolations, are proven to be stable on a model problem and are not based on any scale separation assumption. The schemes are robust in the sense that even if f_F and f_S switch roles (i.e. f_S becomes stiffer than f_F) the solution remains accurate.

Our methodology is based on the introduction of a new framework based on modified equations for (1) and (2) which are good approximations to the original problems but where stiffness is reduced and depends solely on the slow term f_S . Therefore, integration of the modified equations by explicit schemes is cheaper than the original problems. Although in this thesis we mainly focus on stabilized explicit methods, the framework introduced here allows for the development of a whole class of explicit multirate schemes for (1) and (2).

For (1) we define the modified equation

$$y'_\eta = f_\eta(y_\eta), \quad y_\eta(0) = y_0, \quad (3)$$

where f_η depends on a free parameter η . If η is appropriately chosen, then

$$f_\eta = f_F + f_S + \mathcal{O}(\eta) \quad \text{and} \quad \rho_\eta \leq \rho_S,$$

where ρ_η and ρ_S are the spectral radii of the Jacobians of f_η and f_S , respectively. Hence, f_η is an approximation to $f = f_F + f_S$ but whose stiffness is decreased; to do so, we use an average of f along the direction defined by a fast auxiliary problem in a short time interval of size η . Therefore, for each evaluation of f_η the fast auxiliary problem must be solved, however, since f_F is cheap to evaluate, the computation of f_η is not expensive compared to $f_F + f_S$. Since $\rho_\eta \leq \rho_S$, the number of evaluations of f_η needed by an explicit scheme applied to (3) depends solely on the slow dynamics and as the cost of f_η is comparable to the cost of $f_F + f_S$ integrating (3) is much cheaper than (1). The averaged force f_η is such that if the original right-hand side is contractive then, under some hypothesis, f_η is contractive. Thanks to this property the solution to the original problem (1) is very well approximated by the solution to (3).

Therefore, the modified equation (3) provides a framework to develop a new class of explicit multirate methods, where both (3) and the auxiliary problem are approximated by some time discretization scheme, or even by two different schemes. This procedure allows to overcome the bottleneck caused by the stiffness of f_F without sacrificing neither explicitness nor accuracy. Furthermore, the schemes do not need interpolations nor any scale separation assumption.

The main explicit multirate method for (1) developed in this thesis, called mRKC for multirate RKC, is based on the discretization of (3) and of the auxiliary problem by the first-order Runge–Kutta–Chebyshev (RKC) scheme [110, 118, 124, 125] for stiff ordinary differential equations. The RKC scheme is a stabilized explicit method with an extended stability domain growing quadratically with the number of stages s . Furthermore, the stability domain attains the optimal size for an s -stage Runge–Kutta method. Since the spectral radius of the Jacobian of f_η is bounded by ρ_S , the spectral radius of f_S , the number of evaluations of f_η needed by the mRKC method grows as $\sqrt{\rho_S}$, and not as $\sqrt{\rho_F}$ as for (1). The mRKC method is first-order accurate, proven to be stable on two model problems and computational efficiency is proven theoretically and numerically. In order to illustrate the flexibility of the modified equation framework we will discuss, only theoretically, also two other explicit multirate methods based on (3).

For the multirate stochastic differential equation (2) we introduce the modified equation

$$dX_\eta(t) = f_\eta(X_\eta(t)) dt + g_\eta(X_\eta(t)) dW(t), \quad X_\eta(0) = X_0, \quad (4)$$

where f_η is the same as in (3) and g_η is a damped approximation of g , which is as well defined through the solutions to fast auxiliary deterministic problems. A stabilization procedure for the diffusion term g is needed in order to preserve the mean-square stability properties of the original problem (2). Indeed, as f_η has weaker contractivity properties than f , it cannot control the original noise term g ; hence, a damped noise g_η must be introduced.

As (3), also (4) provides a framework for the development of a new class of explicit multirate methods, where both (4) and the auxiliary problems needed to evaluate f_η, g_η are solved by

different integration schemes. Again, this procedure results in interpolation-free schemes which do not need any scale separation assumption.

We will introduce the multirate SK-ROCK (mSK-ROCK) method for (2), where (4) is solved by the SK-ROCK scheme and the auxiliary problems by the RKC scheme. The SK-ROCK scheme for stiff stochastic differential equations (SDE) has been introduced in [2] and it is one of the possible extensions of the RKC scheme to SDEs, see also [8, 16]. However, SK-ROCK is the only scheme which reaches optimal size of the stability domain for an s -stage Runge–Kutta method for SDEs. The mSK-ROCK method inherits the main properties of the mRKC and SK-ROCK schemes: it is explicit, the stability domain grows optimally and quadratically with the number of stages, the number of expensive function evaluations depends solely on the slow terms, it is not based on any scale separation assumption, there is no need of interpolations nor extrapolations and therefore it is straightforward to implement.

Application of the modified equation framework (4) to stochastic differential equations driven by other types of noise than white noise is not introduced in this thesis as it is still under investigation. Nevertheless, a first building block towards constructing multirate methods for chemical reactions with discrete noise has been developed in this thesis [5, 55]. The standard integration method for such equations is the τ -leaping scheme, which is the discrete noise counterpart of the Euler–Maruyama method. A stabilized method for this class of problems exists and is called τ -ROCK [7]. However, this scheme does not have optimal stability properties and is inefficient in capturing the exact invariant measure of ergodic processes. Hence, in Section 1.4.2 we introduce the new SK- τ -ROCK scheme. It is an SK-ROCK-like method for discrete noise equations which has optimal stability properties and shows excellent capabilities of capturing the exact invariant measures of SDEs driven by discrete noise such as chemical reactions.

Part II: Local adaptive discontinuous Galerkin schemes for elliptic equations

In Part II, we propose a local scheme for quasilinear elliptic equations of the form

$$\begin{aligned} -\nabla \cdot (A(u)\nabla u) &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega, \end{aligned} \tag{5}$$

and for the advection-diffusion-reaction equation

$$\begin{aligned} -\nabla \cdot (A\nabla u) + \beta \cdot \nabla u + \mu u &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega, \end{aligned} \tag{6}$$

where Ω is an open bounded polytopal connected subset of \mathbb{R}^d for $d \geq 2$, A is the diffusion tensor, β the velocity field, μ the reaction coefficient and f a forcing term. In (5), the tensor A may depend on the solution u , as we consider a linear and a quasilinear case.

In order to capture strong variations of the exact solution in the numerical approximations of elliptic PDEs, non-uniform grids are usually required. The construction of such grids is often based on an iterative process, where a solution is computed and an a posteriori error estimator is used to mark the regions where the mesh has to be refined, see [19, 25, 101, 122]. In such approach, the solution is computed on the whole domain at each step, even if the mesh has changed only in a small portion of the domain.

The aim of local schemes for elliptic equations is to decouple the solution of the coarse and refined regions and therefore reduce the computational costs. These schemes are computationally more efficient than classical schemes for elliptic PDEs with strong variations for several reasons.

- For linear problems, when using an iterative solver such as the conjugate gradient (CG) method we have smaller problems to compute on the finer meshes. Conversely, the non-local classical schemes need the solution of global linear systems with a large number of degrees of freedom (DOF) (recall that the CG method has a convergence rate that is super-linear with respect to the DOF of the system).
- When solving a linear system arising from PDEs with CG methods, preconditioners are usually needed, a usual choice for CG being the incomplete Cholesky (IC) factorization. For non-local schemes, the high contrast of the PDE leads to systems with a high condition number (due to mesh and data variations). For the local scheme, as each subdomain involves smaller variations of the solution and data, the condition number is smaller, leading to faster convergence of the iterative method.

Literature review on local schemes for elliptic equations. One of the first local methods to appear in the literature is the Local Defect Correction method (LDC) presented in [67], which consists of an iterative process that at each step solves a global problem on a coarse mesh and a local problem on a fine mesh. The solution of the global problem provides artificial boundary conditions to the local problem, whose solution is in turn introduced into the coarse system to estimate its residual. The coarse system is solved again but correcting the right-hand side with the residual of the local solution, leading to a more accurate coarse solution and hence better artificial boundary conditions for the next local problem. Two similar methods are the Fast Adaptive Composite grid algorithm [90] and the Multi-Level Adaptive Technique [29]. In [53] it is shown that under reasonable assumptions the three methods lead to the same solution. In their original form the schemes were defined for finite difference methods but finite volumes or finite element versions exist, see [91, 127]. Only recently the LDC scheme has been coupled with an a posteriori error estimator in order to automatically select the local domains [27].

The aforementioned local methods usually employ finite difference schemes and therefore their application to equations defined on complex geometries is cumbersome. Furthermore, the a priori error analysis lies on strong and rather nonphysical hypothesis, as high smoothness of the solution. Finally, a priori knowledge of the under resolved regions and thus of the local domains is required. The only exception is the LDC scheme, which is endowed with error estimators used to mark the local domains. However, a rigorous a posteriori error analysis is lacking for this method.

Part II, main contributions. Our aim in Part II of the thesis is to design a flexible local adaptive discontinuous Galerkin scheme, for which we can provide an a priori error analysis under minimal regularity assumptions for linear and quasilinear equations as (5), that is, assuming $f \in H^{-1}(\Omega)$ and $u \in H_0^1(\Omega)$. Then, we derive as well an a posteriori error analysis for the same method when applied to (6), which can be used to rigorously bound the numerical error and to identify the subdomains to be refined. This is crucial for practical applications of the local method.

Our local scheme for (5) is based the Symmetric Weighted Interior Penalty (SWIP) discontinuous Galerkin scheme [39, 49] and on a sequence of local subdomains $\{\Omega_k\}_{k=1}^M$, $M \geq 1$, adapted to the variation of the solution. In particular, we have $\Omega_1 = \Omega$, and Ω_k for $k \geq 2$ can be any polytopal subdomain of Ω . A typical example is given by an embedded sequence $\Omega_{k+1} \subset \Omega_k$, but any sequence of subdomains is allowed. The local method for (5) proceeds, informally, as follows.

- 1) Solve (5) on the domain $\Omega_1 = \Omega$ using the SWIP method on a coarse mesh, call u_1 the solution.
- 2) For $k = 2, \dots, M$:

- i) Define a local problem as the restriction of (5) to Ω_k with artificial Dirichlet boundary conditions extracted from u_{k-1} .
- ii) Solve the local problem with the SWIP method on a finer mesh in Ω_k ; call \hat{u}_k the solution.
- iii) Define a new solution on Ω as $u_k = \hat{u}_k + u_{k-1}\chi_{\Omega \setminus \Omega_k}$,

where $\chi_{\Omega \setminus \Omega_k}$ is the indicator function on $\Omega \setminus \Omega_k$.

We readily see two differences with respect to previous local methods: the use of the SWIP scheme and the fact that only local problems need to be solved and no iterations are needed between subdomains, as we define artificial boundary conditions and compute the solution only once in each local domain. Furthermore, when the tensor A is solution dependent the nonlinear problem (5) is solved only on the coarse global mesh, while the subsequent local problems are linearized. Therefore, the local method is particularly efficient for nonlinear problems.

Another main contribution is the analysis of the local scheme. Our local method is shown to converge for linear and quasilinear problems (5) under minimal regularity assumptions, i.e. $f \in H^{-1}(\Omega)$ and $u \in H^1(\Omega)$. This is achieved by using the Gradient Discretization (GD) method [41]; this is a framework for the convergence of discretization schemes, called Gradient Schemes (GS), for diffusion equations. Many popular discretization schemes can be written as a GS; indeed, we recast the SWIP scheme into a GS and then use the GD method to prove convergence of the local SWIP scheme. We stress that the GD framework is needed to prove convergence only; for practical purposes and implementation, we can use the usual discontinuous Galerkin setting.

The local scheme for (6) is essentially the same as the one for (5). However, the local domains Ω_k , for $k \geq 2$, are identified on the fly using a posteriori error estimators. Indeed, we derive an a posteriori error analysis for the local scheme for (6) and then employ the error estimators on the current solution u_k in order to mark the high error regions. In [48] an a posteriori error analysis based on conforming fluxes and potential reconstructions for the SWIP scheme is derived, following the same strategy we analyze the local scheme. The resulting error estimators inherit two key properties of the estimators obtained in [48]: they are robust in singularly perturbed regimes and free of undetermined constants.

Outline of the thesis

Here we detail the structure of this thesis. As Parts I and II are unrelated, the reader is free to choose the order in which to read them.

Outline of Part I. Here we present the modified equation frameworks and the stabilized explicit multirate integrators for deterministic and stochastic differential equations. We begin with an introductory chapter on stabilized explicit methods for deterministic and stochastic differential equations which we recommend to readers not familiar with these schemes. Then, we present a first attempt of an interpolation based stabilized explicit multirate scheme, which is however prone to instabilities. The two following chapters are the core of this part of the thesis, it is there that we introduce and analyze the modified equations (3) and (4) and the mRKC and mSK-ROCK methods. Conclusions are drawn in the last chapter.

In Chapter 1 we introduce stabilized explicit Runge–Kutta methods. We start deriving the optimal stability polynomial for a first-order s -stage explicit Runge–Kutta scheme, discussing as well the consistency and damping properties. Then we point out the computational advantages stemming

from the optimal stability polynomials and we define three of the first-order stabilized explicit schemes, namely: the factorization method, the diagonal method and the RKC method. Next we digress on higher order stabilized schemes, focusing on the second-order RKC method. The chapter is closed by the introduction to stabilized explicit schemes for stiff stochastic differential equations. We first recall the SK-ROCK scheme for equations driven by white noise and then we introduce the new SK- τ -ROCK scheme for equations driven by discrete noise, as chemical reactions. A numerical example showing the high accuracy and effectiveness of the SK- τ -ROCK method is also provided.

In Chapter 2 we discuss an interpolation based stabilized explicit additive Runge–Kutta scheme. The method is based on a splitting of the problem in severely and mildly stiff subproblems, which are then solved independently with an RKC scheme. The number of stages is adapted according to the subproblem’s stiffness and leads to asynchronous integration needing ghost values. Whenever ghost values are needed, linear interpolation in time between stages is employed. One important application of the scheme is for parabolic partial differential equations discretized on a nonuniform grid. However, the goal of this chapter is to introduce the scheme and prove on a model problem that linear interpolations trigger instabilities, corroborating previous results on interpolation based explicit multirate schemes. Furthermore, we show that the same scheme suffers from an order reduction phenomenon. The theoretical results are confirmed numerically.

Chapter 3 is devoted to the modified equation (3) and the mRKC scheme. We start defining the modified equation and studying its properties; as accuracy, decreased stiffness and contractivity. To illustrate how the new framework can be employed to define explicit multirate methods we first introduce the multirate explicit Euler method and study its stability and efficiency compared to the standard explicit Euler scheme. Then, we introduce a semidiscrete scheme where only (3) is discretized by an RKC method, but the auxiliary problem needed to evaluate f_η is solved exactly. We will analyze this scheme and discuss a case where it can be used in combination with exponential integrators and therefore employed for practical problems. Next, we introduce our mRKC scheme and study its theoretical efficiency, we prove its stability on two model problems and show that it is first-order accurate for general equations (1). We also derive a step size control strategy and analyze the scheme under the additional assumption that a clear-cut scales separation exist. Finally, we present a series of numerical experiments, where we verify the theoretical findings and demonstrate the usefulness and efficiency of the mRKC method.

In Chapter 4 we extend the results of Chapter 3 to stochastic differential equations (2). We introduce the stabilizing procedure for the diffusion term and the modified SDE (4). Then, we define the multirate Euler–Maruyama method, defined as the integration of (4) by the Euler–Maruyama scheme and of the auxiliary problems by the explicit Euler scheme. Next, we introduce the mSK-ROCK scheme, analyze its computational efficiency, show its mean-square stability on a model problem and prove that for general equations (2) it has strong order 1/2 and weak order 1. We conclude with a series of numerical experiments, where we verify the theoretical results and confirm the efficiency of the multirate method comparing it to the standard SK-ROCK scheme.

In Chapter 5 we summarize the main results and draw the conclusions of Part I of the thesis.

Outline of Part II. Here we introduce the local adaptive discontinuous Galerkin scheme for elliptic PDEs (5) and (6) with high contrasts. We start by introducing the framework needed for the a priori error analysis of the scheme, then we define the scheme for (5) and prove its convergence to the exact solution. Finally we define the scheme for (6) and derive an a posteriori error analysis. We also prove that for linear purely diffusive problems the schemes for (5) and (6) are equivalent. We would like to note that Chapter 7 depends on Chapter 6, but Chapter 8 is independent from Chapters 6 and 7.

In Chapter 6 we recall the gradient discretization method [41], it is a framework suitable for studying the convergence of gradient schemes for diffusion problems: linear and nonlinear, steady-state or transient. Then we recast the SWIP scheme [39, 49] into a gradient scheme, which gives rise to the Weighed Discontinuous Galerkin Gradient Discretization (WDGGD) scheme. To do so, we closely follow [52], where the Symmetric Interior Penalty method is written as a gradient scheme, called Discontinuous Galerkin Gradient Discretization (DGGD).

In Chapter 7 we define the local WDGGD scheme and show that it converges to the exact solution u under the assumptions that $f \in H^{-1}(\Omega)$ and $u \in H_0^1(\Omega)$. Here, the gradient discretization framework is convenient to decompose the sources of errors in the local problems. Furthermore, applying the pointwise estimates from [36] we can prove (in some particular cases) that the errors coming from the artificial boundary conditions are of higher order and depend only on the local regularity of the solution. The theoretical results are then corroborated by numerical experiments, where we investigate as well the computational efficiency of the scheme.

In Chapter 8 we define the local SWIP scheme for (6) and show its equivalence with the local WDGGD method of Chapter 7; therefore, the local SWIP scheme converges to the exact solution. Then, we develop an a posteriori error analysis for the local SWIP method applied to (6), in which the error estimators are used to bound the numerical error and as well to mark the local domains. A sequence of numerical experiments is devoted to determine heuristically the effectivity index of the error estimators and the computational efficiency of the scheme.

In Chapter 9 we summarize the main results and draw the conclusions of Part II of the thesis.

Stabilized explicit multirate methods for stiff deterministic and stochastic differential equations

Part I

In this part of the thesis, we present the stabilized explicit multirate methods for stiff deterministic and stochastic differential equations.

Stabilized explicit methods, more precisely the Runge–Kutta–Chebyshev (RKC) scheme, play a central role in this part of the thesis. Hence, we begin in Chapter 1 introducing the reader to stabilized explicit methods for deterministic and stochastic differential equations. Since in the following chapters we mostly use first-order schemes, we will focus on those. We briefly review higher order methods but closely study only the second-order RKC scheme, that is used in Chapter 2. Then we recall the SK-ROCK scheme for stochastic differential equations driven by white noise and introduce the new SK- τ -ROCK scheme for stochastic differential equations driven by discrete Poisson noise.

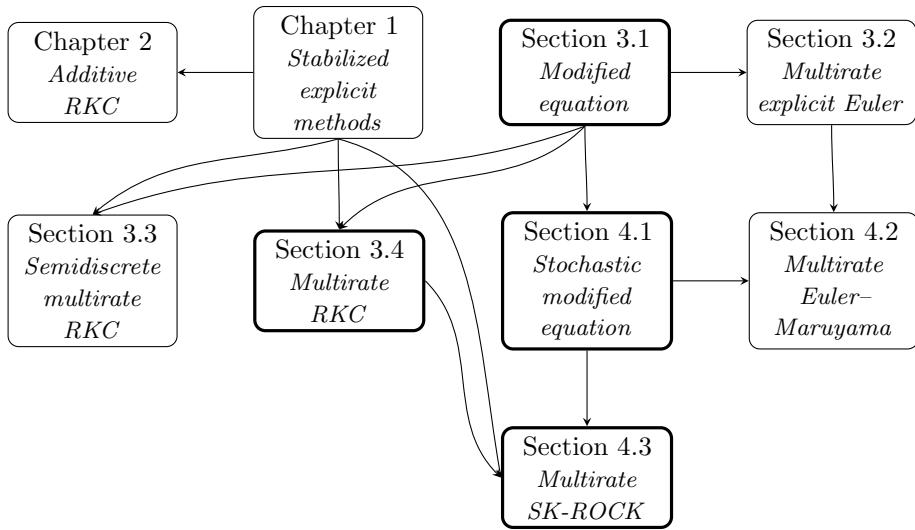
Stabilized schemes are usually more efficient than traditional explicit schemes, indeed their stability domain grows quadratically with the number of stages. Nonetheless, the number of function evaluations is still adversely affected by the problem’s stiffness; therefore, a few severely stiff degrees of freedom destroy the efficiency of stabilized schemes. Hence, in a problem composed by stiff and less stiff terms, the efficiency of stabilized schemes is recovered only if their evaluation is decoupled. One of the aims of this part of the thesis is indeed to develop stabilized explicit schemes which efficiency is not affected by a few severely stiff degrees of freedom.

In Chapter 2 we discuss an interpolation based multirate RKC scheme for stiff ordinary differential equations which, after some manipulation, can be partitioned into two subproblems. However, the literature of explicit multirate schemes contains many unstable interpolation based methods and this one is no exception. Indeed, we show on a model problem that interpolations might bring instabilities into the system and, furthermore, induce an order reduction phenomenon.

Our new methodology starts in Chapter 3, where we introduce a framework which allows for the development of a whole class of interpolation-free explicit multirate methods which do not need any scale separation assumption. We make use of a modified equation which stiffness depends solely on the slow dynamics and hence its integration by explicit methods requires less function evaluations. The modified right-hand side, called averaged force, is defined by the solution to a fast but cheap auxiliary problem. Any explicit scheme can be used to solve the modified equation and the auxiliary problem; in order to illustrate the flexibility of the framework we will present three alternatives: employment of the explicit Euler scheme for both problems, employment of RKC for both problems, or the using the RKC scheme for the modified equation and the exponential Euler method for the auxiliary problem.

In Chapter 4 we extend the presented methodology to multirate stochastic differential equations. To do that, we replace the drift term by the same averaged force used for deterministic equations and we define an approximation of the diffusion term such that the mean-square stability properties of the original problem are inherited by the stochastic modified equation. The approximated diffusion is also defined through solutions to deterministic auxiliary problems, as the averaged force. As for the deterministic case, many possibilities exist for the discretization of the stochastic modified equation and the deterministic fast auxiliary problems; indeed, we will introduce a multirate Euler–Maruyama method and the multirate SK-ROCK method, where the stabilized explicit SK-ROCK scheme is used for the stochastic modified equation and the RKC scheme is employed for the deterministic auxiliary problems. We will also see that not every Runge–Kutta method for stochastic differential equations can be used to derive an explicit multirate scheme.

Since there is an interlaced dependence between the sections of this part of the thesis, we propose a reading diagram here below, so that the reader can follow the preferred flow. Note that the highlighted sections are what we consider the most important ones. Nevertheless, we see in the diagram that, if someone is not interested on stabilized explicit methods, he can read only



Reading diagram illustrating dependencies between sections and the possible reading flows for Part I of the thesis. The highlighted boxes indicate sections containing the main contributions.

sections Sections 3.1 and 3.2 and readily have an idea of what the modified equation framework is and how it can be used to develop new multirate methods.

Chapter 2 is mainly taken from [14], Chapter 3 from [6] and Chapter 4 from [15]. Section 1.4.2 is based on [5, 55].

1 Stabilized explicit Runge–Kutta methods

The first part of this chapter is devoted to the introduction of stabilized explicit Runge–Kutta (RK) methods for ordinary differential equations (ODE)

$$y' = f(y), \quad y(0) = y_0, \quad (1.1)$$

where $y_0 \in \mathbb{R}^n$ is the initial value and $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is smooth enough to ensure the existence of solutions $y(t) \in \mathbb{R}^n$ in a bounded interval $t \in [0, T]$. We also assume that the eigenvalues of the Jacobian of f lie in a narrow strip along the negative real axis. Finally, the restriction to autonomous systems (1.1) is for simplicity only, as any nonautonomous system is recast into an autonomous one by appending the equation $t' = 1$.

Stabilized explicit methods are one-step RK methods with an extended stability domain along the negative real axis; therefore, when applied to (1.1) they have milder stability conditions than classical explicit methods. Many stabilized methods have been proposed in the literature, such as DUMKA methods, based on the composition of Euler steps [83, 84, 92], Runge–Kutta–Chebyshev (RKC) methods, based on the linear combination of Chebyshev polynomials [110, 118, 125] and orthogonal Runge–Kutta–Chebyshev methods (ROCK), based on optimal orthogonal stabilized functions [1, 9]; note that the ROCK and RKC methods differ only beyond order one.

Here we focus on first-order stabilized explicit schemes and start this chapter with the derivation of their optimal stability polynomials. Then we discuss RK schemes realizing such polynomials, putting attention on the first-order RKC scheme, which is the main method used in the chapters that follow. Next, we briefly discuss higher order schemes, concentrating on the second-order RKC scheme.

In the second and last part of this chapter we generalize first-order stabilized explicit methods to stiff stochastic differential equations (SDEs) driven by white or discrete Poisson noise. The S-ROCK schemes is a family of stabilized explicit methods for stochastic differential equations defined by Itô integrals [2, 8, 16], by Stratonovich integrals [3, 4] or by discrete noise [7]. In this chapter we first recall the SK-ROCK method [2], which is the only generalization of the first-order RKC scheme to SDEs which inherit the optimal stability properties of RKC. Then we introduce the new SK- τ -ROCK scheme [5, 55], which is an adaptation of the SK-ROCK scheme to discrete noise equations, driven by Poisson noise.

1.1 Optimal stability polynomials for first-order methods

We introduce the first-order stabilized explicit methods starting from their stability polynomial, which is the main tool for studying their stability properties.

The simplest method for the numerical solution of (1.1) is the explicit Euler method

$$y_{n+1} = y_n + \tau f(y_n),$$

where $\tau > 0$ is the step size, $t_n = t_{n-1} + \tau$, $t_0 = 0$ and y_n is an approximation of $y(t_n)$. When applied to the *Dahlquist test equation*

$$y' = \lambda y, \quad y(0) = y_0, \quad (1.2)$$

with $\lambda \in \mathbb{C}$, it yields $y_{n+1} = R(z)y_n$, where $z = \tau\lambda$ and $R(z) = 1 + z$. Recursive application then yields $y_n = R(z)^n y_0$. Since $y(t_n) = e^{t_n \lambda} y_0$, $|y(t_n)| \leq |y_0|$ if $\operatorname{Re}(\lambda) \leq 0$. A similar property is desirable for the numerical solution and it is clear that $|y_n| \leq |y_0|$ if, and only if, $|R(z)| \leq 1$. A similar derivation holds for any explicit RK method and we call

$$\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$$

the stability domain and $R(z)$ the stability polynomial of the method. For the stable numerical integration of (1.2), the step size τ has to satisfy $z = \tau\lambda \in \mathcal{S}$. For the explicit Euler method, for instance, and $\lambda \in \mathbb{R}^-$, it holds $z \in \mathcal{S}$ if and only if $z \in [-2, 0]$. Thus, for $\lambda < 0$ and $|\lambda|$ large, the stability condition $\tau \leq 2/|\lambda|$ imposes a very stringent restriction on the step size.

Similar step size restrictions hold for standard higher order explicit schemes, see for instance [69, Chapter IV.2]. Indeed, when the number of stages of standard explicit schemes is augmented the aim is, in general, to improve accuracy. In contrast, stabilized Runge–Kutta methods [1, 9, 83, 84, 92, 110, 118, 125] use an increased number of stages to improve stability properties along the negative real axis and thereby relax the stringent constraint of standard explicit RK methods on the step size.

Given $s \in \mathbb{N}$, the optimal stability polynomial $R_s(z)$ of a first-order s -stage stabilized scheme has degree s and solves the optimization problem:

$$R_s(0) = R'_s(0) = 1 \quad \text{and} \quad |R_s(z)| \leq 1 \text{ for } z \in [-\ell_s, 0] \text{ with } \ell_s \text{ maximal.} \quad (1.3)$$

The first condition in (1.3) is necessary to guarantee first-order accuracy, as an RK scheme has first-order if, and only if, its stability polynomial $R(z)$ satisfies $R(z) = e^z + \mathcal{O}(z^2)$. The second condition yields the longest stability domain along the negative real axis, for an s -stage RK method.

The solution to (1.3) was first proposed in [63, 104]. With the change of variables

$$R_s(z) = T_s \left(1 + \frac{2z}{\ell_s} \right) \quad \text{it holds} \quad T_s(1) = R_s(0) = 1 \quad \text{and} \quad 2T'_s(1) = \ell_s R'_s(0) = \ell_s.$$

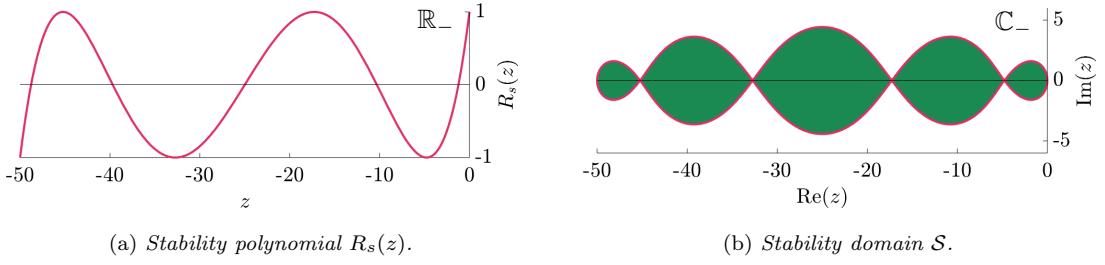
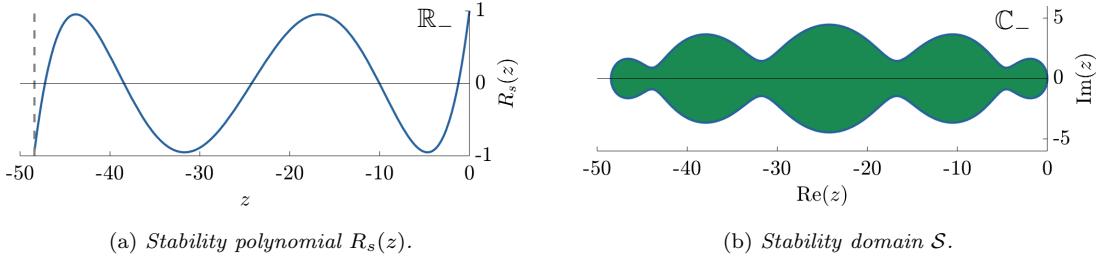
Hence, problem (1.3) is equivalent to solve the optimization problem of finding the polynomial $T_s(x)$ of degree s satisfying

$$T_s(1) = 1, \quad |T_s(x)| \leq 1 \quad \text{for } x \in [-1, 1] \quad \text{such that} \quad T'_s(1) \quad \text{is maximal} \quad (1.4)$$

and letting $R_s(z) = T_s(1 + 2z/\ell_s)$ with $\ell_s = 2T'_s(1)$.

A theorem of Markoff A. (1889) [89] states that a polynomial of degree s satisfying $|T_s(x)| \leq 1$ for $x \in [-1, 1]$ must satisfy $|T'_s(x)| \leq s^2$ for all $x \in [-1, 1]$, hence the solution to (1.4) satisfies $T'_s(1) \leq s^2$. Chebyshev polynomials of the first kind are defined recursively by

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_j(x) = 2xT_{j-1}(x) - T_{j-2}(x). \quad (1.5)$$

Figure 1.1. Stability polynomial and domain of the undamped $R_s(z) = T_s(1 + z/s^2)$ for $s = 5$.Figure 1.2. Stability polynomial and domain of the damped $R_s(z) = T_s(\omega_0)^{-1}T_s(\omega_0 + \omega_1 z)$ for $s = 5$ and $\varepsilon = 0.05$.

It is known that if $x \in [-1, 1]$ then $T_s(x) = \cos(s \arccos(x))$ and therefore $|T_s(x)| \leq 1$, moreover $T_s(1) = 1$ and $T'_s(1) = s^2$. Hence, Chebyshev polynomials solve (1.4) and

$$R_s(z) = T_s\left(1 + \frac{z}{s^2}\right) \quad (1.6)$$

solves (1.3) with $\ell_s = 2s^2$. Uniqueness of $R_s(z)$ follows from an equal ripple property [103], see also [117, Section 4.2.2].

It follows that, along \mathbb{R}^- , the size $\ell_s = 2s^2$ of the stability domain increases *quadratically* with the number of stages, as shown in Figure 1.1 for $s = 5$. However, we illustrate the stability domain of $R_s(z)$ in Figure 1.1(b) and observe distinct points with no stability in the imaginary direction, precisely where $R_s(x) = \pm 1$. Those intersection points can cause instabilities in the presence of nonlinearities or convection, for instance.

To avoid those discrete but critical points of potential instability, Guillou and Lago [63] (see also [124]) introduce a damping parameter $\varepsilon \geq 0$ and replace the second condition in (1.3) by

$$|R_s(z)| \leq 1 - \varepsilon \text{ for } z \in [-\ell_s^\varepsilon, -\delta] \text{ with } \ell_s^\varepsilon \text{ maximal and } \delta > 0 \text{ small.}$$

Note that $R_s(0) = 1$, hence $\delta > 0$ is needed. Approximate solutions to this problem are the scaled and shifted Chebyshev polynomials

$$R_s(z) = \frac{T_s(\omega_0 + \omega_1 z)}{T_s(\omega_0)}, \quad \text{with} \quad \omega_0 = 1 + \frac{\varepsilon}{s^2}, \quad \omega_1 = \frac{T_s(\omega_0)}{T'_s(\omega_0)}. \quad (1.7)$$

In Figure 1.2, we display the damped stability polynomial (1.7) and the associated stability domain for $s = 5$ and $\varepsilon = 0.05$; we note the improved stability at those intersection points. The definition of ω_0, ω_1 ensure that $R_s(0) = R'_s(0) = 1$ and hence the method is first-order accurate. Moreover $T_s(x)$ is increasing for $x \geq 1$, since $T_s(1) = 1$ then $T_s(\omega_0) > 1$ and we have $|R_s(z)| < 1$ for $z \in [-\ell_s^\varepsilon, 0]$, or more precisely:

$$|R_s(z)| \leq 1 - \varepsilon + \mathcal{O}(\varepsilon^2) \text{ for } z \in [-\ell_s^\varepsilon, -\delta].$$

The resulting stability domain now extends along the negative real axis with $\ell_s^\varepsilon = (1 + \omega_0)/\omega_1$, hence, if $|z| \leq \ell_s^\varepsilon$ then $|R_s(z)| \leq 1$. In fact, the size of the stability domain scales like s^2 , as it can be shown [124] that

$$\beta s^2 \leq \ell_s^\varepsilon, \quad \text{with} \quad \beta = 2 - \frac{4}{3}\varepsilon, \quad (1.8)$$

and therefore $|z| \leq \beta s^2$ is a sufficient condition for stability.

For $\varepsilon = 0$, we have $\omega_0 = 1$, $\omega_1 = 1/s^2$ and thus we recover the original (undamped) stability polynomial (1.6).

For $s = 1$ we recover the explicit Euler scheme, indeed we easily verify that $R_1(z) = 1 + z$, for any $\varepsilon \geq 0$.

1.2 First-order stabilized explicit methods

In the previous section we derived the optimal stability polynomials for first-order methods. Here, we introduce three schemes which indeed have $R_s(z)$ as stability polynomial, but before that we explain how the stability theory derived in the previous section extends to more general problems (1.1) and also why stabilized methods are advantageous compared to traditional explicit schemes.

1.2.1 Stability condition and computational advantages

Let us briefly discuss the stability condition for the more general problem (1.1). When (1.1) is linear, the stability condition for RK methods is derived using the Schur decomposition of the Jacobian of f [69, Chapter IV.2]. Hence, if f is linear and its Jacobian has only negative real eigenvalues, a stabilized scheme applied to (1.1) is stable if, and only if,

$$\tau\rho \leq \ell_s^\varepsilon, \quad (1.9)$$

where ρ is the spectral radius of the Jacobian of f . In practice, the sufficient stability condition

$$\tau\rho \leq \beta s^2 \quad (1.10)$$

is used, as it allows to choose s without computing ω_0, ω_1 from (1.7). For nonlinear problems (1.1), stability of RK schemes is studied under additional assumptions on the right-hand side f , namely a one-sided Lipschitz condition [69, Chapter IV.12]. However, upon linearization of f and neglecting higher order terms one recovers condition (1.10). In this thesis we will use conditions (1.9) and (1.10) even for nonlinear f , as is generally done in practice.

The stability condition (1.10) leads to many advantages of stabilized methods compared to standard explicit methods. For instance, when integrating (1.1), we wish to choose the step size τ according to the required accuracy. In Section 1.1 we saw that the step size of the explicit Euler scheme, when applied to the test equation (1.2), is bounded by $\tau \leq 2/|\lambda|$. When it is applied to (1.1) it must satisfy $\tau \leq 2/\rho$, where ρ is the spectral radius of the Jacobian of f . For diffusion dominated problems $2/\rho$ scales as h^2 , where h is the size of the smallest elements in the mesh, for example. Hence, the user is often forced to take a step size much smaller than that which satisfies the required accuracy, thereby incurring a high computational cost. In contrast, the stability condition of stabilized schemes is $\tau\rho \leq \beta s^2$; hence, the step size can be truly chosen according to the desired accuracy, as long as s is large enough.

We can also compare the computational costs of the explicit Euler (EE) and a first-order stabilized explicit (SE) method. Suppose that we integrate (1.1) in the interval $[0, 1]$ and that $\tau \leq 1$ is the

step size needed to attain a desired accuracy. We consider a situation where $\tau\rho \geq 2$; otherwise $s = 1$, the problem is nonstiff and the SE schemes are indeed equivalent to the EE scheme. Therefore, the EE scheme uses a step size $\tilde{\tau} = 2/\rho \leq \tau$ smaller than the one sufficient for the required accuracy. As the cost per time-step of EE is one function evaluation and the number of steps is $1/\tilde{\tau}$, the total integration cost for the EE scheme is

$$C_{\text{EE}}^T = \frac{1}{\tilde{\tau}} = \frac{\rho}{2}.$$

In contrast, the SE schemes can truly take τ as step size. For simplicity we let s vary in \mathbb{R} and consider an undamped stabilized scheme, hence $\beta = 2$. The cost of one time-step, in terms of function evaluations, is

$$C_{\text{SE}} = s = \sqrt{\frac{\tau\rho}{2}}.$$

As the number of time steps is $1/\tau$ then the total cost for an SE scheme is

$$C_{\text{SE}}^T = \frac{1}{\tau} C_{\text{SE}} = \sqrt{\frac{\rho}{2\tau}}. \quad (1.11)$$

We define the theoretical relative speed-up as the ratio between the two costs, hence

$$S = \frac{C_{\text{EE}}^T}{C_{\text{SE}}^T} = \sqrt{\frac{\tau\rho}{2}}. \quad (1.12)$$

In (1.12) we observe two important things. First, since we supposed $\tau\rho \geq 2$ then $S \geq 1$. Hence, as the step size of the EE scheme is limited by stability constraints, then an SE scheme is faster than the EE scheme. Otherwise they are equivalent. Second, the efficiency of SE schemes increases as they take larger step sizes τ ; the reason for that is the quadratic increase, in terms of function evaluations, of the stability domain's size.

Note that for stiff problems $\tau\rho$ is large, hence the relative speed-up S in (1.12) is important. For instance, if we consider the case where f contains a discretized Laplacian and ρ scales as C/h^2 , where h is the smallest mesh size, the efficiency gain in using a stabilized method over the simple Euler method scales as $\sqrt{\tau}/h$. The same arguments apply to classic higher order explicit schemes such as the midpoint, RK4 and DOPRI5 methods, to name a few. For higher order stabilized explicit methods we refer to [1, 9, 92, 110] and Section 1.3 below.

Now, let us introduce three first-order stabilized methods. Following the chronological order of appearance, we start with the factorization scheme of Guillou and Lago (1960) [63] and Saul'ev (1960) [104]. Then we introduce the diagonal method of Van der Houwen (1977) [117] and finally the most recent Runge–Kutta–Chebyshev (RKC) scheme of Van der Houwen and Sommeijer (1980) [118], which is the most used in practice. In the rest of this chapter we mainly follow [69, 118].

1.2.2 Factorization methods

The idea behind factorization methods [63, 104] is to factorize the stability polynomial, hence write

$$R_s(z) = \prod_{j=1}^s (1 + \delta_j z), \quad \text{where} \quad \delta_j = -\frac{1}{z_j} \text{ and } z_j \text{ are roots of } R_s(z). \quad (1.13)$$

Note that z_j , $j = 1, \dots, s$, are easily computable thanks to (1.7) and $T_s(x) = \cos(s \arccos(x))$ for $x \in [-1, 1]$. Hence, the roots z_j are well distributed in the interval $[-\ell_s^\varepsilon, 0]$, see also Figure 1.2(a).

Thanks to (1.13) an RK method having $R_s(z)$ as stability polynomial is realized as a composition of explicit Euler steps

$$k_j = k_{j-1} + \delta_j \tau f(k_{j-1}), \quad j = 1, \dots, s, \quad (1.14)$$

where $k_0 = y_n$ and $y_{n+1} = k_s$. The disadvantage of scheme (1.14) is that although the final stage is stable the internal stages k_j might be unstable. If (1.14) is applied to the test equation (1.2), the stage k_j is given by the internal stability polynomial $R_j(z)$, i.e.

$$k_j = R_j(z)y_n, \quad \text{with} \quad R_j(z) = \prod_{i=1}^j (1 + \delta_i z).$$

If, for instance, the first root z_1 is small in magnitude (Figure 1.2(a)) then the associated Euler step is large and the first stage $R_1(z) = 1 + \delta_1 z$ is unstable. Therefore, the roots z_j must be properly ordered. In [82, 83] Lebedev combines the roots symmetrically and computes a quadratic factor

$$(1 + \delta_j z)(1 + \delta_{s-j+1} z) = 1 + (\delta_j + \delta_{s-j+1})z + \delta_j \delta_{s-j+1} z^2,$$

which is realized by the two-stage scheme

$$\begin{aligned} k_j &= k_{j-1} + \alpha_j \tau f(k_{j-1}), \\ k_{j+1} &= k_j + \alpha_j \tau f(k_j), \\ k_{j+2} &= k_{j+1} - \gamma_j(k_{j+1} + k_{j-1} - 2k_j), \end{aligned} \quad (1.15)$$

where $\alpha_j = (\delta_j + \delta_{s-j+1})/2$ and $\alpha_j^2(1 - \gamma_j) = \delta_j \delta_{s-j+1}$. If s is odd, an additional Euler step is performed. With this approach, the largest Euler step is, approximately, halved. Indeed, a large step δ_j is combined with a small step δ_{s+j-1} and therefore $\alpha_j \approx \delta_j/2$.

The scheme can be further improved optimizing the order in which the two-stage procedures (1.15) are computed and thereby obtaining truly stable internal stages. However, the optimal order is found numerically and depends on the number of stages s , which is not practical for applications. We refer to [69, 82] for further details.

1.2.3 Diagonal methods

Diagonal methods [117] take the form

$$k_j = k_0 + \tilde{\delta}_j \tau f(k_{j-1}), \quad j = 1, \dots, s, \quad (1.16)$$

where $k_0 = y_n$, $y_{n+1} = k_s$. The name comes from their subdiagonal Butcher tableau, this choice has been made in order to reduce to a minimum the memory requirements. When (1.16) is applied to the test equation (1.2) with $y_0 = 1$ it yields

$$k_j = 1 + \sum_{i=1}^j z^i \prod_{l=j-i+1}^j \tilde{\delta}_l \quad (1.17)$$

and the coefficients $\tilde{\delta}_j$ are uniquely identified by comparing (1.17) for $j = s$ with the coefficients of $R_s(z)$. It is pointed out in [118] that one should not use this scheme with values of s higher than 12 due to strong internal instabilities. We note as well that for this method there is no room for reordering nor improvement.

1.2.4 First-order Runge–Kutta–Chebyshev methods

The Runge–Kutta–Chebyshev (RKC) method has been derived in [118] using the recursive definition of the Chebyshev polynomials $T_s(x)$.

Three-term recurrence formulation.

The aim of the RKC method is not only to obtain a stable step but stable internal stages too. In [118], they let

$$R_j(z) = \frac{T_j(\omega_0 + \omega_1 z)}{T_j(\omega_0)}, \quad j = 0, \dots, s-1, \quad (1.18)$$

with ω_0, ω_1 as in (1.7). Similarly to $R_s(z)$, also $|R_j(z)| \leq 1$ for all $z \in [-\ell_s^\varepsilon, 0]$ and $j = 0, \dots, s-1$; it is therefore advantageous to define a scheme which has $R_j(z)$ as internal stability polynomials, i.e. when the scheme is applied to the test equation (1.2), the j th internal stage k_j is given by $R_j(z)y_n$.

It follows from (1.5) that the shifted and scaled Chebyshev polynomials $R_j(z)$ satisfy the recurrence relation

$$\begin{aligned} R_0(z) &= 1, \\ R_1(z) &= \frac{\omega_0 + \omega_1 z}{\omega_0}, \\ R_j(z) &= 2 \frac{T_{j-1}(\omega_0)}{T_j(\omega_0)} (\omega_0 + \omega_1 z) R_{j-1}(z) - \frac{T_{j-2}(\omega_0)}{T_j(\omega_0)} R_{j-2}(z), \quad j = 2, \dots, s. \end{aligned} \quad (1.19)$$

Letting $b_j = 1/T_j(\omega_0)$ for $j = 0, \dots, s$ and

$$\begin{aligned} \mu_1 &= \omega_1/\omega_0, & \mu_j &= 2\omega_1 b_j/b_{j-1}, \\ \nu_j &= 2\omega_0 b_j/b_{j-1}, & \kappa_j &= -b_j/b_{j-2}, \end{aligned} \quad \text{for } j = 2, \dots, s, \quad (1.20)$$

(1.19) can be written as

$$\begin{aligned} R_0(z) &= 1, \\ R_1(z) &= R_0(z) + \mu_1 z R_0(z), \\ R_j(z) &= \nu_j R_{j-1}(z) + \kappa_j R_{j-2}(z) + \mu_j z R_{j-1}(z), \quad j = 2, \dots, s. \end{aligned} \quad (1.21)$$

The first-order RKC method.

From (1.21) follows the first-order RKC scheme:

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau f(k_0), \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j \tau f(k_{j-1}), \quad j = 2, \dots, s, \\ y_{n+1} &= k_s, \end{aligned} \quad (1.22)$$

where $s \in \mathbb{N}$ is such that $\tau\rho \leq \beta s^2$, with ρ the spectral radius of the Jacobian of f . Since, for the test equation (1.2), (1.21) and (1.22) are equivalent it follows that the stability polynomial of the RKC scheme is indeed $R_s(z)$. Furthermore, the internal stability polynomials are $R_j(z)$ and therefore the internal stages are stable, for all choice of s . We also note that $R_j(z) = e^{c_j z} + \mathcal{O}(z^2)$, with $c_j = R'_j(0)$. Hence, the internal stages k_j in (1.22) are first-order approximations to $y(t_n + c_j \tau)$ [125].

For a nonautonomous system $y' = f(t, y)$, the RKC scheme is given by (1.22) but where $f(k_j)$ is replaced by $f(t_n + c_j \tau, k_j)$. In order to avoid the computation of $R'_j(0)$, the coefficients c_j are also computed appending the equation $t' = 1$ to (1.1), which yields

$$\begin{aligned} t_n + c_0 \tau &= t_n, \\ t_n + c_1 \tau &= t_n + \mu_1 \tau, \\ t_n + c_j \tau &= \nu_j (t_n + c_{j-1} \tau) + \kappa_j (t_n + c_{j-2} \tau) + \mu_j \tau, \quad j = 2, \dots, s. \end{aligned} \quad (1.23)$$

Using $\nu_j + \kappa_j = 1$ (1.23) is equivalent to

$$\begin{aligned} c_0 &= 0, \\ c_1 &= \mu_1, \\ c_j &= \nu_j c_{j-1} + \kappa_j c_{j-2} + \mu_j, \quad j = 2, \dots, s. \end{aligned}$$

Modeling internal stage perturbations.

The stability polynomial $R_s(z)$ of the RKC schemes is the main tool for studying the evolution of perturbations from one step to the next one. However, for schemes with a high number of internal stages it is also important to study the evolution of internal perturbations, from one stage to the next one. For RKC schemes this analysis has been done in [125] and the goal was twofold: bound round-off errors introduced at each stage and show that the schemes are stiffly accurate. In this thesis, we will use the same perturbation analysis with a different aim: for proving stability of our multirate schemes, as they can also be seen as perturbed RKC methods. Let us recall the main results obtained in [125].

We consider the RKC scheme (1.22) for the test equation (1.2) and where each stage k_j , $j = 1, \dots, s$, is perturbed by a quantity r_j . Letting $z = \tau\lambda$ it yields

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 z k_0 + r_1, \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j z k_{j-1} + r_j, \quad j = 2, \dots, s, \\ y_{n+1} &= k_s, \end{aligned} \tag{1.24}$$

Using (1.20) it is shown recursively in [125] that

$$\begin{aligned} k_j &= b_j T_j(\omega_0 + \omega_1 z) y_n + \sum_{k=1}^j \frac{b_j}{b_k} U_{j-k}(\omega_0 + \omega_1 z) r_j, \\ &= R_j(z) y_n + \sum_{k=1}^j \frac{b_j}{b_k} U_{j-k}(\omega_0 + \omega_1 z) r_j, \end{aligned} \tag{1.25}$$

where $U_k(x)$ are the Chebyshev polynomials of the second kind of degree k , defined recursively by

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_j(x) = 2xU_{j-1}(x) - U_{j-2}(x). \tag{1.26}$$

They satisfy $U_j(x) \in [-(j+1), j+1]$ for $x \in [-1, 1]$,

$$U_j(1) = j+1, \quad U'_j(1) = \frac{j(j+1)(j+2)}{3}$$

and are increasing for $x \geq 1$. Furthermore, in most of the interval $(-1, 1)$, i.e. away from the extremities, $U_j(x)$ oscillates in $[-1, 1]$.

The above result, (1.25), is used in [125], for instance, to bound the round-off errors. Indeed, it holds

$$\begin{aligned} |U_j(\omega_0 + \omega_1 z)| &\leq |U_j(\omega_0)| = j+1 + \frac{j(j+1)(j+2)}{3} \frac{\varepsilon}{s^2} + \mathcal{O}(\varepsilon^2) \\ &= (j+1) \left(1 + \frac{j(j+2)}{3s^2} \varepsilon \right) + \mathcal{O}(\varepsilon^2) \end{aligned}$$

and $|U_j(\omega_0 + \omega_1 z)| \leq (j+1)(1+C\varepsilon)$, for a small constant $C > 0$ and ε sufficiently small; similar bounds hold for the coefficients b_j [125]. From (1.25) follows

$$|y_{n+1}| \leq |y_n| + \sum_{k=1}^s (s-k+1)(1+C\varepsilon)|r_j| \leq |y_n| + \frac{1}{2}s(s+1)(1+C\varepsilon) \max_{j=1,\dots,s} |r_j|. \quad (1.27)$$

Estimate (1.27) shows that round-off errors r_j committed in the internal stages k_j , see (1.24), are amplified by a factor growing as s^2 . Hence, if a round-off error has size 10^{-16} and $s \leq 10^3$ then at the end of the integration step it is amplified to, approximately, 10^{-10} . Which is still small. Nevertheless, as we already said, in this thesis we will use relations (1.24) and (1.25) mainly for studying stability of the multirate methods.

1.3 Higher order methods

Here we briefly comment on higher order stabilized explicit methods, which enjoy an extended stability domain along the negative real axis that grows quadratically with the number of stages, as the first-order stabilized methods seen in Section 1.2. Among those, we define only the second-order Runge–Kutta–Chebyshev scheme.

1.3.1 An overview of higher order stabilized methods

Optimal stability polynomials $R_s^p(z)$ with higher order of consistency, i.e. $R_s^p(z) = e^z + \mathcal{O}(z^{p+1})$ for $p > 1$, are defined as solutions to an optimization problem similar to (1.3). In [103] it is shown that such polynomials exist and are unique; however, an analytical expression is not known and therefore the derivation of optimal higher order stabilized RK methods is more involved.

For second-order methods, Lebedev and Medovikov [84] express the optimal stability polynomial as an elliptic integral. They find that the optimal stability domain grows as $0.82s^2$, approximately. Identifying numerically the roots of the stability polynomials they derive the DUMKA methods [83, 84] as factorization methods. Medovikov derives also third- and fourth-order DUMKA methods in [92]. However, the internal stability of DUMKA type schemes depends on the order of evaluation of the two-stage scheme (1.15), which depends on s and is found numerically. Second-order Runge–Kutta–Chebyshev methods are based on the linear combination of Chebyshev polynomials [110, 118, 125], their stability polynomial is not optimal but leads to a stability domain that grows as $0.65s^2$, hence covering roughly 80% of the optimal stability domain. Compared to the DUMKA methods, the coefficients of the RKC schemes are defined analytically and need not reordering. Abdulle and Medovikov [9] derive the second-order orthogonal Runge–Kutta–Chebyshev methods (ROCK), based on optimal orthogonal stabilized functions. They are based on a three-term recurrence relation as the RKC schemes and therefore the internal stages are stable. The recurrence coefficients of the ROCK methods are computed numerically beforehand and taken from a table whenever needed. The stability domain of the second-order ROCK scheme grows as $0.81s^2$, hence it is essentially optimal. In [1], Abdulle derives the fourth-order ROCK schemes, whose stability domain grows as $0.35s^2$ and is as well nearly optimal. Note that the first-order ROCK and RKC methods are equivalent.

1.3.2 Second-order Runge–Kutta–Chebyshev methods

We define here the second-order RKC method, which is similar to the first-order RKC method but with additional degrees of freedom so that second-order accuracy is achieved; at the cost of shortening the stability domain. First, we derive the stability polynomial and then we define the method.

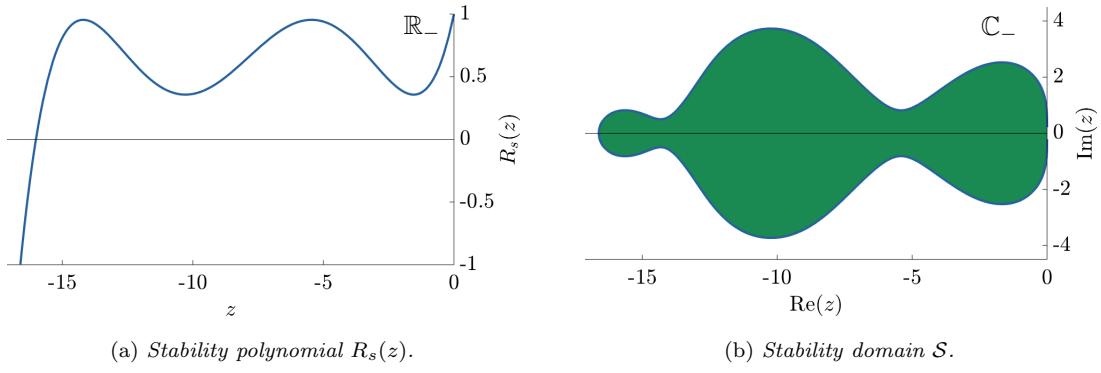


Figure 1.3. *Stability polynomial and domain of the damped $R_s(z) = a_s + b_s T_s(\omega_0 + \omega_1 z)$ for $s = 5$ and $\varepsilon = 0.15$.*

Derivation of the stability polynomial

It was pointed out by Bakker (1971) [24] (see also [117, Section 4.2.3]) that appropriate linear combinations of Chebyshev polynomials lead to high order consistent polynomials (i.e. $R_s^p(z) = e^z + \mathcal{O}(z^{p+1})$ for $p > 1$) with relatively long stability domains; for instance, the polynomial

$$\bar{R}_s(z) = \frac{2s^2 + 1}{3s^2} + \frac{s^2 - 1}{3s^2} T_s \left(1 + \frac{3z}{s^2 - 1} \right) \quad (1.28)$$

satisfies the second-order conditions

$$\bar{R}_s(0) = \bar{R}'_s(0) = \bar{R}''_s(0) = 1 \quad (1.29)$$

and has a stability domain that grows as $2(s^2 - 1)/3$. We recall that if a stability polynomial satisfy (1.29) then an associated RK scheme is second-order accurate [69, IV.2].

However, $\bar{R}_s(z)$ as defined in (1.28) is not strongly stable, i.e. there are some points z where $|\bar{R}_s(z)| = 1$, and whenever $|\bar{R}_s(z)| = 1$ there is no stability along the imaginary direction; as in Figure 1.1 for first-order methods. In order to obtain a strongly stable stability polynomial satisfying $|\bar{R}_s(z)| \leq 1 - \varepsilon$ for $\varepsilon > 0$ and z away from zero, inspired by (1.28) in [118] the authors let

$$\bar{R}_s(z) = a_s + b_s T_s(\omega_0 + \omega_1 z) \quad (1.30)$$

and uniquely identify $a_s, b_s, \omega_0, \omega_1$ from the stability condition $a_s + b_s = 1 - \varepsilon$ and the order conditions (1.29). However, this leads to rather complicated expressions for $a_s, b_s, \omega_0, \omega_1$. Here, we consider the simpler set of coefficients

$$\omega_0 = 1 + \frac{\varepsilon}{s^2}, \quad \omega_1 = \frac{T'_s(\omega_0)}{T''_s(\omega_0)}, \quad b_s = \frac{T''_s(\omega_0)}{T'_s(\omega_0)^2}, \quad a_s = 1 - b_s T_s(\omega_0) \quad (1.31)$$

proposed in [125], which satisfies the order conditions (1.29) and yields $a_s + b_s = 1 - \varepsilon/3 + \mathcal{O}(\varepsilon^2)$ with $a_s + b_s < 1$. Thus, $|\bar{R}_s(z)| \leq 1$ for all $z \in [-\ell_s^\varepsilon, 0]$ and $|\bar{R}_s(z)| \leq 1 - \varepsilon/3 + \mathcal{O}(\varepsilon^2)$ for $z \in [-\ell_s^\varepsilon, -\delta]$ and $\delta > 0$ small, where [125]

$$\ell_s^\varepsilon = \frac{1 + \omega_0}{\omega_1} \approx \bar{\beta}(s^2 - 1), \quad \text{with} \quad \bar{\beta} = \frac{2}{3} \left(1 - \frac{2}{15}\varepsilon \right). \quad (1.32)$$

Typically, $\varepsilon = 0.15$ and $\bar{\beta} \approx 0.65$, thus $\ell_s^\varepsilon \approx 0.65(s^2 - 1)$. We display in Figure 1.3 the stability polynomial and domain of the second-order RKC scheme for $s = 5$ and $\varepsilon = 0.15$.

Three-term recurrence formulation.

In [118] the recurrence relation of the Chebyshev polynomials is used in order to derive an RK scheme which realizes $\bar{R}_s(z)$ and have stable internal stages, as the first-order RKC method. Nevertheless, in [118] the internal stages are only first-order accurate. In [111], it was noted that the polynomials

$$\bar{R}_j(z) = a_j + b_j T_j(\omega_0 + \omega_1 z), \quad j = 0, \dots, s-1, \quad (1.33)$$

defined by

$$\begin{aligned} a_j &= 1 - b_j T_j(\omega_0), & \text{for } j = 0, \dots, s-1, \\ b_0 = b_1 = b_2, & \quad b_j = \frac{T''_j(\omega_0)}{T'_j(\omega_0)}, & \text{for } j = 2, \dots, s-1 \end{aligned} \quad (1.34)$$

satisfy $\bar{R}'_j(0) = c_j$ and $\bar{R}''_j(0) = c_j^2$ for $j = 2, \dots, s$, with

$$c_j = \frac{T''_j(\omega_0)T'_s(\omega_0)}{T'_j(\omega_0)T''_s(\omega_0)}, \quad (1.35)$$

see also [125]. Hence $\bar{R}_j(z) = e^{c_j z} + \mathcal{O}(z^3)$ and it is therefore convenient to have $\bar{R}_j(z)$ as internal stability polynomial, as the associated internal stage would be second-order accurate at the point $t_n + \tau c_j$. From (1.5) follows that $\bar{R}_j(z)$ for $j = 0, \dots, s$ defined as in (1.30), (1.31), (1.33) and (1.34) satisfy the recurrence relations

$$\begin{aligned} \bar{R}_0(z) &= 1, \\ \bar{R}_1(z) &= 1 + \mu_1 z, \\ \bar{R}_j(z) &= \nu_j \bar{R}_{j-1}(z) + \kappa_j \bar{R}_{j-2}(z) + \mu_j z \bar{R}_{j-1}(z) - \mu_j a_{j-1} z + 1 - \nu_j - \kappa_j, \quad j = 2, \dots, s, \end{aligned} \quad (1.36)$$

with

$$\begin{aligned} \mu_1 &= b_1 \omega_1, & \mu_j &= 2\omega_1 \frac{b_j}{b_{j-1}}, \\ \nu_j &= 2\omega_0 \frac{b_j}{b_{j-1}}, & \kappa_j &= -\frac{b_j}{b_{j-2}}, & j &= 2, \dots, s. \end{aligned}$$

The second-order RKC method

From (1.36) follows the second-order RKC scheme

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau f(k_0), \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + (1 - \nu_j - \kappa_j) k_0 \\ &\quad + \mu_j \tau f(k_{j-1}) - \mu_j a_{j-1} \tau f(k_0), \quad j = 2, \dots, s, \\ y_{n+1} &= k_s, \end{aligned} \quad (1.37)$$

where s is such that $\tau \rho \leq \bar{\beta}(s^2 - 1)$ (see (1.32)), with ρ the spectral radius of the Jacobian of f .

Since the stages k_j are approximations to $y(t_n + c_j \tau)$, for nonautonomous systems $y' = f(t, y)$ in (1.37) we replace $f(k_j)$ by $f(t_n + c_j \tau, k_j)$. In order to avoid the computation of (1.35), the coefficients c_j for $j = 0, \dots, s$ are again easily computed recursively. Integrating $t' = 1$ with scheme (1.37) yields

$$\begin{aligned} t_n + c_0 \tau &= t_n, \\ t_n + c_1 \tau &= t_n + \mu_1 \tau, \\ t_n + c_j \tau &= \nu_j (t_n + c_{j-1} \tau) + \kappa_j (t_n + c_{j-2} \tau) + \mu_j \tau - \mu_j a_{j-1} \tau + (1 - \nu_j - \kappa_j) t_n, \quad j = 2, \dots, s, \end{aligned}$$

which is equivalent to

$$\begin{aligned} c_0 &= 0, \\ c_1 &= \mu_1, \\ c_j &= \nu_j c_{j-1} + \kappa_j c_{j-2} + \mu_j - \mu_j a_{j-1}, \quad j = 2, \dots, s. \end{aligned}$$

1.4 Stabilized explicit methods for stochastic differential equations

We end this chapter presenting stabilized explicit methods for stiff stochastic differential equations. In Section 1.4.1 we consider equations driven by white noise and we recall the SK-ROCK method already introduced in [2]. In Section 1.4.2 we consider discrete Poisson noise and introduce the SK- τ -ROCK scheme [5, 55], which is an adaptation of the SK-ROCK scheme to discrete noise.

1.4.1 The second kind orthogonal Runge–Kutta–Chebyshev method

We recall here the SK-ROCK scheme for the stiff stochastic differential equation

$$dX = f(X) dt + g(X) dW, \quad X(0) = X_0, \quad (1.38)$$

where $X(t)$ is a stochastic process in \mathbb{R}^n , $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the drift term, $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ is the diffusion term and $W(t)$ is an m -dimensional Wiener process.

Since (1.38) is stiff, classical explicit schemes as Euler–Maruyama or the Milstein–Talay method are overly expensive and one must resort to stabilized explicit schemes [2, 8, 16]. As the first-order RKC scheme can be seen as a stabilization of the explicit Euler method by Chebyshev polynomials, the S-ROCK schemes stabilize the Euler–Maruyama scheme (or higher order schemes) with Chebyshev polynomials as well. Depending on how the noise term is introduced in the Chebyshev iteration, different mean-square stability properties arise. In [8, 16], for instance, the stability domain's size L_s grows quadratically with the number of function evaluations, i.e. $L_s \approx \beta s^2$, but the constant β is much smaller than the optimal value of 2. For the SK-ROCK method [2], instead, the noise term is introduced in a way that the optimal stability domain's size of RKC schemes is recovered.

In the remaining of this section we will first define the SK-ROCK scheme and then we point out its main stability and accuracy properties. At the end we closely compare the stability properties of the SK-ROCK scheme and the Euler–Maruyama method.

The SK-ROCK method

Given the spectral radius ρ of the Jacobian of f and a step size $\tau > 0$ let $s \in \mathbb{N}$ such that $\tau\rho \leq \beta s^2$, μ_j, ν_j, κ_j for $j = 2, \dots, s$ be as in (1.20) and

$$\mu_1 = \frac{\omega_1}{\omega_0}, \quad \nu_1 = s \frac{\omega_1}{2}, \quad \kappa_1 = s \frac{\omega_1}{\omega_0}. \quad (1.39)$$

One step of the SK-ROCK method is given by

$$\begin{aligned} K_0 &= X_n, \\ K_1 &= K_0 + \mu_1 \tau f(K_0 + \nu_1 Q) + \kappa_1 Q, \\ K_j &= \nu_j K_{j-1} + \kappa_j K_{j-2} + \mu_j \tau f(K_{j-1}), \quad j = 2, \dots, s, \\ X_{n+1} &= K_s, \end{aligned} \quad (1.40)$$

where $Q = g(K_0)\Delta W_n$ and $\Delta W_n = W(t_{n+1}) - W(t_n)$. The SK-ROCK method has strong order 1/2 and weak order 1, see [2].

Mean-square stability analysis

The stability properties of the SK-ROCK method are derived by considering the *stochastic test equation*

$$dX(t) = \lambda X(t) dt + \mu X(t) dW(t), \quad X(0) = X_0, \quad (1.41)$$

with $X(t) \in \mathbb{R}$ and $\lambda, \mu \in \mathbb{C}$. The exact solution to (1.41) is called mean-square stable if $\lim_{t \rightarrow \infty} \mathbb{E}(|X(t)|^2) = 0$ and this holds if, and only if, $(\lambda, \mu) \in \mathcal{S}^{MS}$, where

$$\mathcal{S}^{MS} = \{(\lambda, \mu) \in \mathbb{C}^2 : \operatorname{Re}(\lambda) + \frac{1}{2}|\mu|^2 < 0\}. \quad (1.42)$$

Let $p = \lambda\tau$, $q = \mu\tau^{1/2}$ and $\xi_n \sim \mathcal{N}(0, 1)$. If the SK-ROCK method is applied to (1.41) it yields

$$\begin{aligned} K_0 &= X_n, \\ K_1 &= K_0 + \mu_1 p K_0 + r_1, \\ K_j &= \nu_j K_{j-1} + \kappa_j K_{j-2} + \mu_j p K_{j-1}, \quad j = 2, \dots, s, \end{aligned} \quad (1.43)$$

where $r_1 = (\mu_1 \nu_1 p + \kappa_1) q \xi_n X_n$. We observe that (1.43) is a perturbed RKC scheme, as (1.24). From (1.25) we thus deduce

$$K_j = b_j T_j(\omega_0 + \omega_1 p) X_n + \frac{b_j}{b_1} U_{j-1}(\omega_0 + \omega_1 p) r_1.$$

The identity $T'_n(x) = n U'_{n-1}(x)$, $b_j = T_j(\omega_0)^{-1}$, (1.7), (1.39) and $X_{n+1} = K_s$ imply

$$X_{n+1} = R_s(p, q, \xi_n) X_n, \quad \text{with} \quad R_s(p, q, \xi) = A_s(p) + B_s(p) q \xi \quad (1.44)$$

and

$$A_s(p) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)}, \quad B_s(p) = \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)} \left(1 + \frac{\omega_1}{2} p\right). \quad (1.45)$$

Therefore, the SK-ROCK scheme is mean-square stable, i.e. $\lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^2) = 0$, if, and only if, $(p, q) \in \mathcal{S}_{num}$, where

$$\mathcal{S}_{num} = \{(p, q) \in \mathbb{C}^2 : \mathbb{E}(|R_s(p, q, \xi)|^2) < 1\}, \quad \text{with} \quad \mathbb{E}(|R_s(p, q, \xi)|^2) = A_s(p)^2 + B_s(p)^2 |q|^2.$$

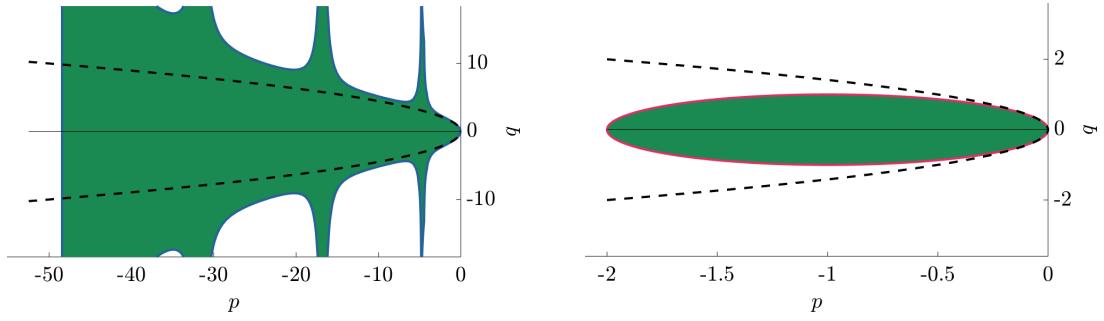
In [2] the following stability result is proved.

Theorem 1.1. *Let the damping $\varepsilon = 0$ and $(p, q) \in \mathcal{S}^{MS}$. For all $s \in \mathbb{N}$ such that $|p| \leq \ell_s^\varepsilon$ we have $\mathbb{E}(|R_s(p, q, \xi)|^2) \leq 1$, i.e. the SK-ROCK method is mean-square stable.*

For bounded $\varepsilon > 0$ the same result is proved under an additional condition on s and numerical evidences show that stability holds for any parameters ε and s satisfying the stability condition.

Observe that $(p, q) \in \mathcal{S}^{MS}$ and $(\lambda, \mu) \in \mathcal{S}^{MS}$ are equivalent. Hence, Theorem 1.1 states that, if the continuous equation (1.41) is mean-square stable, i.e. $(\lambda, \mu) \in \mathcal{S}^{MS}$, then the method is stable provided that s satisfies the stability condition $\tau|\lambda| \leq \ell_s^\varepsilon$. Therefore, we can say that the stability condition of the SK-ROCK method is the same as the one for the RKC scheme.

We display in Figure 1.4(a) the stability domain \mathcal{S}_{num} of the SK-ROCK scheme with $s = 5$ stages and damping $\varepsilon = 0.05$, where the dashed lines represent the border of the mean-square stability



(a) *Stability domain \mathcal{S}_{num} of the SK-ROCK scheme with $s = 5$ stages and damping $\varepsilon = 0.05$. The dashed lines are partially inside \mathcal{S}_{num} .*
 (b) *Stability domain \mathcal{S}_{num}^{EM} of the Euler–Maruyama scheme. The dashed lines are never inside \mathcal{S}_{num}^{EM} .*

Figure 1.4. *Stability domains of the SK-ROCK and Euler–Maruyama schemes. The dashed lines represent the boundaries of the mean-square stability domain \mathcal{S}^{MS} defined in (1.42).*

domain \mathcal{S}^{MS} of the stochastic test equation (1.41). We observe that \mathcal{S}_{num} contains a portion of \mathcal{S}^{MS} . More precisely, $\mathcal{S}_{\ell_s^{\varepsilon}}^{MS} \subset \mathcal{S}_{num}$, where

$$\mathcal{S}_a^{MS} = \mathcal{S}^{MS} \cap [-a, 0] \times \mathbb{R}, \quad (1.46)$$

i.e. \mathcal{S}_a^{MS} is \mathcal{S}^{MS} truncated at $p = -a$.

Comparison with Euler–Maruyama

We end this section pointing out two key differences, by a stability point of view, between the SK-ROCK method (1.40) and the Euler–Maruyama (EM) scheme

$$X_{n+1} = X_n + \tau f(X_n) + g(X_n) \Delta W_n. \quad (1.47)$$

When (1.47) is applied to the stochastic test equation (1.41) it yields

$$X_{n+1} = (1 + p + q\xi_n) X_n,$$

with $p = \lambda\tau$, $q = \mu\tau^{1/2}$ and $\xi_n \sim \mathcal{N}(0, 1)$. Hence, the method is mean-square stable if, and only if, $(p, q) \in \mathcal{S}_{num}^{EM}$, where

$$\mathcal{S}_{num}^{EM} = \{(p, q) \in \mathbb{C}^2 : \mathbb{E}(|R(p, q, \xi)|^2) < 1\}, \quad \text{with} \quad \mathbb{E}(|R(p, q, \xi)|^2) = (1 + p)^2 + |q|^2.$$

Let $\lambda, \mu \in \mathbb{R}$, replacing $p = \lambda\tau$, $q = \mu\tau^{1/2}$ into the stability condition $\mathbb{E}(|R(p, q, \xi)|^2) < 1$ yields

$$\tau < \frac{|2\lambda + \mu^2|}{\lambda^2} = \frac{2}{|\lambda|} - \frac{\mu^2}{\lambda^2}. \quad (1.48)$$

In (1.48) we observe two things. First, that $\tau < 2/|\lambda|$ and hence the stability condition for the EM scheme is stronger than for the explicit Euler scheme. In contrast, the SK-ROCK and RKC schemes share the same stability condition. Furthermore, if, for instance, (λ, μ) is close to the boundary of \mathcal{S}^{MS} , then $|2\lambda + \mu^2| \approx 0$ and τ must be taken extremely small, much smaller than $2/|\lambda|$. On the other hand, the SK-ROCK method is not affected by the distance between (λ, μ) and the border of \mathcal{S}^{MS} .

The reason for these additional downsides of the EM scheme, other than those inherited by explicit Euler, is that the stability domain \mathcal{S}_{num}^{EM} does not contain a portion \mathcal{S}_a^{MS} of \mathcal{S}^{MS} , with \mathcal{S}_a^{MS} as in (1.46). We display in Figure 1.4(b) the stability domain of the EM scheme and see that indeed \mathcal{S}_a^{MS} is not contained in \mathcal{S}_{num}^{EM} , for any $a > 0$.

1.4.2 The second kind τ -leap orthogonal Runge–Kutta–Chebyshev method

The modeling of kinetic chemical processes very often involves multiple chemical species reacting at disparate time-scales, therefore stiffness is a very common phenomenon for such problems. Standard explicit methods for such systems driven by discrete noise, as the τ -leaping method [61], face similar step size restrictions as the explicit Euler or the Euler–Maruyama schemes. Furthermore, the amplification properties of standard explicit methods prevent to capture the correct statistics of the process. Implicit or semi-implicit schemes also exists [32, 34, 99], however as the invariant measure of the process is not trivial (not Dirac) also implicit schemes fail in capturing the exact statistics due to their strong damping properties [34], see also [85] for a similar discussion in the white noise case. An exception is the trapezoidal rule, which for linear equations captures the exact statistics; however, it is shown in [85] that this property is lost when the scheme is applied to nonlinear equations. Hence, in [7] the authors propose a stabilized explicit scheme for discrete noise equations, called τ -ROCK, based on the S-ROCK [8] scheme for (1.38). For mean-square stable problems the τ -ROCK scheme represents an important improvement to the standard τ -leaping scheme. However, for non mean-square stable problems its effectiveness depends on a tuning of the damping parameter ε and for problems with too large variance the τ -ROCK scheme shows no improvement when compared to the τ -leaping scheme. We will introduce in this section an SK- τ -ROCK-like scheme for SDEs with discrete noise which inherits the stability properties of the SK-ROCK scheme and shows excellent capabilities of capturing the statistics of non mean-square stable problems thanks to a post-processing procedure already developed in [2].

In the remaining of this section we will briefly discuss the modeling of chemical systems, then we introduce the τ -leaping method and finally the SK- τ -ROCK method and its stability analysis. A numerical experiment comparing the accuracy and efficiency of the SK- τ -ROCK method, its post-processed version, the τ -ROCK method and the trapezoidal rule closes this section.

This section is based on [5, 55].

Exact simulation of chemical reaction systems

Here we give a short introduction to the modeling of a chemical reaction system and discuss the computational challenges of its exact simulation. We follow [70] and we refer the reader to the same article for a more detailed presentation and derivation of the different models.

Suppose that we are concerned with a chemical system composed by N species (of molecules) S_1, \dots, S_N which interact in M reactions, denoted R_1, \dots, R_M , and that we are interested in the number of molecules of each specie in an instant of time t . Hence, we denote by $X(t) = (X_1(t), \dots, X_N(t))^\top$ the state vector, where $X_j(t) \in \mathbb{N}$ is the number of molecules of specie S_j at time t .

The evolution of such system can be described by an initial condition describing the position and velocity of each particle and appropriate laws of physics describing collisions and interactions. However, this approach is obviously too expensive and a probabilistic approach is preferred. Therefore, we will neglect the position and velocities of the molecules and suppose that the system is well stirred, at thermal equilibrium and the volume of the physical domain is constant. Under these assumptions, each reaction R_j is characterized by

$$a_j : \mathbb{N}^N \rightarrow \mathbb{R}_{\geq 0}, \quad \boldsymbol{\nu}_j \in \mathbb{Z}^N,$$

where a_j is called propensity function and $\boldsymbol{\nu}_j$ is the state-change vector. Given a state X , the propensity function $a_j(X)$ characterizes the probability that reaction R_j occurs; as a reaction is

more likely to fire if the number of molecules of the involved species is high, then $a_j(X)$ contains products of the number of molecules of the species participating in the reaction. The state-change vector ν_j describes the change in $X(t)$ when reaction R_j is fired, i.e. reaction R_j has the effect of changing the state vector from $X(t)$ to $X(t) + \nu_j$. We will denote by $a(X) = (a_1(x), \dots, a_M(X))^\top$ the vector of propensity functions and by $\nu = (\nu_1, \dots, \nu_M)$ the stoichiometric matrix.

Example 1.2. We provide here an example in order to illustrate how to derive the propensity functions $a(X)$ and the stoichiometric matrix ν from the description of a chemical system. To do so, we consider the famous Michaelis–Menten system describing the mechanism of enzymatic catalysis. The model consists in four species: a substrate S_1 , an enzyme S_2 , a complex enzyme-substrate S_3 and the product S_4 . The reactions may be written as



The constants c_1, c_2, c_3 are called rate constants and define the speed, or frequency, at which the associated reaction occurs. The state vector is $X(t) = (X_1(t), X_2(t), X_3(t), X_4(t))^\top$ and represents the number of molecules of each species S_1, S_2, S_3, S_4 at time t . Let us see how to derive the stoichiometric matrix ν from (1.49). If the first reaction is fired, the value of $X_3(t)$ is increased by one molecule and $X_1(t), X_2(t)$ are decreased by one molecule each, hence the state vector is updated as $X(t) + \nu_1$, where $\nu_1 = (-1, -1, 1, 0)^\top$. In the same manner we define $\nu_2 = (1, 1, -1, 0)^\top$ and $\nu_3 = (0, 1, -1, 1)^\top$. The propensity function $a_1(X)$ describes the probability that the first reaction takes place and is proportional to the number of involved molecules, hence $a_1(X) = c_1 X_1 X_2$. The product $X_1 X_2$ represents the probability that S_1, S_2 collide and c_1 the fact that not all collisions result in a reaction. Similarly $a_2(X) = c_2 X_3$ and $a_3(X) = c_3 X_3$. There is also a third type of reaction, not present in (1.49), of the type $2S_n \xrightarrow{c_j} S_m$. The propensity function of such reaction is $a_j(X) = c_j X_n (X_n - 1)/2$ and its form comes from a combinatoric argument.

Given the propensity functions $a(X)$ and the stoichiometric matrix ν the system evolves following two simple rules.

- i) Given a state vector $X(t)$ at time t , in an infinitesimal time dt the probability that reaction R_j fires is given by $a_j(X(t)) dt$.
- ii) If dt is sufficiently small so that in the interval $[t, t + dt]$ only one reaction fires and this reaction is R_j , then the system is updated as $X(t + dt) = X(t) + \nu_j$.

Departing from these principles one can derive a system of ODEs called chemical master equation (CME). The CME describes the exact evolution, in time, of the probabilities of the system being in a certain state $x \in \mathbb{N}^N$. Hence, there is one ODE for each state x and the number of states depends on the number of different species and the total number of molecules in the system. Therefore, the CME very quickly becomes an enormous system of ODEs and sheer size forces the employment of some approximation.

Instead of computing the exact probability of being in a certain state x at time t , as the CME does, the stochastic simulation algorithm (SSA) [59, 60] computes samples $X(t)$ of the same distribution. The SSA closely follows the two principles i), ii) too and proceeds, informally, as follows. Given a state $X(t)$, it defines the probability distribution of the waiting time until the next reaction fires; then, it samples dt from this distribution. Next, it picks one of the reactions R_j , with the rule that the probability of sampling the j th reaction is proportional to $a_j(X(t)) dt$. Finally, it updates the state vector as $X(t + dt) = X(t) + \nu_j$, where R_j is the sampled reaction.

The waiting time until the next reaction takes place, dt , follows an exponentially distributed random variable with rate $a_0(X(t)) = \sum_{k=1}^M a_j(X(t))$, hence it can be extremely small if the number of molecules is large and therefore $a_0(t)$ is large. As a consequence, reactions in the SSA fire too frequently and the algorithm becomes practically impossible. Hence, a coarser-grained model is necessary.

The τ -leaping method

We recall here the τ -leaping method [61], to do so we follow [7].

The τ -leaping method speeds the simulation by fixing a time step τ and firing the reactions that would occur in the time interval $[t, t + \tau]$ simultaneously. One step, of size τ , of the τ -leaping method is given by

$$X_{n+1} = X_n + \sum_{j=1}^M \nu_j \mathcal{P}_j(a_j(X_n)\tau), \quad (1.50)$$

where the $\mathcal{P}_j(a_j(X_n)\tau)$ are random variables representing the number of times that reaction R_j fires in the interval $[t, t + \tau]$. Hence, $\mathcal{P}_j(a_j(X_n)\tau)$ is a counting process and follows a Poisson distribution with rate $a_j(X_n)\tau$. For the τ -leaping strategy to be accurate, the exact propensity functions $a(X(t))$ cannot change considerably in the interval $[t, t + \tau]$, as the propensity function is frozen at X_n . This assumption is satisfied when the number of molecules in the system is large and τ is sufficiently small.

The τ -leaping step (1.50) can also be written as

$$X_{n+1} = X_n + f(X_n)\tau + Q(X_n, \tau), \quad (1.51)$$

where

$$f(X) = \sum_{j=1}^M \nu_j a_j(X), \quad Q(X, \tau) = \sum_{j=1}^M \nu_j (\mathcal{P}_j(a_j(X)\tau) - a_j(X)\tau). \quad (1.52)$$

Since the mean (and variance) of $\mathcal{P}_j(a_j(X)\tau)$ is $a_j(X)\tau$, then $\mathbb{E}(Q(X, \tau)) = 0$. Thus, we note that (1.51) is very similar to the Euler–Maruyama scheme (1.47), except for the different noise. It is shown in [34] that the τ -leaping algorithm faces severe restrictions on the step size when the system is stiff.

The SK- τ -ROCK algorithm

The SK-ROCK scheme (1.40) can be seen as a stabilization procedure for the Euler–Maruyama method (1.47). In this section, we define the SK- τ -ROCK scheme as a stabilization procedure for the τ -leaping method (1.50) written in the Euler–Maruyama-like formulation (1.51). Then we define the postprocessing procedure, which improves the accuracy of the SK- τ -ROCK scheme for non mean-square stable problems. The content of this section and the remaining ones is taken from [5, 55].

One step, of size τ , of the SK- τ -ROCK scheme is given by

$$\begin{aligned} K_0 &= X_n, \\ K_1 &= K_0 + \mu_1 \tau f(K_0 + \nu_1 Q(K_0, \tau)) + \kappa_1 Q(K_0, \tau), \\ K_j &= \nu_j K_{j-1} + \kappa_j K_{j-2} + \mu_j \tau f(K_{j-1}), \quad j = 2, \dots, s, \\ X_{n+1} &= K_s, \end{aligned} \quad (1.53)$$

where f, Q are given in (1.52) and the coefficients μ_j, ν_j, κ_j , for $j = 1, \dots, s$, are the same as for the SK-ROCK scheme, hence given in (1.20) and (1.39).

In chemical reactions, one is often interested in the stationary state of a given system. Hence, capturing accurately the invariant measure of (1.50) is of considerable interest. In order to obtain second order accuracy for the invariant measure of (1.50) at a certain time $t_n = n\tau$ the postprocessor

$$\bar{X}_n = X_n + \frac{1}{2s}Q(X_n, \tau) \quad (1.54)$$

is employed. In what follows, we call PSK- τ -ROCK the postprocessed SK- τ -ROCK scheme defined by (1.53) and (1.54). Note that the PSK- τ -ROCK scheme does not need (1.54) to advance the solution in time but only to obtain higher accuracy at a specified time t_n .

Stability analysis on the reversible isomerization reaction model

We analyze here the stability and accuracy of the SK- τ -ROCK and PSK- τ -ROCK scheme on a test equation.

In the context of numerical integrators for discrete noise equations, the isomerization reaction model

$$S_1 \xrightleftharpoons[c_2]{c_1} S_2, \quad (1.55)$$

first introduced in [34], plays the role of test equation. This is a reversible system and therefore the number of molecules is constant, i.e. $X_1(t) + X_2(t) = X^T$. As a consequence specie S_2 can be neglected, we consider only specie S_1 and denote $X(t) = X_1(t)$. The system is described by the two reactions

$$a_1(x) = c_1x, \quad a_2(x) = c_2(X^T - x), \quad \nu_1 = -1, \quad \nu_2 = 1.$$

We also define $\lambda = -(c_1 + c_2)$, which represents the stiffness in the system. From the CME one can compute the exact distribution of (1.55) and find that the system has a stationary state X_∞ following a binomial distribution $B(n, p)$, with $n = X^T$ and $p = c_2/|\lambda|$. Hence,

$$\mathbb{E}(X_\infty) = \frac{c_2}{|\lambda|}X^T, \quad \text{Var}(X_\infty) = \frac{c_1c_2}{|\lambda|^2}X^T,$$

where for a random variable X we denote as $\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$ its variance. Note that if $c_2 = 0$, i.e. (1.55) is not reversible, then $\mathbb{E}(X_\infty) = \text{Var}(X_\infty) = 0$ and the problem is considered to be mean-square stable.

A numerical method applied to (1.55) is said to have absolutely stable mean and variance if, and only if, $\mathbb{E}(X_n)$ and $\text{Var}(X_n)$ remain bounded as $n \rightarrow \infty$. We state now a theorem showing stability and accuracy properties of the SK- τ -ROCK method for (1.55), for the proof see [5, 55]. We remind that ℓ_s^ε represents the size of the stability domain of the RKC scheme, see Section 1.2.1 and (1.8).

Theorem 1.3. *If $\tau|\lambda| \leq \ell_s^\varepsilon$ and $\varepsilon > 0$, the mean and variance of the SK- τ -ROCK scheme (1.53) are absolutely stable. Moreover, the mean and variance satisfy*

$$\lim_{n \rightarrow \infty} \mathbb{E}(X_n) = \mathbb{E}(X_\infty), \quad \lim_{n \rightarrow \infty} \text{Var}(X_n) = \frac{-2zB_s(z)^2}{1 - A_s(z)^2} \text{Var}(X_\infty),$$

where $z = \tau\lambda$ and A_s, B_s are as in (1.45).

We see in Theorem 1.3 that the SK- τ -ROCK has the same stability condition as the RKC and SK-ROCK schemes, i.e. $|\tau|\lambda| \leq \ell_s^\varepsilon$. The condition on the damping parameter, $\varepsilon > 0$, is only needed to ensure that $|A_s(z)| < 1$; as for some points along the negative real axis $|A_s(z)| = 1$ is possible. Furthermore, we observe that the SK- τ -ROCK scheme is able to capture the true expectation. In contrast, the variance is off. The same downside is seen when the SK-ROCK method is applied to the Ornstein–Uhlenbeck process; hence, in [2] a post-processing strategy which allows to capture the exact variance of the Ornstein–Uhlenbeck process is developed. Applying the same methodology to the SK- τ -ROCK method we defined the PSK- τ -ROCK scheme (1.53) and (1.54). The following theorem shows that the PSK- τ -ROCK scheme captures the exact invariant measure of (1.55), for the proof see [5, 55].

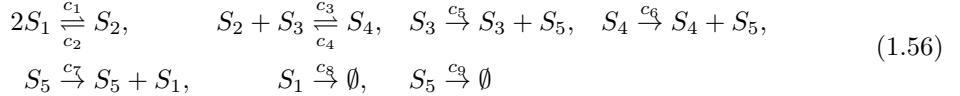
Theorem 1.4. *Let \bar{X}_n be the solution of the PSK- τ -ROCK scheme (1.53) and (1.54) with damping parameter $\varepsilon = 0$ applied to (1.55). If $|\tau|\lambda| \leq \ell_s^\varepsilon$ and $|A_s(z)| < 1$, with $z = \tau\lambda$, then*

$$\lim_{n \rightarrow \infty} \mathbb{E}(\bar{X}_n) = \mathbb{E}(X_\infty), \quad \lim_{n \rightarrow \infty} \text{Var}(\bar{X}_n) = \text{Var}(X_\infty).$$

However, the SK- τ -ROCK scheme with damping $\varepsilon = 0$ is not robust to use as condition $|A_s(z)| < 1$ can not be guaranteed. Therefore, in practice a small damping should be employed.

Numerical experiment on the genetic positive feedback loop

We close the chapter presenting a numerical example (taken from [5, 55]) where we compare the accuracy and computational efficiency of the SK- τ -ROCK, PSK- τ -ROCK, τ -ROCK [7] scheme and the trapezoidal rule [34]. To do so, consider a biological system of stiff chemical reactions called genetic positive feedback loop, taken from [128]. The system is described by the set of reactions



and the rate constants are given by $c_1 = 50$, $c_2 = 10^3$, $c_3 = 50$, $c_4 = 10^3$, $c_5 = 1$, $c_6 = 10$, $c_7 = 3$, $c_8 = 1$ and $c_9 = 6$. We set the final time to $T = 100$ and the initial condition as $X(0) = (10, 0, 20, 0, 0)^\top$.

We start by observing that the system contains the nonlinear reversible reaction



which induce significant fluctuations in X_1, X_2 when compared to other species. The same reaction has been considered in [7, Example 3] and it was found that at equilibrium the variance of X_1, X_2 is relatively large. Hence, for the stability of the τ -ROCK scheme a very small step size τ or a very high damping parameter ε combined with a very high number of stages were needed, see also [55, Section 6.3.3]. Therefore, for such reaction the τ -ROCK scheme is no improvement compared to the τ -leaping method. For (1.56) the same considerations hold and indeed we found that simulation of (1.56) with the τ -ROCK scheme is unstable. Therefore, we will discard the τ -ROCK scheme from the rest of the experiment.

To test the accuracy and efficiency of the SK- τ -ROCK, PSK- τ -ROCK and the trapezoidal rule we apply these methods to (1.56); using a fixed step size $\tau = 0.05$ and computing 10^4 samples. For the SK- τ -ROCK and PSK- τ -ROCK method we used a damping $\varepsilon = 0.05$ and $s = 23$ stages, which is the minimal number of stages which provides a stable integration, for such nonlinear problem.

	X_1	X_2	X_3	X_4	X_5
SSA (reference)	92.25	212.69	1.69	18.31	30.66
Trapezoidal	16.7	2438.4	1.6	18.4	31.0
SK- τ -ROCK	92.37	211.36	1.73	18.27	30.79
PSK- τ -ROCK	92.41	211.30	1.73	18.27	30.71

(a) *Empirical means.*

	X_1	X_2	X_3	X_4	X_5
SSA (reference)	9.94	18.83	1.25	1.25	5.55
Trapezoidal	313.7	9209.2	1.3	1.3	5.6
SK- τ -ROCK	7.14	18.66	0.74	0.74	5.50
PSK- τ -ROCK	9.53	18.73	1.17	1.17	5.55

(b) *Empirical standard deviations.*

Table 1.1. *Genetic positive feedback loop. Empirical means and standard deviations at time $T = 100$, computed over 10^4 samples and using a step size $\tau = 0.05$.*

	CPU [sec.]
Trapezoidal	45767
SK- τ -ROCK	5855
PSK- τ -ROCK	5931

Table 1.2. *Genetic positive feedback loop. Computational times taken by different methods.*

We display the means and the standard deviations of the three methods in Table 1.1, we report as well the result given by the SSA over the same number of samples. As the SSA samples from the exact distribution it is regarded as a reference solution. We observe in Table 1.1(a) that both the SK- τ -ROCK and the PSK- τ -ROCK method approximate very well the mean of all the species. In contrast, the trapezoidal rule completely fails in capturing the mean of S_1 and S_2 . We saw that these two species are involved in the nonlinear reversible reaction (1.57) and have a large variance at equilibrium. In [34] it is shown that the trapezoidal rule captures the exact invariant measure of (1.55), i.e. a linear problem. This example shows that it fails in capturing the statistics of nonlinear problems as (1.57). The same phenomenon is observed in [85] for the trapezoidal rule for (1.38). In Table 1.1(b) we observe that the trapezoidal rule is also completely off in estimating the variances of S_1 and S_2 . In contrast, the other species not involved in (1.57) are well simulated. In the same table we display the results for SK- τ -ROCK and PSK- τ -ROCK. We observe that the variance of SK- τ -ROCK is always smaller than the reference one, this result is in line with the fact that the amplification factor in Theorem 1.3 is smaller than one. The PSK- τ -ROCK instead is very precise in estimating the variance of all the species.

We end the section displaying in Table 1.2 the computational times taken by the three methods. We note that SK- τ -ROCK and PSK- τ -ROCK take a similar amount of time. In contrast, due to its implicit nature the trapezoidal rule is almost 8 times slower than the stabilized explicit schemes.

This numerical experiment shows the accuracy and performance of the SK- τ -ROCK and PSK- τ -ROCK methods. We see that these methods provide very accurate solutions, in contrast to the τ -ROCK scheme that needs either a high number of function evaluations or can not be employed due to stability issues for problems with large variance. We observe as well that the trapezoidal rule, which is exact for linear problems, is completely off for this nonlinear problem and about 8 times more expensive than the SK- τ -ROCK and PSK- τ -ROCK methods.

2 An interpolation based additive Runge–Kutta–Chebyshev method

In this chapter we introduce an additive RKC method for stiff ordinary differential equations (ODEs)

$$y' = f(y), \quad y(0) = y_0, \quad (2.1)$$

where $y(t) \in \mathbb{R}^n$ with $n \geq 2$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a smooth function and $y_0 \in \mathbb{R}^n$ is the initial value. We suppose that f can be split in a severely stiff and a mildly stiff term in the following sense: there is a diagonal matrix $D \in \mathbb{R}^{n \times n}$ such that $D_{ii} = 0$ or $D_{ii} = 1$ for $i = 1, \dots, n$ and Df , $(I - D)f$ are a severely stiff and a mildly stiff term, respectively.¹ A typical example are parabolic problems spatially discretized on a locally refined grid; indeed, the Jacobian's eigenvalues depend on the mesh size. Coarse and refined regions lead to mildly and severely stiff terms, respectively, hence setting D as $D_{ii} = 1$ if, and only if, the i th degree of freedom is in the refined region leads to a severely stiff term Df and a mildly stiff term $(I - D)f$.

Since there is a stiff term Df then integration of (2.1) becomes expensive, even if stiffness is induced by a few components only. Multirate methods exploit the special structure of the problem in order to reduce the computational cost. This is often achieved by adapting the Runge–Kutta (RK) method or the step size to the specific partition of the system and employing interpolations or extrapolations for coupling the components together. It is known that the coupling strategy between the stiff and nonstiff terms strongly affects the stability of the system; indeed, a major difficulty in the field is to construct stable multirate methods (see for instance [58, 64, 65, 73, 81]).

The goal of this chapter is to discuss the properties of the interpolation based additive RKC scheme defined in Section 2.1, which turns out to be very similar to the method described in [95]. In Section 2.2 we will first show that the linear interpolations employed in the scheme might render the integration process unstable and then we discuss theoretically an order reduction phenomenon observed in numerical experiments. Numerical experiments are provided in Section 2.3.

2.1 The additive Runge–Kutta–Chebyshev method

We present here an additive method which uses two RKC schemes simultaneously. Depending on the choice of coefficients the method can be first- or second-order accurate. The scheme preserves the explicitness of the RKC schemes and does not need any predictor step, but makes use of linear interpolations in time between stages.

Before introducing the additive RKC scheme in Section 2.1.2, in Section 2.1.1 we split (2.1) into a

¹We are aware that “severely stiff” and “mildly stiff” are qualitative somewhat imprecise characterizations. This is meant to indicate that the fastest dynamics are in the severely stiff terms. Since the slower scales can still be fast enough to prevent the use of classical explicit schemes, we call them mildly stiff.

stiff and a mildly stiff problem, which are then integrated independently using two RKC schemes.

2.1.1 Equation splitting

Let $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix such that $D_{ii} = 1$ or $D_{ii} = 0$ and $E = I - D$, where $I \in \mathbb{R}^{n \times n}$ is the identity matrix. Then (2.1) can be written as

$$y' = Df(Dy + Ey) + Ef(Dy + Ey), \quad y(0) = y_0 \quad (2.2)$$

and multiplying (2.2) either by D or E yields

$$Dy' = Df(Dy + Ey), \quad Dy(0) = Dy_0, \quad (2.3a)$$

$$Ey' = Ef(Dy + Ey), \quad Ey(0) = Ey_0, \quad (2.3b)$$

as $DE = ED = 0$, $D^2 = D$ and $E^2 = E$. Letting $y_F = Dy$, $y_S = Ey$, $f_F = Df$ and $f_S = Ef$ we can rewrite (2.3)

$$y'_F = f_F(y_F + y_S), \quad y_F(0) = Dy_0, \quad (2.4a)$$

$$y'_S = f_S(y_F + y_S), \quad y_S(0) = Ey_0. \quad (2.4b)$$

Usually, the matrix D is chosen such that f_F is stiff (F for fast) and f_S is less stiff compared to f_F (S for slow).

2.1.2 The additive RKC algorithm

The additive RKC (aRKC) scheme integrates the two problems in (2.4) separately, applying an RKC method of first- or second-order to each equation. Integration is performed simultaneously and linear interpolation is employed for the equations' coupling. Here, we introduce the aRKC method using the second-order RKC schemes of Section 1.3.2, the aRKC method with first-order RKC schemes of Section 1.2.4 has the same formulation. The only difference is on the stability condition and the choice of coefficients. We will comment again on this at the end of the section.

Given ρ_F, ρ_S the spectral radii of the Jacobians of f_F, f_S , respectively, choose s and m such that

$$\tau\rho_F \leq \bar{\beta}(m^2 - 1) \quad \text{and} \quad \tau\rho_S \leq \bar{\beta}(s^2 - 1), \quad (2.5)$$

where conditions (2.5) derive from the stability conditions (1.32) of the second-order RKC scheme. The aRKC scheme integrates (2.4a) using m stages and (2.4b) using s stages. In the following we call $\mu_i, \nu_i, \kappa_i, a_i$ and c_i the coefficients of an s -stage second-order RKC method as defined in Section 1.3.2. The coefficients of the m -stage second-order RKC method are noted $\alpha_j, \beta_j, \gamma_j, \bar{a}_j$ and d_j . In order to alleviate the notation we motivate the aRKC method considering the first step only; hence, in what follows k_i is an approximation to $y_S(c_i\tau)$ and l_j an approximation to $y_F(d_j\tau)$.

If $y_F(t)$ was known, we could integrate (2.4b) with the scheme

$$\begin{aligned} k_0 &= y_S(0), \\ k_1 &= k_0 + \mu_1\tau f_S(y_F(0) + k_0), \\ k_i &= \nu_i k_{i-1} + \kappa_i k_{i-2} + (1 - \nu_i - \kappa_i)k_0 \\ &\quad + \mu_i\tau f_S(y_F(c_{i-1}\tau) + k_{i-1}) - \mu_i a_{i-1}\tau f_S(y_F(0) + k_0), \quad i = 2, \dots, s. \end{aligned} \quad (2.6a)$$

Alternatively, if $y_S(t)$ was known, we could integrate (2.4a) with the scheme

$$\begin{aligned} l_0 &= y_F(0), \\ l_1 &= l_0 + \alpha_1 \tau f_F(l_0 + y_S(0)), \\ l_j &= \beta_j l_{j-1} + \gamma_j l_{j-2} + (1 - \beta_j - \gamma_j) l_0 \\ &\quad + \alpha_j \tau f_F(l_{j-1} + y_S(d_{j-1} \tau)) - \alpha_j \bar{a}_{j-1} \tau f_F(l_0 + y_S(0)), \quad j = 2, \dots, m. \end{aligned} \quad (2.6b)$$

However, as neither y_F nor y_S are known they must be approximated. Since, for $d_{j-1} < c_i \leq d_j$,

$$y_F(c_i \tau) = y_F(d_{j-1} \tau) + \frac{c_i - d_{j-1}}{d_j - d_{j-1}} (y_F(d_j \tau) - y_F(d_{j-1} \tau)) + \mathcal{O}(|(d_j - d_{j-1})\tau|^2)$$

and $l_j = y_F(d_j \tau) + \mathcal{O}(|d_j \tau|^2)$, then we approximate $y_F(c_i \tau)$ by \tilde{l}_i defined by

$$\tilde{l}_i = l_{j-1} + \frac{c_i - d_{j-1}}{d_j - d_{j-1}} (l_j - l_{j-1}), \quad \text{where } d_{j-1} < c_i \leq d_j. \quad (2.7a)$$

A similar strategy is used for $y_S(d_j \tau)$, we approximate it by

$$\tilde{k}_j = k_{i-1} + \frac{d_j - c_{i-1}}{c_i - c_{i-1}} (k_i - k_{i-1}), \quad \text{where } c_{i-1} < d_j \leq c_i. \quad (2.7b)$$

Replacing in (2.6) the exact values $y_F(c_i \tau)$, $y_S(d_j \tau)$ by the approximations \tilde{l}_i , \tilde{k}_j , respectively, yields a fully discrete scheme. One step of the aRKC method is given by

$$\begin{aligned} k_0 &= E y_n, \\ k_1 &= k_0 + \mu_1 \tau f_S(l_0 + k_0), \\ k_i &= \nu_i k_{i-1} + \kappa_i k_{i-2} + (1 - \nu_i - \kappa_i) k_0 \\ &\quad + \mu_i \tau f_S(\tilde{l}_{i-1} + k_{i-1}) - \mu_i a_{i-1} \tau f_S(l_0 + k_0), \quad i = 2, \dots, s \end{aligned} \quad (2.8a)$$

and

$$\begin{aligned} l_0 &= D y_n, \\ l_1 &= l_0 + \alpha_1 \tau f_F(l_0 + k_0), \\ l_j &= \beta_j l_{j-1} + \gamma_j l_{j-2} + (1 - \beta_j - \gamma_j) l_0 \\ &\quad + \alpha_j \tau f_F(l_{j-1} + \tilde{k}_{j-1}) - \alpha_j \bar{a}_{j-1} \tau f_F(l_0 + k_0), \quad j = 2, \dots, m, \end{aligned} \quad (2.8b)$$

where \tilde{k}_j, \tilde{l}_i are defined in (2.7) and $y_{n+1} = k_s + l_m$ is an approximation to $y(t_{n+1})$.

Observe that the conditions on c_i, d_j in interpolations (2.7) impose an interlaced evaluation order for the stages k_i, l_j in (2.8). For instance, the algorithm can compute both k_1, l_1 as k_0 and l_0 are known. But then it can compute k_2 only if $c_1 \leq d_1$. Indeed, the computation of k_2 requires \tilde{l}_1 and the latter needs l_j , where j is such that $c_1 \leq d_j$. Since only l_1 has been computed, the scheme can compute k_2 only if $c_1 \leq d_1$. Otherwise it computes l_2 , which can be computed if $d_1 \leq c_1$. Hence, at each iteration the algorithm verifies which condition (2.7a) or (2.7b) on c_i, d_j is satisfied and computes k_i or l_j accordingly. An illustrative example is provided in Figure 2.1.

The actual implementation of the scheme is fairly simple and a pseudo-code is given in Algorithm 1 below.

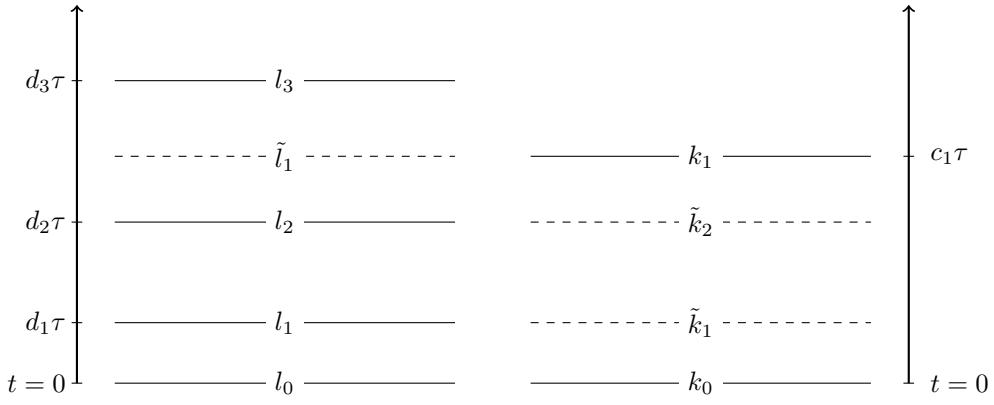


Figure 2.1. Illustration of the aRKC algorithm, solid lines represent the stages k_i, l_j while dashed lines represent the interpolations \tilde{k}_j, \tilde{l}_i . In this example, the algorithm proceeds as follows. As k_0 and l_0 are known the scheme can compute k_1 and l_1 , which are approximations at times $c_1\tau$ and $d_1\tau$, respectively, with $d_1 < c_1$. Then, it cannot compute k_2 , as it would need \tilde{l}_1 , which is an interpolation of l_2 and l_3 , that are not yet computed. But, it can compute l_2 for which \tilde{k}_1 , an interpolation between k_1 and k_0 , can be computed. Once l_2 is computed, k_2 can still not be computed as l_3 is missing. But it can compute l_3 , which requires an interpolation \tilde{k}_2 of k_1 and k_0 (note that \tilde{k}_1, \tilde{k}_2 are both interpolations of k_0 and k_1 , but at different times $d_1\tau, d_2\tau$, respectively). Once l_3 is known \tilde{l}_1 can be computed and k_2 can be evaluated. Informally, we observe that the rule is to advance the variable which is behind in time.

Algorithm 1 aRKC

```

Set  $s, m$  the smallest integers satisfying  $\tau\rho_F \leq \bar{\beta}(m^2 - 1)$  and  $\tau\rho_S \leq \bar{\beta}(s^2 - 1)$ .
 $k_0 = E y_n$ 
 $l_0 = D y_n$ 
 $k_1 = k_0 + \mu_1 \tau f_S(l_0 + k_0)$ 
 $l_1 = l_0 + \alpha_1 \tau f_F(l_0 + k_0)$ 
 $i = j = 1$ 
while  $i < s$  or  $j < m$  do
    if  $d_j < c_i$  then
         $\tilde{k}_j = k_{i-1} + \frac{d_j - c_{i-1}}{c_i - c_{i-1}}(k_i - k_{i-1})$ 
         $j = j + 1$ 
         $l_j = \beta_j l_{j-1} + \gamma_j l_{j-2} + (1 - \beta_j - \gamma_j)l_0$ 
         $+ \alpha_j \tau f_F(l_{j-1} + \tilde{k}_{j-1}) - \alpha_j \bar{a}_{j-1} \tau f_F(l_0 + k_0)$ 
    else if  $c_i \leq d_j$  then
         $\tilde{l}_i = l_{j-1} + \frac{c_i - d_{j-1}}{d_j - d_{j-1}}(l_j - l_{j-1})$ 
         $i = i + 1$ 
         $k_i = \nu_i k_{i-1} + \kappa_i k_{i-2} + (1 - \nu_i - \kappa_i)k_0$ 
         $+ \mu_i \tau f_S(\tilde{l}_{i-1} + k_{i-1}) - \mu_i a_{i-1} \tau f_S(l_0 + k_0)$ 
    end if
end while
 $y_{n+1} = k_s + l_m$ 

```

The first-order aRKC scheme is formulated very similarly. First, the stability condition (2.5) is replaced by $\tau\rho_F \leq \beta m^2$ and $\tau\rho_S \leq \beta s^2$, which follow from the stability condition (1.10) of the first-order RKC schemes. Then, the coefficients μ_i, ν_i, κ_i and c_i of the s -stage scheme are chosen as in Section 1.2.4. Similarly, we set $\alpha_j, \beta_j, \gamma_j$ and d_j as the coefficients of an m -stage first-order RKC scheme. Noting that for first-order RKC schemes $1 - \nu_i - \kappa_i = 1 - \beta_j - \gamma_j = 0$ and setting $a_i = \bar{a}_j = 0$, the first-order aRKC can be written as in (2.7) and (2.8), or Algorithm 1.

2.2 Instabilities and order reduction

In this section we will discuss how linear interpolation adversely affects the stability and accuracy of the aRKC method.

2.2.1 Instability on a model problem

Now, we study the stability properties of the aRKC scheme when applied to a 2×2 system. We consider the equation

$$y' = Ay, \quad y(0) = y_0, \quad (2.9)$$

where $y_0 \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2 \times 2}$ is a symmetric matrix defined by

$$A = \begin{pmatrix} \zeta & \sigma \\ \sigma & \lambda \end{pmatrix}, \quad (2.10)$$

with $\lambda, \zeta \leq 0$ and $\sigma^2 \leq \lambda\zeta$. Under these conditions A is nonpositive definite. We will study the stability of the aRKC scheme when applied to (2.9). Let

$$D = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

and $E = I - D$. In this setting it holds $f_F(y) = DAy$ and $f_S(y) = EAy$ with $\rho_F = |\lambda|$ and $\rho_S = |\zeta|$. Since the system is linear, applying Algorithm 1 yields

$$y_1 = R_{s,m}(\tau DA, \tau EA)y_0,$$

where $R_{s,m}(\tau DA, \tau EA)$ is the iteration matrix. If the spectral radius of $R_{s,m}(\tau DA, \tau EA)$ is bounded by one, the aRKC scheme is stable.

Let us fix $s, m \in \mathbb{N}$, if $\sigma = 0$ then the two equations defined by (2.9) are independent and the scheme is stable for all $\tau\lambda$ and $\tau\zeta$ inside the stability domain of the two RKC schemes, i.e. $(\tau\lambda, \tau\zeta) \in [-\beta m^2, 0] \times [-\beta s^2, 0]$ for the first-order aRKC method and $(\tau\lambda, \tau\zeta) \in [-\bar{\beta}(m^2 - 1), 0] \times [-\bar{\beta}(s^2 - 1), 0]$ for the second-order aRKC method. We want to investigate the stability of the scheme when $\sigma \neq 0$, hence with coupling. Let $z = \tau\lambda$, $w = \tau\zeta$, $u = \tau\sigma$, then

$$B := \tau A = \begin{pmatrix} w & u \\ u & z \end{pmatrix} \quad \text{and} \quad R_{s,m}(\tau DA, \tau EA) = R_{s,m}(DB, EB).$$

Since $\sigma^2 \leq \lambda\zeta$ then $u^2 \leq zw$ and in the following we consider $u = \theta\sqrt{zw}$ with $\theta \in [-1, 1]$. Thus, we define

$$B_\theta = \begin{pmatrix} w & \theta\sqrt{zw} \\ \theta\sqrt{zw} & z \end{pmatrix}$$

and denote the stability domain of the additive RKC method by

$$\mathcal{S} = \{(z, w) \in \mathbb{R}^2 : \rho(R_{s,m}(DB_\theta, EB_\theta)) \leq 1\},$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix and \mathcal{S} depends implicitly on the coupling strength θ . We will study the stability of the aRKC methods for different coupling strengths θ , from $\theta = 0$ corresponding to the absence of coupling to $\theta = \pm 1$, the maximal coupling. First, we consider the first-order aRKC method and let (z, w) vary in the rectangle $[-\beta m^2, 0] \times [-\beta s^2, 0]$, which is the stability domain of the method when there is no coupling, i.e. $\theta = 0$. The method is considered to be stable if, and only if, $[-\beta m^2, 0] \times [-\beta s^2, 0] \subset \mathcal{S}$ for all coupling strength $\theta \in [-1, 1]$.

Observe that the matrix $R_{s,m}(DB_\theta, EB_\theta)$ can be computed replacing y_0 by I , $f_F(y)$ by $DB_\theta y$ and $f_S(y)$ by $EB_\theta y$ in Algorithm 1. Hence, for some fixed s , m and θ values we display in Figures 2.2 and 2.3 the stability domain \mathcal{S} of the first-order aRKC methods by computing the spectral radius of $R_{s,m}(DB_\theta, EB_\theta)$ for varying $(z, w) \in [-\beta m^2, 0] \times [-\beta s^2, 0]$. The shaded regions represent the stability domains, while the dashed black lines represent the box $B_{s,m} = [-\beta m^2, 0] \times [-\beta s^2, 0]$, which is the region where the method is stable in absence of coupling, i.e. for $\theta = 0$. In Figure 2.2 we show the results for the first-order aRKC method with $m = 8$ and $s = 4$. We observe in Figure 2.2(a) that for $\theta = 0$ and a standard damping parameter $\varepsilon = 0.05$, the method is stable in the box $B_{s,m}$, as expected. In Figures 2.2(b) and 2.2(c) we increase the coupling factor θ and observe that instability regions appear inside the box $B_{s,m}$. In Figure 2.2(d) we try to increase the damping parameter $\varepsilon = 0.2$ and notice that it is not enough to fully stabilize the method. We observed that taking an even larger damping parameter does not stabilize the method. We perform the same experiment in Figure 2.3 but taking $s = 40$ and $m = 10$, we see again that if $\theta > 0$ the method has instability regions inside the box $B_{s,m}$ and increasing the damping parameter ε does not help in stabilizing the scheme. Moreover, comparing Figures 2.2 and 2.3 we remark that the pattern of the instability region is very different and hence not predictable. We perform the same experiment in Figure 2.4 using the second-order aRKC method and obtain similar results.

Figures 2.2 to 2.4 illustrate that the additive RKC method discussed here is not stable. Furthermore, the location of the instability regions is not easy to characterize, thus changing the values of s and m does not help in stabilizing the scheme in a given region.

2.2.2 Order reduction in the second-order additive RKC scheme

Here we show how linear interpolation triggers an unexpected order reduction phenomenon in the second-order aRKC method.

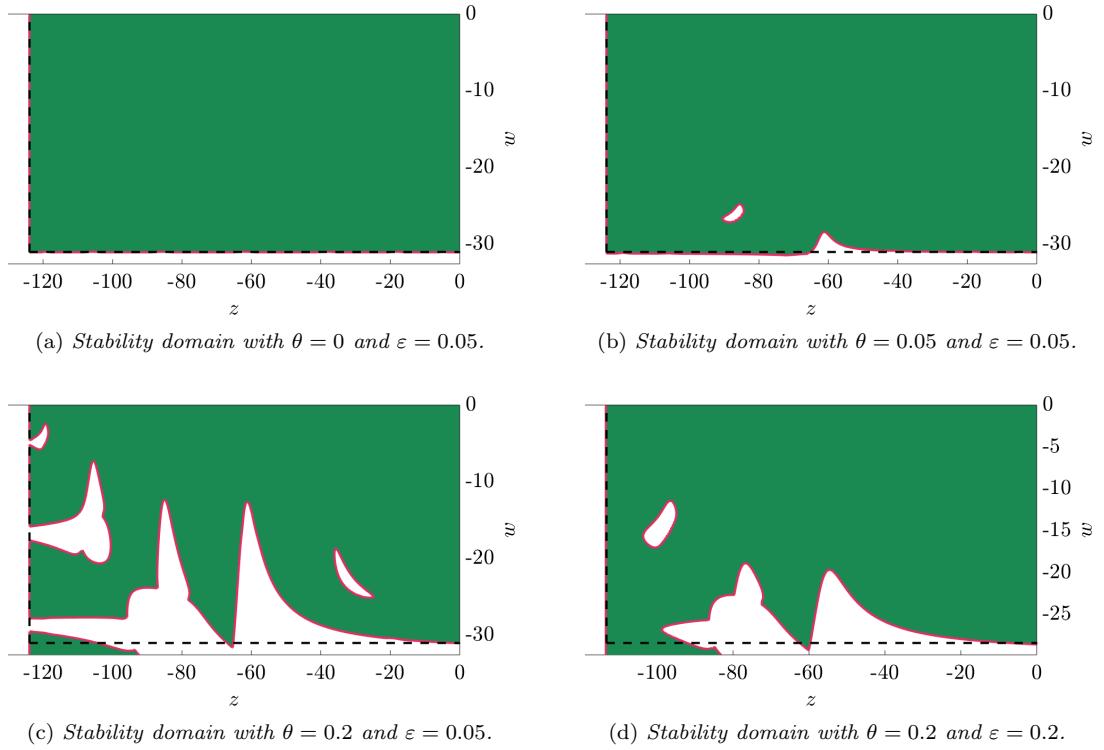
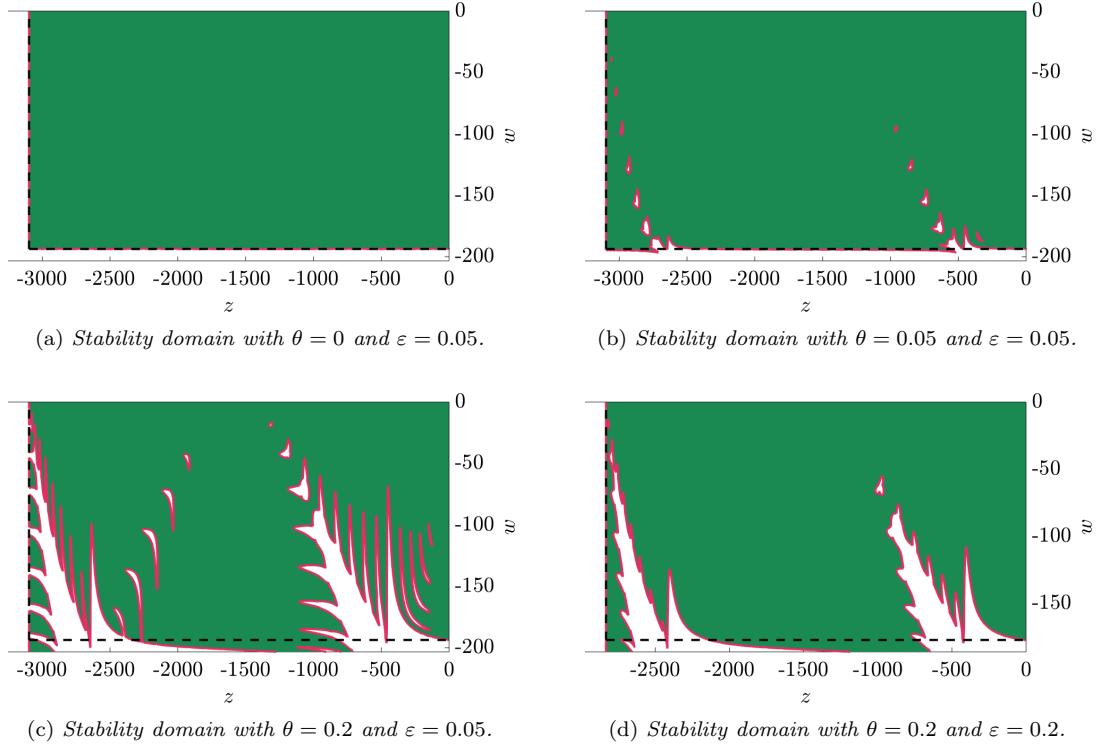
For simplicity, we will motivate the order reduction phenomenon using a semidiscrete method, where the stage values l_j are known beforehand and are exact, that is $l_j = y_F(d_j \tau)$. Furthermore, we assume that f_S is integrated exactly by the second-order RKC scheme. In this situation, Algorithm 1 reduces to computing the exact solution $\tilde{y}_S(\tau)$ of

$$\tilde{y}'_S = f_S(\tilde{y}_F + \tilde{y}_S) \quad t \in (0, \tau], \quad y_S(0) = E y_0,$$

where $\tilde{y}_F(t) : [0, \tau] \rightarrow \mathbb{R}^n$ is the piecewise interpolation of $y_F(d_j \tau)$ for $j = 0, \dots, m$. This scheme is clearly more accurate than the general aRKC scheme (i.e. when l_j are not known and f_S is not integrated exactly) and an order reduction for this semidiscrete method will imply the order reduction for the full second-order aRKC method.

Let us define $E_F(t) = y_F(t) - \tilde{y}_F(t)$ and $E_S(t) = y_S(t) - \tilde{y}_S(t)$ for $t \in [0, \tau]$. For $t \in [d_{j-1}\tau, d_j\tau]$, from a standard linear interpolation result we get

$$|E_F(t)| \leq C_y((d_j - d_{j-1})\tau)^2,$$

Figure 2.2. Stability domains of the first-order aRKC method for $m = 8$, $s = 4$.Figure 2.3. Stability domains of the first-order aRKC method for $m = 40$, $s = 10$.

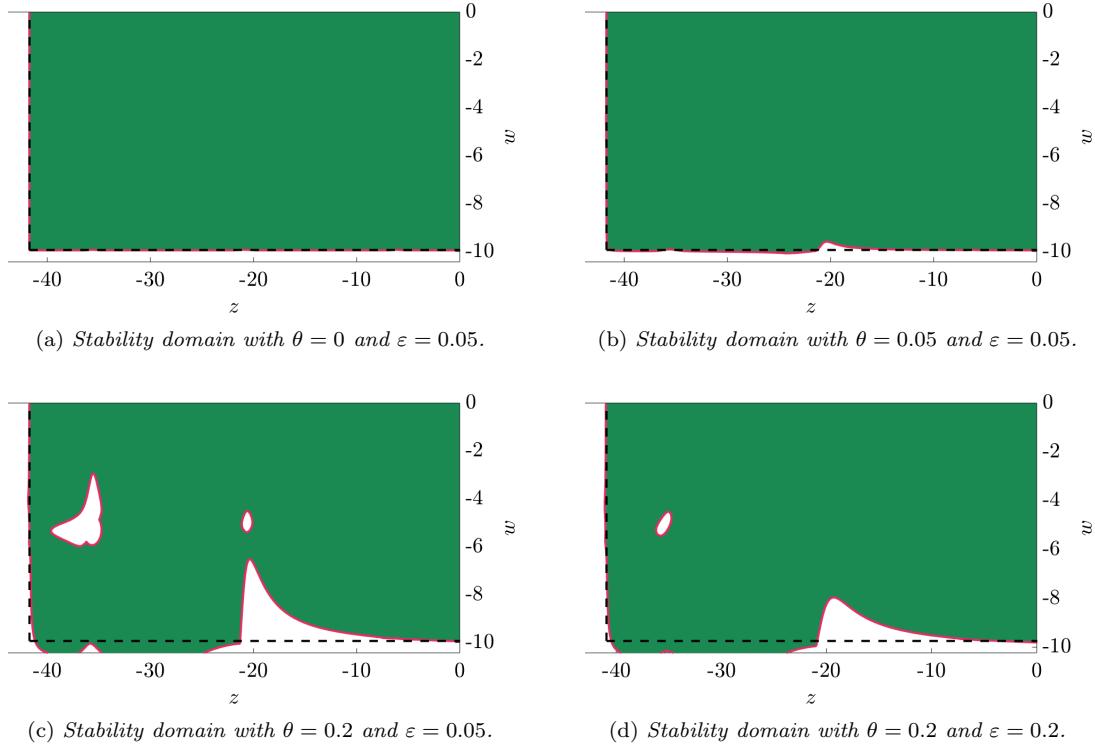


Figure 2.4. Stability domains of the second-order aRKC method for $m = 8$, $s = 4$.

where C_y is a constant dependent on $\max_{t \in [0, \tau]} |y_F''(t)|$. For $t \in [0, \tau]$ it holds

$$\begin{aligned} E_S(t) &= \int_0^t f_S(y_F(s) + y_S(s)) - f_S(\tilde{y}_F(s) + \tilde{y}_S(s)) \, ds \\ &= \int_0^t \int_0^1 \frac{\partial f_S}{\partial y}(\bar{y}(r, s))(E_F(s) + E_S(s)) \, dr \, ds, \end{aligned}$$

with $\bar{y}(r, s)$ in the segment $[y_F(s) + y_S(s), \tilde{y}_F(s) + \tilde{y}_S(s)]$. Supposing $\|\frac{\partial f_S}{\partial y}\| \leq M_S$ we get

$$|E_S(t)| \leq M_S \tau \max_{s \in [0, \tau]} |E_F(s)| + M_S \int_0^t |E_S(s)| \, ds$$

and using Gronwall's lemma we obtain

$$\begin{aligned} |E_S(t)| &\leq M_S \tau e^{\tau M_S} \max_{s \in [0, \tau]} |E_F(s)| \\ &\leq C_y M_S e^{\tau M_S} \max_{j=1, \dots, m} |d_j - d_{j-1}|^2 \tau^3 \\ &= C_S(\tau) \max_{j=1, \dots, m} |d_j - d_{j-1}|^2 \tau^3, \end{aligned} \tag{2.11}$$

where $C_S(\tau) = C_y M_S e^{\tau M_S}$ is bounded from below by $C_y M_S$. Let us now estimate the quantity $\max_{j=1, \dots, m} |d_j - d_{j-1}|^2$ in the nonstiff and the stiff regime.

In a nonstiff regime $\tau \rho_F$ is small, where we recall that ρ_F is the spectral radius of the Jacobian of f_F . Since the stability condition of the second-order RKC method is $\tau \rho_F \leq \bar{\beta}(m^2 - 1)$ then m

is a small number in this regime. It follows that the discretization of the interval $[0, 1]$ by the nodes $\{d_j\}_{j=1}^m$ is coarse and the estimate

$$\max_{j=1,\dots,m} |d_j - d_{j-1}|^2 \leq 1$$

is fairly accurate, implying that

$$|E_S(t)| \leq C_S(\tau)\tau^3 \quad (2.12)$$

is tight. Therefore, in a nonstiff regime the interpolation error introduces a third-order local error in the numerical solution, without deteriorating the global second-order accuracy of the aRKC scheme.

In contrast in a stiff regime $\tau\rho_F$ is large and therefore m is large as well. For a damping parameter $\varepsilon = 0$ we have $d_j = (j^2 - 1)/(m^2 - 1)$ (see [125]) and thus

$$\max_{j=1,\dots,m} |d_j - d_{j-1}| = (d_m - d_{m-1}) = \frac{m^2 - 1 - (m-1)^2 + 1}{m^2 - 1} = \frac{2m - 1}{m^2 - 1} \approx \frac{2}{m}, \quad (2.13)$$

where the last approximation comes from the fact that m is large. Using $\tau\rho_F \leq \bar{\beta}(m^2 - 1) \leq \bar{\beta}m^2$ and (2.13) yield

$$\max_{j=1,\dots,m} |d_j - d_{j-1}|^2 \approx \frac{4}{m^2} \leq \frac{4\bar{\beta}}{\tau\rho_F}. \quad (2.14)$$

Hence, from (2.11) and (2.14) we obtain, approximately,

$$|E_S(t)| \leq C_S(\tau)\frac{4\bar{\beta}}{\rho_F}\tau^2. \quad (2.15)$$

Let us now discuss both interpolation error (2.12) and (2.15). We observe that in the stiff regime $\tau\rho_F \gg 1$, estimate (2.15) is a second-order interpolation error. In the nonstiff regime $\tau\rho_F = \mathcal{O}(1)$, estimate (2.12) is accurate and represents a third-order interpolation error. If we now fix ρ_F and vary τ (as done in Figure 2.5(b)) the transition from stiff to nonstiff regime occurs when the step size τ is sufficiently small so that $\tau\rho_F = \mathcal{O}(1)$. This is what is seen in Figure 2.5(b), where a second-order local interpolation error yields a first-order convergence of the aRKC, while for sufficiently small τ (nonstiff regime) a second-order convergence is recovered thanks to the third-order local interpolation error.

We note that second-order interpolation techniques for the stage values could lead to a genuine third-order interpolation error (also in the stiff regime). However, we observed that second-order interpolations techniques completely destroy the stability of the scheme and we are not aware of a strategy to avoid such instabilities.

2.3 Numerical experiments

In this section we present two numerical experiments that support the results of Sections 2.2.1 and 2.2.2.

2.3.1 Instability on the model problem

We show numerically that the first-order aRKC scheme is unstable (a similar example can be derived for the second-order aRKC scheme as well).

We consider (2.9) with $y_0 = (1, 1)^\top$ and A as in (2.10). We want to choose $s, m, \tau, \lambda, \zeta, \sigma$ such that s, m are the smallest integers satisfying $\tau|\lambda| \leq \beta m^2$, $\tau|\zeta| \leq \beta s^2$ but $\rho(R_{s,m}(\tau PA, \tau QA)) > 1$.

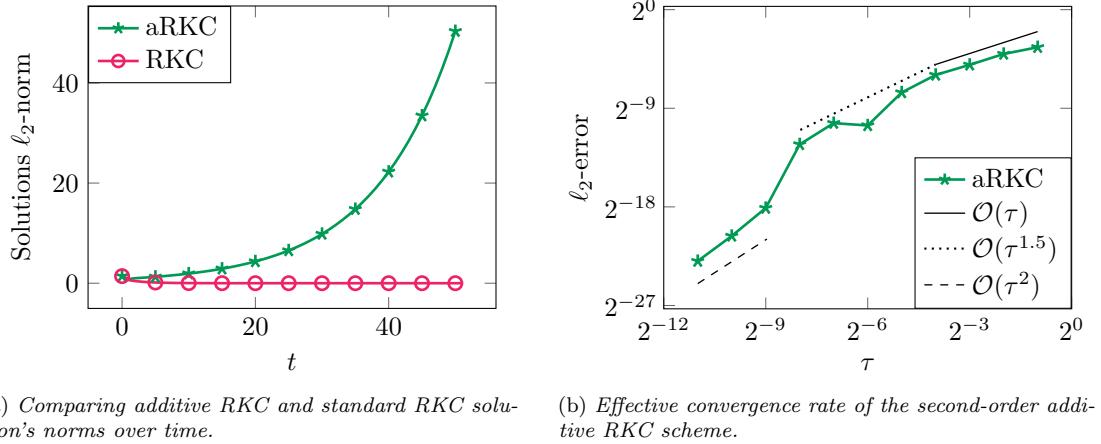


Figure 2.5. Illustrating instabilities and order reduction phenomenon of the additive RKC method.

Looking at Figure 2.2(c) we see that the couple $(z, w) = (-100, -28)$ is outside of the stability domain and $m = 8$, $s = 4$ are the smallest integers satisfying $|z| \leq \beta m^2$ and $|w| \leq \beta s^2$ (recall that $\beta \approx 2$). Hence, if we set $\lambda = -100$, $\zeta = -28$, $\sigma = 0.2\sqrt{\lambda}\zeta$, $\tau = 1$ and integrate (2.9) with the aRKC scheme then it will set $m = 8$ and $s = 4$. Since $\rho(R_{s,m}(\tau PA, \tau QA)) > 1$ we expect that the solution explodes. We display in Figure 2.5(a) the norm of the solutions given by the first-order aRKC and the first-order RKC method, indeed we observe that the aRKC method is unstable.

2.3.2 Order reduction on the heat equation

We consider the heat equation

$$\begin{aligned} \partial_t u - \Delta u &= g & (x, t) \in [0, e] \times [0, 1], \\ u(0, t) &= u(e, t) = 0 & t \in [0, 1], \\ u(x, 0) &= u_0(x) & x \in [0, e], \end{aligned} \quad (2.16)$$

with g , u_0 such that the exact solution is $u(x, t) = e^{-t}x(\log(x) - 1)$. We discretize the domain $\Omega = [0, e]$ with second-order finite differences. Since u has a spatial singularity in $x = 0$ we use a uniform mesh size $H \approx 1/2^4$ in $\Omega_S = (0.005e, e)$ and a uniform mesh size $h \approx H/200$ in $\Omega_F = (0, 0.005e)$. After space discretization, (2.16) can be written as

$$y' = Ay + G \quad t \in (0, 1], \quad y(0) = y_0. \quad (2.17)$$

Let D be a diagonal matrix of the same size as A such that $D_{ii} = 1$ if the i th node is in $\bar{\Omega}_F$ and $D_{ii} = 0$ else. We define $f_F(t, y) = D(Ay + G(t))$ and $f_S(t, y) = (I - D)(Ay + G(t))$. We verify the effective order of convergence of the second-order aRKC scheme integrating (2.17) using different step sizes $\tau = 1/2^k$, with $k = 1, \dots, 11$, comparing the numerical solution against a reference solution computed on the same mesh. We do not use the exact solution since we are only interested in time discretization errors. The results are shown in Figure 2.5(b), we observe that for large enough τ the rate of convergence is one, then there is a transition phase and finally for small τ the second-order convergence rate is recovered. This result is in line with the findings of Section 2.2.2.

3 Stabilized explicit multirate methods for stiff differential equations

In this chapter we introduce stabilized explicit multirate methods for stiff ordinary differential equations (ODE) as

$$y' = f(y) := f_F(y) + f_S(y), \quad y(0) = y_0, \quad (3.1)$$

where $f_F, f_S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and the eigenvalues of the Jacobians of f_F, f_S are distributed in a narrow strip along the negative real axis. Furthermore, we suppose that f_S is an expensive but only mildly stiff term (“S” for slow) and f_F is inexpensive but severely stiff (“F” for fast). Hence, the spectral radius ρ_S of the Jacobian of f_S is relatively small compared to the spectral radius ρ_F of the Jacobian of f_F ; still, we do not assume that f_F represents fast dynamics only and its Jacobian can have small eigenvalues too, in magnitude, as depicted in Figure 3.1(a). Therefore, we are not making any scale separation assumption, this is very important for applications where a clear separation of fast and slow dynamics is not available, as for spatially discretized parabolic problems.

We saw in Section 1.2.1 that the number of function evaluations needed by a stabilized explicit method grows as $\sqrt{\rho}$, where ρ is the spectral radius of f (see (1.11)). Hence, the cost of integrating (3.1) grows as $\sqrt{\rho} c_f$, with c_f the cost of evaluating f . Since in (3.1) the term inducing stiffness is f_F we can suppose $\rho \approx \rho_F$ and since the expensive term is f_S we can suppose $c_f \approx c_S$, where c_S is the cost of evaluating f_S . Therefore, the integration cost of a stabilized explicit scheme for (3.1) grows as $\sqrt{\rho_F} c_S$. Thus, very few degrees of freedom inducing a severe stiffness in f_F can completely destroy the efficiency of the method; albeit f_F is cheap to evaluate f_S is evaluated concurrently. The same discussion obviously apply to classical explicit methods.

In order to decouple the evaluation of f_F and f_S a popular strategy is to use interpolations or extrapolations, but this approach might render the integration unstable. Indeed, in Chapter 2 we proposed an interpolation based additive RKC method and proved on a model problem that the scheme is unstable; other explicit multirate methods face the same stability issues

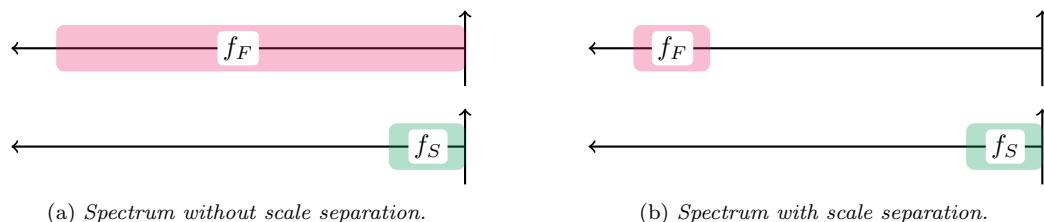


Figure 3.1. Stiff problems with identical spectral radii but different eigenvalues distribution.

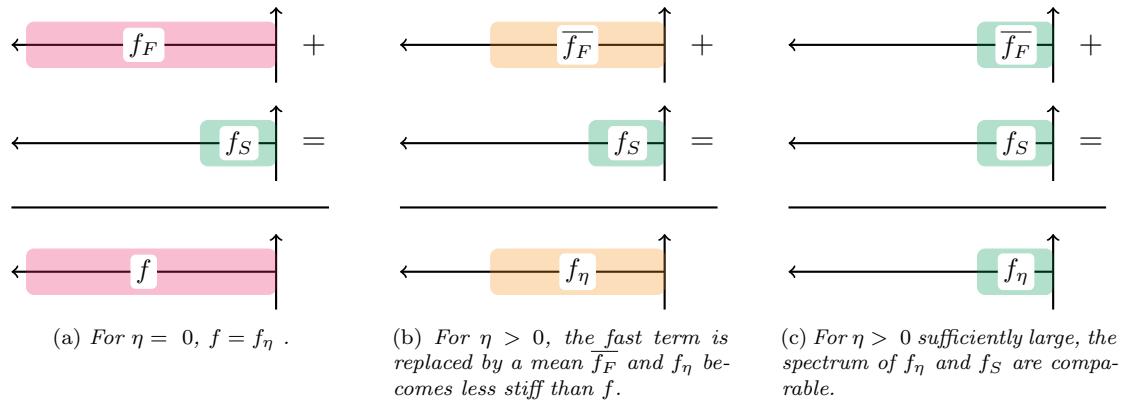


Figure 3.2. Spectrum of f and of the averaged force f_η for varying η .

[20, 21, 22, 44, 58, 73, 81, 102]. One could resort to multiscale schemes [42, 57], however these methods are employable only if there is a clear-cut separation of fast and slow dynamics as illustrated in Figure 3.1(b). Finally, implicit or multirate implicit methods [38, 50, 51, 64, 65, 94, 105, 106, 107, 108, 115] could be used, but they are more involved to implement, require the solution to possibly large and nonlinear systems and their efficiency depends on the availability of effective preconditioners.

In this chapter we propose an alternative approach based on a *modified equation*

$$y'_\eta = f_\eta(y_\eta), \quad y_\eta(0) = y_0, \quad (3.2)$$

where $f = f_F + f_S$ in (3.1) is replaced by an *averaged force*, f_η , which depends on a free parameter $\eta \geq 0$. For $\eta = 0$ it holds $f_\eta = f$ whereas for $\eta > 0$, the spectrum of f_η is compressed and thus f_η is less stiff than f , see Figure 3.2. In fact for $\eta > 0$ sufficiently large, the spectral radius ρ_η of the Jacobian of f_η is bounded by the spectral radius ρ_S of the Jacobian of f_S , i.e. $\rho_\eta \leq \rho_S$; then, the stiffness of the modified equation depends solely on f_S and its integration by any explicit method is cheaper than (3.1) integrated with the same method. Indeed, if (3.2) is solved with the explicit Euler scheme the step size restriction depends on the stiffness of f_S only. If it is solved with a stabilized explicit scheme the number of stages depends also on f_S only. In practice, evaluation of f_η requires the solution to a fast *auxiliary problem* in a short time interval; nonetheless, since f_F is cheap and the time interval is short, evaluation of f_η is not expensive when compared to $f = f_F + f_S$. Moreover, since the condition $\rho_\eta \leq \rho_S$ is already satisfied for η relatively small, f_η actually remains a good approximation of f .

The modified equation (3.2) is the starting point to derive a new class of explicit multirate RK schemes, where both the modified equation (3.2) and the auxiliary problem are discretized by a RK method, or even two different RK methods. This approach does not require interpolations nor extrapolations and is not based on any scale separation assumption. Moreover, even if the role of f_F and f_S changes in time, i.e. f_S becomes the stiff term, the solution remains accurate. Following this strategy, in this chapter we derive a multirate explicit Euler (mEE) method and the multirate RKC (mRKC) method, where both (3.2) and the auxiliary problem are solved with the explicit Euler or the RKC method, respectively. We briefly discuss a hybrid exponential Euler-RKC approach as well.

The chapter is structured as follows. In Section 3.1 we define the modified equation (3.2) and study its properties, in particular the stiffness and how well the exact solution y of (3.1) is approximated by the solution y_η of (3.2). In Section 3.2 we introduce the mEE method with its

efficiency, stability and accuracy analysis; however, the aim of this section is purely pedagogical and for illustration purposes, as the mRKC method is a more efficient scheme. In Section 3.3 we introduce the semidiscrete multirate RKC method, this scheme is defined as the time discretization of (3.2) with an RKC scheme. Since, in general, f_η cannot be evaluated exactly this scheme is not employable in practice. Still, we briefly discuss a case where f_η is computable, which gives rise to the multirate exponential Euler-RKC scheme. In Section 3.4 we present the mRKC scheme and study its theoretical efficiency. Then we prove stability on a scalar model problem and on a 2×2 model problem too, first-order accuracy for general problems (3.1) is also shown. Next we present the step size control strategy and only at the end of the section we consider problems with well separated time-scales and study the mRKC scheme under this additional assumption. Section 3.5 is dedicated to numerical experiments, where we confirm the theoretical findings and prove the efficiency of the method. The chapter is closed by Section 3.6, where we prove some technical results needed in Sections 3.2 and 3.4. Indeed, we postpone to the end of the chapter all the proofs that are purely technical and do not help in the understanding of the scheme. Nevertheless, the lemmas for which the proof has been postponed to Section 3.6 are accompanied by a figure corroborating the statement.

3.1 The modified equation

We define here the modified equation (3.2). We will study its accuracy, the damping properties and then how stiffness depends on the parameter η .

3.1.1 The averaged force

We begin this section defining f_η .

Definition 3.1. Let $\eta \geq 0$, $u_0 \in \mathbb{R}^n$ and $u : [0, \eta] \rightarrow \mathbb{R}^n$ be defined by the *auxiliary problem*

$$u' = f_F(u) + f_S(u_0) \quad t \in [0, \eta], \quad u(0) = u_0. \quad (3.3)$$

For $\eta > 0$, we define the *averaged force* f_η as

$$f_\eta(u_0) = \frac{1}{\eta}(u(\eta) - u_0) \quad (3.4)$$

and for $\eta = 0$ we define $f_0 = f$.

From (3.3) and (3.4) it follows that

$$f_\eta(u_0) = \frac{1}{\eta} \int_0^\eta u'(s) ds = f_S(u_0) + \frac{1}{\eta} \int_0^\eta f_F(u(s)) ds. \quad (3.5)$$

Hence, f_η is an average of f over the time interval $[0, \eta]$ along the *auxiliary solution* u .

Now, we wish to understand the properties of f_η and the role of the parameter η . First, we wish to know which is the effect of the contractivity of f_F on f_η . Then we show that f_η satisfies a one-sided Lipschitz condition, which is the fundamental property for proving convergence and contractivity for general nonlinear problems [69, Section IV.12]. Next, thanks to this property, we shall prove in Theorem 3.5, bounds on the error introduced by solving (3.2) instead of (3.1), which are independent of the problem's stiffness.

We denote by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ the dot product and the Euclidean norm in \mathbb{R}^n , respectively. To begin, we prove that if f_F satisfies a one-sided Lipschitz condition then it has a smoothing effect on f_η .

Lemma 3.2. Let $\mu_F \in \mathbb{R}$ and f_F satisfy

$$\langle f_F(z) - f_F(y), z - y \rangle \leq \mu_F \|z - y\|^2 \quad \forall z, y \in \mathbb{R}^n. \quad (3.6)$$

Then

$$\|f_\eta(u_0)\| \leq \varphi(\eta\mu_F) \|f(u_0)\|, \quad \text{where} \quad \varphi(z) = \frac{e^z - 1}{z} \quad \text{for } z \neq 0 \quad (3.7)$$

and $\varphi(0) = 1$ is defined by continuous extension. If, moreover, $f_F(y) = A_F y$ with $A_F \in \mathbb{R}^{n \times n}$ then

$$f_\eta(u_0) = \varphi(\eta A_F) f(u_0). \quad (3.8)$$

Proof. Let $v : [0, \eta] \rightarrow \mathbb{R}^n$ defined by $v(s) = u_0$ for all s . We set

$$\delta = \|v'(s) - f_F(v(s)) - f_S(u_0)\| = \|f(u_0)\|.$$

Since the logarithmic norm of the Jacobian of f_F is bounded by μ_F , we obtain from a classical result on differential inequalities (see [68, Chapter I.10, Theorem 10.6])

$$\|u(\eta) - u_0\| = \|u(\eta) - v(\eta)\| \leq e^{\eta\mu_F} \int_0^\eta e^{-s\mu_F} \delta \, ds = \eta \varphi(\eta\mu_F) \|f(u_0)\|,$$

which yields (3.7) by (3.4). Now, let us suppose that $f_F(u) = A_F u$ with $A_F \in \mathbb{R}^{n \times n}$ nonsingular. The variation-of-constants formula and replacing $f_F(u) = A_F u$ in (3.3) yield

$$u(\eta) = e^{A_F \eta} \left(u_0 + \int_0^\eta e^{-A_F s} f_S(u_0) \, ds \right) = e^{A_F \eta} u_0 + A_F^{-1} (e^{A_F \eta} - I) f_S(u_0),$$

with I the identity matrix. Hence,

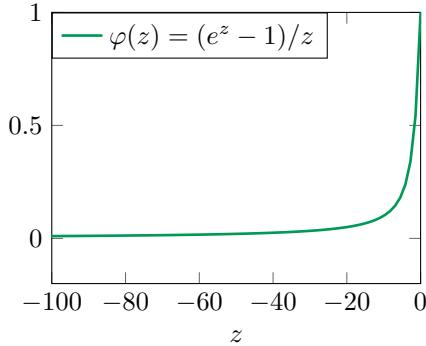
$$u(\eta) = e^{A_F \eta} u_0 + \eta \varphi(\eta A_F) f_S(u_0). \quad (3.9)$$

Since φ has no poles, $u(\eta)$ as in (3.9) is well-defined and satisfies (3.3) for all matrices A_F . By using (3.4), we thus obtain

$$f_\eta(u_0) = \frac{1}{\eta} (e^{A_F \eta} - I) u_0 + \varphi(\eta A_F) f_S(u_0) = \varphi(\eta A_F) (A_F u_0 + f_S(u_0)) = \varphi(\eta A_F) f(u_0). \quad \blacksquare$$

Remark 3.3. The entire function $\varphi(z)$ is very common in the theory of exponential integrators [72]. Indeed, when $f_F(y) = A_F y$, the solution (3.9) to the auxiliary problem (3.3) is nothing else than one step of the *exponential Euler* method applied to (3.1). Although we briefly discuss an approach using the exponential Euler scheme in Section 3.3, the framework presented here is very different from exponential integrators for several reasons. First, $u(\eta)$ is only an auxiliary solution used to compute f_η ; what we want to approximate is the solution y_η to the modified equation (3.2). Next, we must not use an exponential integrator to compute $u(\eta)$ but any explicit or implicit scheme is allowed; hence, in general $\varphi(\eta A_F)$ is not computed explicitly. Finally, here η is not the step size but a free parameter.

The function $\varphi(z)$ satisfies $\varphi(0) = 1$, $\lim_{z \rightarrow -\infty} \varphi(z) = 0$ and $\varphi(z) \in (0, 1)$ for $z < 0$, see Figure 3.3. Hence, if A_F is negative definite, multiplication of f by $\varphi(\eta A_F)$ in (3.8) has a smoothing effect, which can be adjusted by choosing $\eta \geq 0$ accordingly — see Theorem 3.7. A similar property holds for a nonlinear f_F which is contractive, i.e. with $\mu_F \leq 0$ in (3.6), because of (3.7). Next, we show that under certain assumptions on the Jacobians of f_F and f_S , the averaged force f_η satisfies a one-sided Lipschitz condition.

Figure 3.3. Illustration of $\varphi(z)$.

Theorem 3.4. Let $A_F \in \mathbb{R}^{n \times n}$ be symmetric and $f_F(y) = A_F y$. We suppose

$$\langle \frac{\partial f}{\partial y}(w)(z - y), z - y \rangle \leq \mu \|z - y\|^2 \quad \forall w, y, z \in \mathbb{R}^n \quad (3.10)$$

with $\mu \leq 0$ and that $A_F \frac{\partial f_S}{\partial y}(w) = \frac{\partial f_S}{\partial y}(w)A_F$ for all $w \in \mathbb{R}^n$. Then,

$$\langle f_\eta(z) - f_\eta(y), z - y \rangle \leq \mu_\eta \|z - y\|^2,$$

where $\mu_\eta = \mu \min_{\lambda \in \lambda(A_F)} \varphi(\eta \lambda) \leq 0$ and $\lambda(A_F)$ is the spectrum of A_F .

Proof. Let $y, z \in \mathbb{R}^n$ and $w(r) = rz + (1 - r)y$ for $r \in [0, 1]$. We have

$$\begin{aligned} \langle f_\eta(z) - f_\eta(y), z - y \rangle &= \langle \varphi(\eta A_F)(f(z) - f(y)), z - y \rangle \\ &= \langle \varphi(\eta A_F)^{1/2} \int_0^1 \frac{\partial f}{\partial y}(w(r))(z - y) dr, \varphi(\eta A_F)^{1/2}(z - y) \rangle. \end{aligned}$$

Since $\varphi(z) > 0$ for all z , $\varphi(\eta A_F)$ is symmetric positive definite and $\varphi(\eta A_F)^{1/2}$ exists. By hypothesis, $\varphi(\eta A_F)^{1/2}$ and $\frac{\partial f}{\partial y}(w(r))$ commute. Therefore

$$\begin{aligned} \langle f_\eta(z) - f_\eta(y), z - y \rangle &= \int_0^1 \langle \frac{\partial f}{\partial y}(w(r))\varphi(\eta A_F)^{1/2}(z - y), \varphi(\eta A_F)^{1/2}(z - y) \rangle dr \\ &\leq \mu \|\varphi(\eta A_F)^{1/2}(z - y)\|^2 \leq \mu \min_{\lambda \in \lambda(A)} \varphi(\eta \lambda) \|z - y\|^2. \end{aligned} \quad \blacksquare$$

Theorem 3.4 shows that f_η indeed satisfies a one-sided Lipschitz condition if the Jacobians of f_F and f_S commute and (3.10) holds, which is slightly stronger than asking that f is one-sided Lipschitz. Indeed, f is one-sided Lipschitz if, and only if, (3.10) holds for all z, y and $w \in [z, y]$, see [68, I.10]. The next theorem bounds the error between the solutions to (3.1) and (3.2).

Theorem 3.5. Under the assumptions of Theorem 3.4, it holds

$$\|y(t) - y_\eta(t)\| \leq \max_{\lambda \in \lambda(A_F)} |1 - \varphi(\eta \lambda)| \int_0^t e^{\mu_\eta(t-s)} \|f(y(s))\| ds, \quad (3.11)$$

with $\mu_\eta = \mu \min_{\lambda \in \lambda(A_F)} \varphi(\eta \lambda) \leq 0$.

Proof. Let y be the solution to (3.1). From Theorem 3.4, we have

$$\begin{aligned}\|y'(t) - f_\eta(y(t))\| &= \|f(y(t)) - f_\eta(y(t))\| = \|(I - \varphi(\eta A_F))f(y(t))\| \\ &\leq \max_{\lambda \in \lambda(A_F)} |1 - \varphi(\eta\lambda)| \|f(y(t))\| := \delta(t).\end{aligned}$$

Since the logarithmic norm of the Jacobian of f_η is bounded by $\mu_\eta = \mu \min_{\lambda \in \lambda(A_F)} \varphi(\eta\lambda)$, as implied by Theorem 3.4, the estimate (3.11) follows from a classical result on differential inequalities (see [68, Chapter I.10, Theorem 10.6]). \blacksquare

Observe that the bound in (3.11) is independent of the stiffness of the problem and since $\varphi(\eta\lambda) = 1 + \mathcal{O}(\eta)$ as $\eta \rightarrow 0$ then $\|y(t) - y_\eta(t)\| \leq C\eta$ as $\eta \rightarrow 0$.

3.1.2 Multirate test equation and stability

In this section we want to study the stiffness of the modified problem (3.2), i.e. the spectral radius ρ_η of the Jacobian of f_η . In particular, we wish to know under which conditions it holds that $\rho_\eta \leq \rho_S$, with ρ_S the spectral radius of the Jacobian of f_S , and hence stiffness of the modified problem solely depends on the slow components.

Under the hypotheses of Theorem 3.4, the Jacobians of f_F and f_S commute; thus, they are simultaneously triangularizable. Therefore, the stability analysis of (3.1) and (3.2) can be reduced to a scalar equation. Thus, we henceforth consider the *multirate test equation*

$$y' = \lambda y + \zeta y, \quad y(0) = y_0, \quad (3.12)$$

with $\lambda, \zeta \leq 0$ and $y_0 \in \mathbb{R}$, which corresponds to setting $f_F(y) = \lambda y$ and $f_S(y) = \zeta y$; thus, $\rho_F = |\lambda|$ and $\rho_S = |\zeta|$. Since we do not make any scale separation assumption, λ can take any nonpositive value. Equation (3.12) satisfies the hypothesis of Lemma 3.2 with $A_F = \lambda$, thus

$$u(\eta) = (e^{\eta\lambda} + \varphi(\eta\lambda)\eta\zeta)u_0, \quad (3.13)$$

$$f_\eta(u_0) = \varphi(\eta\lambda)(\lambda + \zeta)u_0 \quad (3.14)$$

and (3.2) becomes

$$y'_\eta = \varphi(\eta\lambda)(\lambda + \zeta)y_\eta, \quad y_\eta(0) = y_0. \quad (3.15)$$

In the rest of this section, we study the conditions on η, λ and ζ so that

$$|\varphi(\eta\lambda)(\lambda + \zeta)| \leq |\zeta|$$

holds, as the stiffness of (3.15) then depends exclusively on $\rho_S = |\zeta|$. The next lemma is used to prove Theorem 3.7 below.

Lemma 3.6. *Let $w \leq 0$, it holds that $\varphi(z)(z+w) \in [w, 0]$ for all $z \leq 0$ if, and only if, $\varphi'(0)|w| \geq 1$. As $\varphi'(0) = 1/2$, then $|w| \geq 2$.*

Proof. Since $z, w \leq 0$, it clearly holds $\varphi(z)(z+w) \leq 0$. Hence, we must show

$$\varphi(z)(z+w) \geq w. \quad (3.16)$$

For $z = 0$ it is obvious. Suppose that (3.16) holds for $z < 0$, in the neighborhood of $z = 0$ it yields

$$\begin{aligned}0 &\leq \varphi(z)(z+w) - w \\ &= (\varphi(0) + \varphi'(0)z + \mathcal{O}(z^2))(z+w) - w \\ &= z(1 + \varphi'(0)(z+w)) + \mathcal{O}(z^2(z+w)),\end{aligned} \quad (3.17)$$

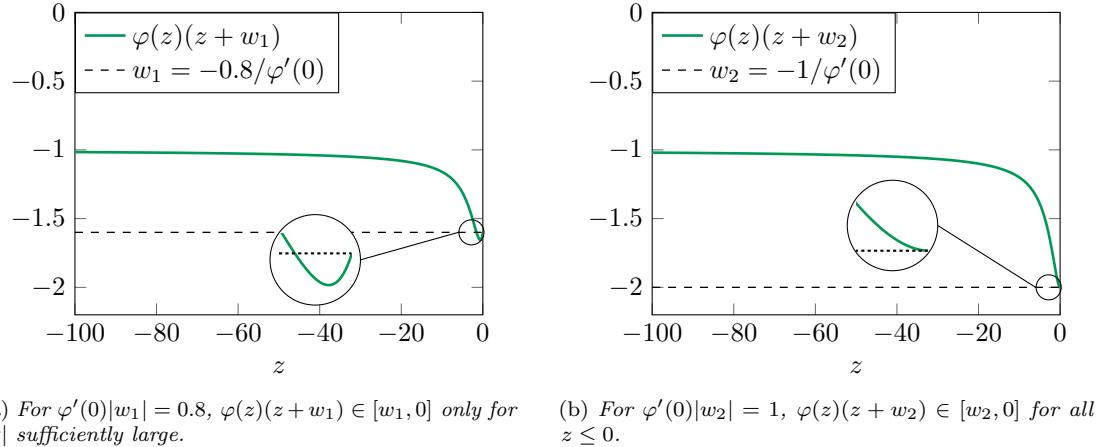


Figure 3.4. Illustration of the condition $\varphi(z)(z + w) \in [w, 0]$ for all $z \leq 0$, with two different values of w .

where we used $\varphi(0) = 1$. Dividing (3.17) by $z < 0$ and letting $z \rightarrow 0$ yields $\varphi'(0)w \leq -1$, hence $\varphi'(0)|w| \geq 1$. The identity $\varphi'(0) = 1/2$ is easily obtained from the definition of $\varphi(z)$ in (3.7).

Now, let us suppose $\varphi'(0)|w| \geq 1$ and thus $w \leq -1/\varphi'(0) = -2$, we multiply (3.16) by $z < 0$ and show the equivalent condition

$$\alpha(z) = zw + (1 - e^z)(z + w) \geq 0 \quad \forall z < 0.$$

It holds

$$\alpha'(z) = 1 + w - (1 + w + z)e^z \quad \text{and} \quad \alpha''(z) = -(2 + w + z)e^z,$$

thus $\alpha(0) = \alpha'(0) = 0$ and

$$\alpha(z) = \int_0^z \int_0^s \alpha''(r) dr ds = - \int_z^0 \int_s^0 (2 + w + r)e^r dr ds \geq 0,$$

indeed $w \leq -2$ and therefore $2 + w + r \leq 0$ for all $r \leq 0$. ■

In Figure 3.4 we fix w , let $z \leq 0$ vary and verify if condition $\varphi(z)(z + w) \in [w, 0]$ holds for all $z \leq 0$. In Figure 3.4(a) we let $\varphi'(0)|w| < 1$ and note that $\varphi(z)(z + w) \in [w, 0]$ holds only for z away from zero. In Figure 3.4(b) we have $\varphi'(0)|w| = 1$ and see that $\varphi(z)(z + w) \in [w, 0]$ holds for all $z \leq 0$.

Theorem 3.7. Let $\zeta < 0$. Then, $\varphi(\eta\lambda)(\lambda + \zeta) \in [\zeta, 0]$ for all $\lambda \leq 0$ if, and only if, $\eta \geq 2/|\zeta|$.

Proof. Setting $z = \eta\lambda$ and $w = \eta\zeta$, we have that

$$\varphi(\eta\lambda)(\lambda + \zeta) \in [\zeta, 0] \quad \text{is equivalent to} \quad \varphi(z)(z + w) \in [w, 0].$$

In view of Lemma 3.6, this holds for all $\lambda \leq 0$, if and only if $\eta|\zeta| = |w| \geq 2$. ■

Theorem 3.7 implies that for $\eta \geq 2/\rho_S$ the stiffness of (3.15) depends only on the slow term f_S . Since the result holds for all $\lambda \leq 0$, there is no need for any assumption on scale separation.

3.2 The multirate explicit Euler method

In this section we illustrate, in a simple case, how the modified equation framework can be used to derive an explicit multirate scheme. Integrating (3.2) and (3.3) with the explicit Euler scheme we derive the multirate explicit Euler (mEE) method. We will define the method, study its efficiency and then prove its stability and accuracy; however, the aim of this section is purely illustrative and no numerical experiments are provided for the mEE method.

3.2.1 The mEE algorithm

Let $\tau >$ be the step size used to integrate (3.2), $\Delta\tau > 0$ be the micro step size used for (3.3) and $N \in \mathbb{N}$ the number of micro time steps satisfying

$$\tau\rho_S \leq 2, \quad \Delta\tau\rho_F \leq 2, \quad N \geq 1 + \frac{\tau}{\Delta\tau}, \quad \eta = N\Delta\tau, \quad (3.18)$$

with ρ_S, ρ_F the spectral radii of the Jacobians of f_S, f_F , respectively. One step of the mEE scheme is given by

$$y_{n+1} = y_n + \tau\bar{f}_\eta(y_n), \quad (3.19)$$

where

$$\bar{f}_\eta(u_0) = \frac{1}{\eta}(u_N - u_0) \quad (3.20)$$

is the numerical counterpart of f_η in (3.4). The approximation u_N of $u(\eta)$ is given by N steps of size $\Delta\tau$ of the explicit Euler scheme applied to (3.3), hence

$$u_{j+1} = u_j + \Delta\tau(f_F(u_j) + f_S(u_0)), \quad j = 0, \dots, N-1. \quad (3.21)$$

Note that $N \geq 2$ for any $\tau, \Delta\tau$ and in general $\Delta\tau \leq \tau$, as $\rho_S \leq \rho_F$. If, because of nonlinearities, at some point during integration it holds $\rho_S > \rho_F$ then we use $\Delta\tau = \tau$ and $N = 2$. Therefore, $\Delta\tau \leq \tau$ for any ρ_S, ρ_F .

Stability and first-order accuracy of the mEE scheme (3.18) to (3.21) are proved in Theorems 3.12 and 3.13 in Section 3.2.3 below.

3.2.2 Efficiency of the mEE method

Given ρ_S, ρ_F we now are interested in the theoretical efficiency of the mEE scheme compared to the standard explicit Euler (EE) scheme. As we consider a wide range of values for ρ_F we suppose that the spectral radius of the Jacobian of f is $\rho = \rho_F + \rho_S$, hence for large ρ_F it holds $\rho \approx \rho_F$ and for small ρ_F holds $\rho \approx \rho_S$. We denote the cost of evaluating f_F, f_S relatively to the cost of evaluating f itself by $c_F, c_S \in [0, 1]$, with $c_F + c_S = 1$. We also consider a situation where the step size is limited by stability, hence we neglect accuracy requirements when choosing τ .

The EE scheme needs a step size $\tau = 2/\rho = 2/(\rho_F + \rho_S)$ and therefore, supposing that we integrate (3.1) in the interval $[0, 1]$, it needs $N_{\text{EE}} = 1/\tau = (\rho_F + \rho_S)/2$ time steps. At each time step it evaluates f_F and f_S , hence the integration cost is

$$C_{\text{EE}} = N_{\text{EE}}(c_F + c_S) = \frac{\rho_F + \rho_S}{2}.$$

From (3.18) follows that for the mEE method we have $\tau = 2/\rho_S$, $\Delta\tau = 2/\rho_F$ and $N \in \mathbb{N}$ such that $N \geq 1 + \tau/\Delta\tau = 1 + \rho_F/\rho_S$. Letting $r_\rho = \rho_F/\rho_S$ it holds

$$N = 1 + \left\lceil \frac{\rho_F}{\rho_S} \right\rceil = 1 + \lceil r_\rho \rceil,$$

with $\lceil x \rceil$ denoting the smallest integer greater than or equal to $x \in \mathbb{R}$. The number of time steps taken by the mEE method is $N_{\text{mEE}} = 1/\tau = \rho_S/2$ and the cost of one step is one evaluation of f_S and N evaluations of f_F (see (3.21)); therefore, the total integration cost of the mEE scheme is

$$C_{\text{mEE}} = N_{\text{mEE}}(c_F N + c_S) = \frac{\rho_S}{2} (c_F(1 + \lceil r_\rho \rceil) + 1 - c_F) = \frac{\rho_S}{2} (1 + c_F \lceil r_\rho \rceil), \quad (3.22)$$

where we used $c_F + c_S = 1$. The relative speed-up S is defined as the ratio between the two costs, hence

$$S = \frac{C_{\text{EE}}}{C_{\text{mEE}}} = \frac{\rho_F + \rho_S}{\rho_S(1 + c_F \lceil r_\rho \rceil)} = \frac{1 + r_\rho}{1 + c_F \lceil r_\rho \rceil} \quad (3.23)$$

with $r_\rho = \rho_F/\rho_S \in [0, \infty)$. Let us consider, for the time being, $r_\rho \in \mathbb{N}$ and therefore $\lceil r_\rho \rceil = r_\rho$; then it holds $S \in [1, 1 + r_\rho]$ and thus the mEE method is always faster than the EE scheme. We plot S in function of c_F for some values of $r_\rho \in \mathbb{N}$ in Figure 3.5(a), we observe that indeed $S > 1$ for any value of $c_F < 1$ and S approaches $1 + r_\rho$ as $c_F \rightarrow 0$. Finally, we observe that for $r_\rho \in \mathbb{N}$ it holds, from (3.22),

$$C_{\text{mEE}} = \frac{\rho_S}{2} \left(1 + c_F \frac{\rho_F}{\rho_S} \right) = \frac{c_F \rho_F + \rho_S}{2}$$

and therefore the stiffness of f_F , which increases the computational cost, is weighted by the evaluation cost of f_F , which is low. Hence, if stiffness of f_F is induced by a few degrees of freedom the efficiency of the mEE scheme is not severely affected.

Now, we are interested on the case where $r_\rho \in \mathbb{R}$, as in this situation $S < 1$ is possible. We solve the inequality $S > 1$ with S as in (3.23) in order to identify the maximal value for c_F that still leads to reduced cost in using the mEE scheme over the EE scheme. It holds $S > 1$ if, and only if,

$$c_F < c_F^{\max} = \frac{r_\rho}{\lceil r_\rho \rceil}.$$

We display in Figure 3.5(b) the value of c_F^{\max} for varying r_ρ . In general $\rho_S < \rho_F$ and thus $r_\rho > 1$; we see in Figure 3.5(b) that $c_F^{\max} > 0.5$ as $r_\rho > 1$ and rapidly approaches 1 as r_ρ grows. Therefore, although in our hypothesis evaluation of f_F is cheap, in practice it can be relatively expensive. We conclude from this complexity analysis that as long as f_F is stiffer than f_S the mEE method will always outperform the EE scheme, except in extreme cases where $c_F \approx 1$. For $\rho_F < \rho_S$ and thus $r_\rho < 1$ then c_F^{\max} must be relatively small but not necessarily zero; this case is relevant for nonlinear problems, where f_F, f_S can temporarily switch their role.

3.2.3 Stability and convergence analysis

This section is devoted to the analysis of the mEE method, we show that it is stable and first-order accurate.

Stability analysis on the multirate test equation

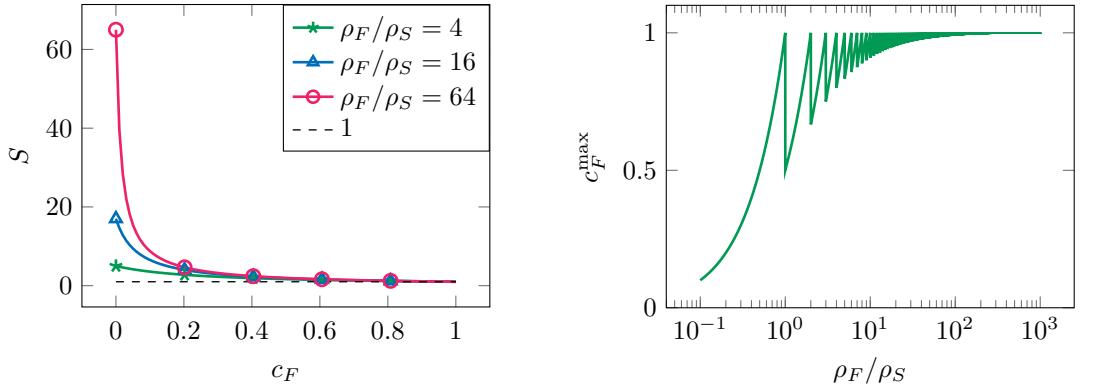
We will show stability of the method when applied to the multirate test equation (3.12), hence we will compute the stability polynomial of the mEE scheme and show that under conditions (3.18) it is bounded by one. We start computing a closed expression for u_N in (3.21).

Lemma 3.8. *Let $\lambda, \zeta \in \mathbb{R}$, $f_F(y) = \lambda y$ and $f_S(y) = \zeta y$. Then the solution u_N of (3.21) is given by*

$$u_N = ((1 + \Delta\tau\lambda)^N + N\Phi_N^{EE}(\Delta\tau\lambda)\Delta\tau\zeta)u_0,$$

where

$$\Phi_N^{EE}(z) = \frac{1}{N} \frac{(1+z)^N - 1}{z} \quad \text{for } z \neq 0 \quad (3.24)$$



(a) Theoretical speed-up of the mEE method over the standard EE scheme, for varying c_F and fixed $\rho_F/\rho_S \in \mathbb{N}$.

(b) Maximal c_F which still yields speed-up $S > 1$, w.r.t. $\rho_F/\rho_S \in \mathbb{R}$.

Figure 3.5. The relative speed-up S of the mEE method over the EE scheme with respect to c_F and the maximal value for c_F which still leads to an efficiency gain.

and $\Phi_N^{EE}(0) = \lim_{z \rightarrow 0} \Phi_N^{EE}(z) = 1$.

Proof. Replacing $f_F(y) = \lambda y$ and $f_S(y) = \zeta y$ into (3.21), yields

$$u_{j+1} = (1 + \Delta\tau\lambda)u_j + \Delta\tau\zeta u_0, \quad j = 0, \dots, N-1$$

and by recursion we obtain

$$\begin{aligned} u_N &= (1 + \Delta\tau\lambda)^N u_0 + \sum_{k=0}^{N-1} (1 + \Delta\tau\lambda)^k \Delta\tau\zeta u_0 \\ &= (1 + \Delta\tau\lambda)^N u_0 + \frac{(1 + \Delta\tau\lambda)^N - 1}{\Delta\tau\lambda} \Delta\tau\zeta u_0 \\ &= (1 + \Delta\tau\lambda)^N u_0 + N\Phi_N^{EE}(\Delta\tau\lambda)\Delta\tau\zeta u_0. \end{aligned}$$

■

Since $\eta = N\Delta\tau$, then

$$\bar{f}_\eta(u_0) = \frac{1}{\eta}(u_N - u_0) = \frac{(1 + \Delta\tau\lambda)^N - 1}{N\Delta\tau} u_0 + \Phi_N^{EE}(\Delta\tau\lambda)\zeta u_0 = \Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta)u_0 \quad (3.25)$$

and thus, from (3.19),

$$y_{n+1} = (1 + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta))y_n, \quad (3.26)$$

which motivates the following definition.

Definition 3.9. Let $\lambda, \zeta \leq 0$, $\tau, \Delta\tau > 0$ and $N \in \mathbb{N}$. The stability polynomial of the mEE method (3.18) to (3.21) is given by

$$R_N(\lambda, \zeta, \tau, \Delta\tau) = 1 + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta).$$

Hence, from (3.26) follows that the mEE scheme is stable if, and only if, $|R_N(\lambda, \zeta, \tau, \Delta\tau)| \leq 1$. In order to prove stability in Theorem 3.12 below we make use of the next two lemmas, their proof is purely technical and postponed to Section 3.6.1. The next result is the numerical counterpart of $\varphi(z) \in [0, 1]$ for all $z \leq 0$, with $\varphi(z)$ as in (3.7).

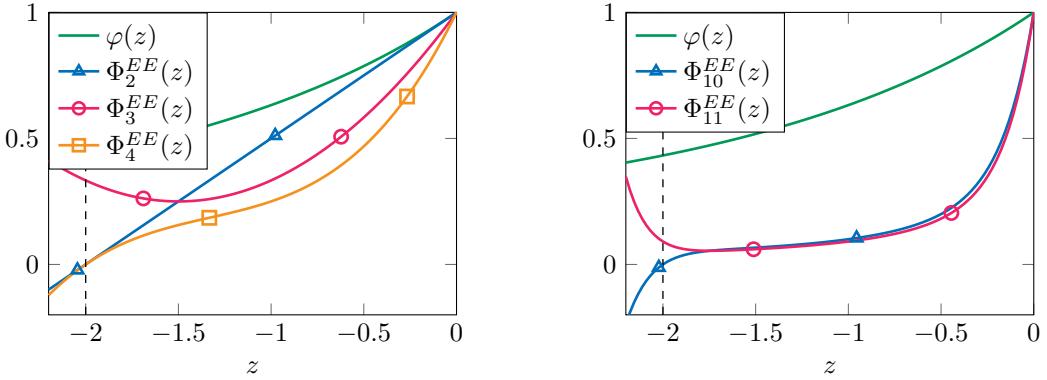


Figure 3.6. Illustration of $\varphi(z)$ and $\Phi_N^{EE}(z)$ for $N = 2, 3, 4$ (left) and $N = 10, 11$ (right). The dashed line indicates the end of the stability domain.

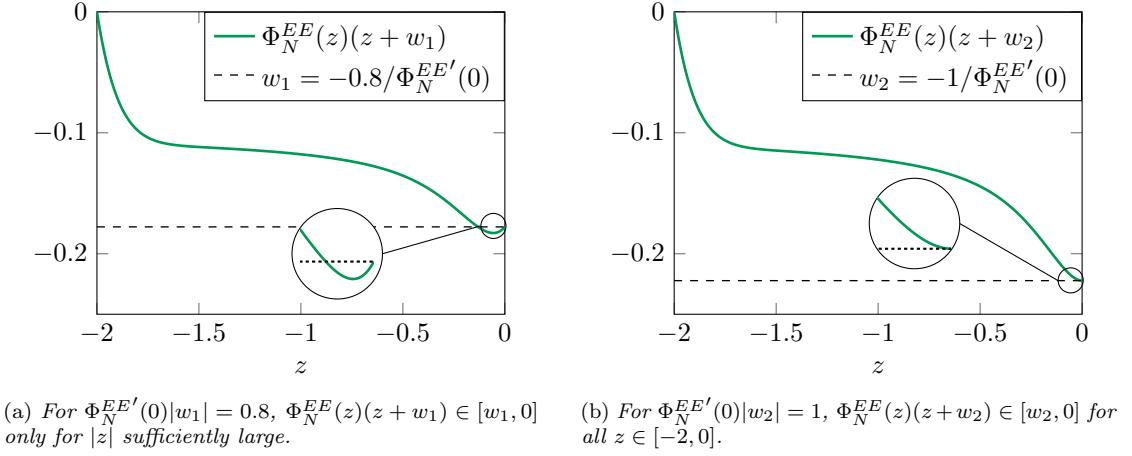


Figure 3.7. Illustration of the condition $\Phi_N^{EE}(z)(z + w) \in [w, 0]$ for all $z \in [-2, 0]$, with $N = 10$ and two different values for w .

Lemma 3.10. It holds $\Phi_N^{EE}(z) \in [0, 1]$ for all $N \in \mathbb{N}$ if, and only if, $z \in [-2, 0]$.

In Figure 3.6 we display $\Phi_N^{EE}(z)$ for some values of N and we note that $\Phi_N^{EE}(z) \in [0, 1]$ is indeed satisfied for $z \in [-2, 0]$.

The next result is the discrete counterpart of Lemma 3.6.

Lemma 3.11. Let $w < 0$ and $N \geq 2$, it holds $\Phi_N^{EE}(z)(z + w) \in [w, 0]$ for all $z \in [-2, 0]$ if, and only if, $\Phi_N^{EE'}(0)|w| \geq 1$. As $\Phi_N^{EE'}(0) = (N-1)/2$, then $|w| \geq 2/(N-1)$.

In Figure 3.7 we display condition $\Phi_N^{EE}(z)(z + w) \in [w, 0]$ for all $z \in [-2, 0]$, with $N = 10$ and two different values for w . We see in Figure 3.7(a) that $\Phi_{10}^{EE}(z)(z + w) \in [w, 0]$ only for z away from zero, indeed $\Phi_{10}^{EE'}(0)|w| \geq 1$ is not satisfied. In contrast, in Figure 3.7(b) we consider $\Phi_{10}^{EE'}(0)|w| = 1$ and $\Phi_{10}^{EE}(z)(z + w) \in [w, 0]$ holds for all $z \in [-2, 0]$.

Let us now prove that the mEE scheme is stable.

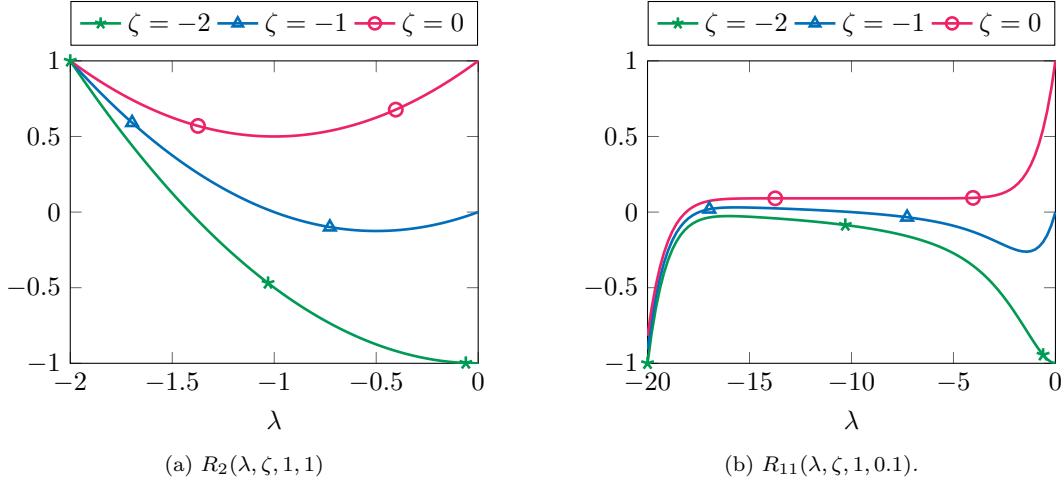


Figure 3.8. Stability polynomial $R_N(\lambda, \zeta, \tau, \Delta\tau)$ of the mEE method vs. λ varying in $[-2/\Delta\tau, 0]$, for $\zeta = -2, -1$ or 0 and $\tau = 1$. We let $\Delta\tau = 1$ (left) and $\Delta\tau = 0.1$ (right), with N as in (3.27).

Theorem 3.12. Let $\lambda, \zeta \leq 0$, for

$$\tau|\zeta| \leq 2, \quad \Delta\tau|\lambda| \leq 2, \quad N \geq 1 + \frac{\tau}{\Delta\tau} \quad (3.27)$$

it holds $|R_N(\lambda, \zeta, \tau, \Delta\tau)| \leq 1$, i.e. the multirate explicit Euler method is stable.

Proof. Inequality $|R_N(\lambda, \zeta, \tau, \Delta\tau)| \leq 1$ is equivalent to $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \in [-2, 0]$, since $\Phi_N^{EE}(z) \geq 0$ for all $z \in [-2, 0]$ then $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \leq 0$ and it is sufficient to show $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq -2$. We have $\Delta\tau\zeta \geq -2\Delta\tau/\tau$ and thus

$$\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) = \frac{\tau}{\Delta\tau}\Phi_N^{EE}(\Delta\tau\lambda)(\Delta\tau\lambda + \Delta\tau\zeta) \geq \frac{\tau}{\Delta\tau}\Phi_N^{EE}(\Delta\tau\lambda)(\Delta\tau\lambda - 2\frac{\Delta\tau}{\tau}) \geq -2,$$

where the last inequality follows from Lemma 3.11 with $z = \Delta\tau\lambda$, $w = -2\Delta\tau/\tau$ and thus $|w| \geq 2/(N-1)$. \blacksquare

In Figure 3.8 we display the stability polynomial $R_N(\lambda, \zeta, \tau, \Delta\tau)$ as a function of λ for different values of $\zeta, \tau, \Delta\tau, N$ satisfying (3.27). We see that $|R_N(\lambda, \zeta, \tau, \Delta\tau)| \leq 1$ is indeed satisfied and the mEE scheme is therefore stable.

Convergence analysis

Here we prove first-order accuracy of the mEE method. It is known that a RK scheme is first-order accurate if, and only if, it is first order-accurate for linear equations [69, Section IV.2]. Hence, it is sufficient to prove first-order accuracy on the multirate test equation.

Theorem 3.13. The mEE method is first-order accurate.

Proof. Let $y_1(\tau, \Delta\tau) = R_N(\lambda, \zeta, \tau, \Delta\tau)y_0 = (1 + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta))y_0$, hence

$$\nabla y_1(\tau, \Delta\tau) = \begin{pmatrix} \Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \\ \tau\lambda\Phi_N^{EE'}(\Delta\tau\lambda)(\lambda + \zeta) \end{pmatrix} y_0 \quad \text{and} \quad \nabla y_1(0, 0) = \begin{pmatrix} \lambda + \zeta \\ 0 \end{pmatrix} y_0.$$

Thus, Taylor expanding $y_1(\tau, \Delta\tau)$ in $(\tau, \Delta\tau) = (0, 0)$ and using $\Delta\tau \leq \tau$ yields

$$y_1(\tau, \Delta\tau) = y_1(0, 0) + \nabla y_1(0, 0) \cdot (\tau, \Delta\tau) + \mathcal{O}(\tau^2 + \Delta\tau^2) = y_0 + \tau(\lambda + \zeta)y_0 + \mathcal{O}(\tau^2).$$

Therefore, $y_1(\tau, \Delta\tau) - y(\tau) = \mathcal{O}(\tau^2)$, which implies first-order accuracy. ■

3.3 The semidiscrete multirate RKC method

In order to prepare for the multirate RKC method, we first present a semidiscrete method, called semidiscrete multirate RKC method, defined as the time discretization of (3.2) with an s -stage RKC scheme, while (3.3) is supposed to be integrated exactly. We will as well discuss a possible realization of such a method in practice, then we study its stability and accuracy.

3.3.1 The semidiscrete multirate RKC algorithm

The semidiscrete multirate RKC method is defined as a first-order RKC scheme (see Section 1.2.4) applied to (3.2), with f_η as in Definition 3.1 and hence (3.3) is integrated exactly. The number of stages s is chosen according to the stiffness of the slow term f_S and the parameter η is chosen such that the spectrum of τf_η fits inside the stability domain of the s -stage RKC method. Thus, η depends on the slow term f_S but only indirectly; indeed, it depends on the s -stage RKC scheme, which in turn depends on f_S .

More precisely, let ρ_S be the spectral radius of the Jacobian of f_S , s such that

$$\tau\rho_S \leq \beta s^2 \quad \text{and} \quad \eta = \frac{2\tau}{\beta s^2}. \quad (3.28)$$

One step of the semidiscrete multirate RKC scheme is given by

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau f_\eta(k_0), \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j \tau f_\eta(k_{j-1}), \quad j = 2, \dots, s, \\ y_{n+1} &= k_s, \end{aligned} \quad (3.29)$$

where f_η is given in Definition 3.1.

Stability and first-order accuracy of the semidiscrete multirate RKC scheme (3.28) and (3.29) are proved in Theorems 3.14 and 3.15 in Section 3.3.3 below.

3.3.2 The multirate exponential Euler-RKC method

For general problems (3.1) equation (3.3) cannot be solved exactly but must be approximated, hence f_η cannot be evaluated and method (3.29) is not employable in practice. However, if $f_F(y) = A_F y$, for some matrix $A_F \in \mathbb{R}^{n \times n}$, we saw in Lemma 3.2 that

$$u(\eta) = e^{A_F \eta} u_0 + \eta \varphi(\eta A_F) f_S(u_0), \quad (3.30a)$$

$$f_\eta(u_0) = \varphi(\eta A_F) f(u_0), \quad (3.30b)$$

where (3.30a) is equivalent to an exponential Euler step, of size η , applied to (3.3). In Remark 3.3 we indeed pointed out that the exponential-like function $\varphi(z)$ is largely used in the context of exponential integrators and numerous efficient routines approximating the matrix function

$\varphi(\eta A_F)$, or its product with a vector, exist — see [71, 72] for a review. Hence, $f_\eta(u_0)$ as in (3.30b) can be efficiently approximated by

$$\bar{f}_\eta(u_0) = \bar{\varphi}(\eta A_F) f(u_0), \quad (3.31)$$

where $\bar{\varphi}(\eta A_F)$ approximates $\varphi(\eta A_F)$. Replacing f_η in (3.29) by \bar{f}_η as in (3.31) gives rise to the multirate exponential Euler-RKC method:

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau \bar{f}_\eta(k_0), \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j \tau \bar{f}_\eta(k_{j-1}), \quad j = 2, \dots, s, \\ y_{n+1} &= k_s. \end{aligned} \quad (3.32)$$

Precise stability conditions for scheme (3.32) are not derived here as they depend on the spectrum of the Jacobian of \bar{f}_η and therefore on the linear algebra routine used to compute $\bar{\varphi}(\eta A_F)$. In contrast, when $\varphi(\eta A_F)$ can be evaluated exactly then the multirate exponential Euler-RKC method (3.32) is equivalent to the semidiscrete RKC method (3.29) and the stability conditions are given by (3.28).

A precise efficiency study of (3.32) is also not performed here as it depends on many factors as the routine used for $\bar{\varphi}(\eta A_F)$, the structure of A_F , the values of η, τ and s ; hence we can only give some qualitative comments. Assuming that s, η are chosen as in (3.28), if we compare (3.32) to the standard exponential Euler scheme $y_{n+1} = e^{\tau A_F} y_n + \tau \bar{\varphi}(\tau A_F) f_S(y_n)$ we see that the multirate scheme computes $\bar{\varphi}(\eta A_F)$ instead of $\bar{\varphi}(\tau A_F)$. In general, convergence of a linear algebra routine for $\bar{\varphi}(\tau A_F)$ gets faster as τ decreases, hence, since $\eta \ll \tau$ (see (3.28)) then $\bar{\varphi}(\eta A_F)$ is computed faster than $\bar{\varphi}(\tau A_F)$; on the other hand $\bar{\varphi}(\eta A_F)$ is computed s times. Compared to a standard RKC scheme, the multirate exponential Euler-RKC method has a number of stages s depending on the slow scale only (see (3.28)), hence it needs much less function evaluations. In contrast, evaluation of \bar{f}_η is more involved than $f = f_F + f_S$.

3.3.3 Stability and convergence analysis

Here we show that the semidiscrete multirate RKC scheme applied to the test equation (3.12) is stable and that the scheme is first-order accurate for general equations (3.1).

Stability analysis on the multirate test equation

As we study stability of scheme (3.28) and (3.29) when applied to the test equation (3.12), it holds $f_\eta(k_j) = \varphi(\eta\lambda)(\lambda + \zeta)k_j$ (see (3.8)) and thus (3.29) becomes

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau \varphi(\eta\lambda)(\lambda + \zeta)k_0, \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j \tau \varphi(\eta\lambda)(\lambda + \zeta)k_{j-1}, \quad j = 2, \dots, s, \\ y_{n+1} &= k_s. \end{aligned} \quad (3.33)$$

From Section 1.2.4 follows $y_{n+1} = R_s(\tau \varphi(\eta\lambda)(\lambda + \zeta))y_n$, where $R_s(z)$ is the stability polynomial of the RKC scheme, given in (1.7). Hence, scheme (3.33) is stable if, and only if, $|R_s(\tau \varphi(\eta\lambda)(\lambda + \zeta))| \leq 1$.

Theorem 3.14. *Let $\lambda \leq 0$ and $\zeta < 0$. For all $\tau > 0$, s such that*

$$\tau|\zeta| \leq \beta s^2 \quad \text{and} \quad \eta = \frac{2\tau}{\beta s^2}$$

then $|R_s(\tau\varphi(\eta\lambda)(\lambda + \zeta))| \leq 1$, i.e. the semidiscrete multirate RKC scheme is stable.

Proof. If $|\tau\varphi(\eta\lambda)(\lambda + \zeta)| \leq \beta s^2$ then $|R_s(\tau\varphi(\eta\lambda)(\lambda + \zeta))| \leq 1$, see Section 1.1. Since $\varphi(\eta\lambda)(\lambda + \zeta) \leq 0$ equation

$$|\tau\varphi(\eta\lambda)(\lambda + \zeta)| \leq \beta s^2 \quad \text{is equivalent to} \quad \varphi(\eta\lambda)(\eta\lambda + \eta\zeta) \in [w(\eta), 0],$$

with $w(\eta) = -\eta\beta s^2/\tau$. We have $\tau|\zeta| \leq \beta s^2$ and hence $\zeta \geq -\beta s^2/\tau$, therefore it yields

$$\eta\zeta \geq w(\eta) \quad \text{and} \quad 0 \geq \varphi(\eta\lambda)(\eta\lambda + \eta\zeta) \geq \varphi(z)(z + w(\eta)),$$

with $z = \eta\lambda$. Thus, $\varphi(z)(z + w(\eta)) \geq w(\eta)$ for all $z \leq 0$ is sufficient for stability. From Lemma 3.6 it holds if $|w(\eta)| \geq 2$, which is true by choice of η . \blacksquare

Convergence analysis

We end this section showing that the semidiscrete multirate RKC scheme is first-order accurate.

Theorem 3.15. *The semidiscrete multirate RKC scheme has first-order of accuracy.*

Proof. Since the RKC scheme is first-order accurate and the semidiscrete multirate RKC scheme is defined as the integration of (3.2) by an RKC scheme, then $y_1 = y_\eta(\tau) + \mathcal{O}(\tau^2)$. From (3.5) it also follows that $f_\eta(u_0) = f(u_0) + \mathcal{O}(\eta)$ and hence

$$y_\eta(\tau) = y_0 + \tau f_\eta(y_0) + \mathcal{O}(\tau^2) = y_0 + \tau f(y_0) + \mathcal{O}(\tau^2 + \tau\eta) = y(\tau) + \mathcal{O}(\tau^2 + \tau\eta).$$

Since $\eta = 2\tau/(\beta s^2)$, then $\mathcal{O}(\tau\eta) = \mathcal{O}(\tau^2)$ and $y_1 - y(\tau) = \mathcal{O}(\tau^2)$, which implies first-order accuracy. \blacksquare

3.4 The multirate RKC method

We finally introduce the multirate RKC (mRKC) method, it is a third example of how the modified equation framework of Section 3.1 is used to derive explicit multirate methods. The mRKC method is obtained by discretizing (3.2) with an s -stage RKC method and approximating f_η (given in Definition 3.1) by solving (3.3) with one step of an m -stage RKC method. We will define the mRKC algorithm ((3.34) to (3.37) below) and then compare its efficiency to that of the standard RKC method of Section 1.2.4. Next we prove stability on the multirate test equation (3.12) and on a 2×2 model problem as well, then we show first-order accuracy for general problems (3.1). Later we discuss a time step control strategy and the mRKC method for problems with well separated scales. Numerical experiments are provided in Section 3.5.

3.4.1 The mRKC algorithm

Let $\tau > 0$ be the step size and ρ_F, ρ_S the spectral radii of the Jacobians of f_F, f_S , respectively (they can be cheaply estimated employing nonlinear power methods [86, 123]). Now, let the stages s, m be the smallest integers satisfying

$$\tau\rho_S \leq \beta s^2, \quad \eta\rho_F \leq \beta m^2, \quad \text{with} \quad \eta = \frac{6\tau}{\beta s^2} \frac{m^2}{m^2 - 1}, \quad (3.34)$$

$\beta = 2 - 4\varepsilon/3$ and $\varepsilon = 0.05$, typically. One step of the mRKC scheme is given applying the first-order RKC scheme (1.22) to (3.2), hence

$$\begin{aligned} k_0 &= y_n, \\ k_1 &= k_0 + \mu_1 \tau \bar{f}_\eta(k_0), \\ k_j &= \nu_j k_{j-1} + \kappa_j k_{j-2} + \mu_j \tau \bar{f}_\eta(k_{j-1}), \quad j = 2, \dots, s, \\ y_{n+1} &= k_s, \end{aligned} \tag{3.35}$$

where the parameters μ_j, ν_j, κ_j are defined in (1.20) and

$$\bar{f}_\eta(u_0) = \frac{1}{\eta}(u_\eta - u_0) \tag{3.36}$$

corresponds to the numerical counterpart of $f_\eta(u_0)$ in (3.4). The approximation u_η of $u(\eta)$ is computed at each evaluation of \bar{f}_η by applying one step, of size η , of the m -stage RKC scheme to (3.3). Hence, u_η is given by

$$\begin{aligned} u_1 &= u_0 + \alpha_1 \eta(f_F(u_0) + f_S(u_0)), \\ u_j &= \beta_j u_{j-1} + \gamma_j u_{j-2} + \alpha_j \eta(f_F(u_{j-1}) + f_S(u_0)), \quad j = 2, \dots, m, \\ u_\eta &= u_m. \end{aligned} \tag{3.37}$$

Here, the parameters $\alpha_j, \beta_j, \gamma_j$ of the m -stage RKC scheme (3.37) are given by

$$v_0 = 1 + \varepsilon/m^2, \quad v_1 = T_m(v_0)/T'_m(v_0), \quad a_j = 1/T_j(v_0) \quad \text{for } j = 0, \dots, m \tag{3.38}$$

and

$$\begin{aligned} \alpha_1 &= v_1/v_0, & \alpha_j &= 2v_1 a_j / a_{j-1}, \\ \beta_j &= 2v_0 a_j / a_{j-1}, & \gamma_j &= -a_j / a_{j-2}, \quad \text{for } j = 2, \dots, m. \end{aligned} \tag{3.39}$$

To compute m, η in (3.34), we let $\eta = 6\tau m^2 / (\beta s^2(m^2 - 1))$ in $\eta\rho_F \leq \beta m^2$, which implies

$$6\tau\rho_F \leq \beta^2 s^2(m^2 - 1). \tag{3.40}$$

Thus, we use (3.40) to compute m and then (3.34) to determine η .

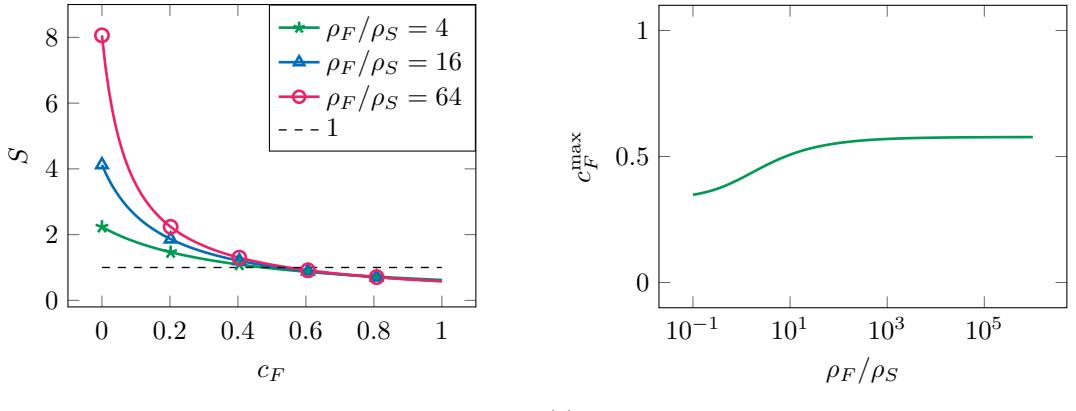
Stability and first-order accuracy of the mRKC scheme (3.34) to (3.37) are proved in Theorems 3.20 and 3.21 in Section 3.4.3 below.

3.4.2 Efficiency of the mRKC method

Given the spectral radii ρ_F and ρ_S of the Jacobians of f_F and f_S , respectively, we now evaluate the theoretical speed-up in using the mRKC method (3.34) to (3.37) over the standard RKC method of Section 1.2.4. In doing so, we set $\varepsilon = 0$ and let s, m vary in \mathbb{R} . We denote the cost of evaluating f_F and f_S relatively to the cost of evaluating f itself by c_F and c_S , respectively, with $c_F, c_S \in [0, 1]$ and $c_F + c_S = 1$. As we did in Section 3.2.2, we suppose that the spectral radius ρ of the Jacobian of f is $\rho = \rho_F + \rho_S$; as we consider a wide range of values for ρ_F we cannot set $\rho = \rho_F$ otherwise $\rho = 0$ for $f_F = 0$, instead of $\rho = \rho_S$.

Since the RKC scheme requires $s = \sqrt{\tau\rho/2}$ evaluations of f per time step, its cost per time step is

$$C_{\text{RKC}} = s(c_F + c_S) = \sqrt{\frac{\tau(\rho_F + \rho_S)}{2}}. \tag{3.41}$$



(a) Theoretical speed-up S of the mRKC method over the standard RKC scheme, with respect to c_F and ρ_F/ρ_S .
(b) Maximal c_F which still yields speed-up $S > 1$, w.r.t. ρ_F/ρ_S .

Figure 3.9. The relative speed-up S of the mRKC method over the RKC scheme with respect to c_F and the maximal value for c_F which still leads to an efficiency gain.

From (3.34) with $\beta = 2$, we infer that the mRKC scheme needs $s = \sqrt{\tau\rho_S/2}$ external stages and from (3.40) that the number of internal stages is $m = \sqrt{3\rho_F/\rho_S + 1}$. Since mRKC needs s evaluations of f_S and $s m$ evaluations of f_F , its cost per time step is

$$C_{\text{mRKC}} = s c_S + s m c_F = (1 - c_F) \sqrt{\frac{\tau\rho_S}{2}} + c_F \sqrt{\frac{3\tau\rho_F}{2} + \frac{\tau\rho_S}{2}}. \quad (3.42)$$

The ratio between (3.41) and (3.42) yields the relative speed-up

$$S = \frac{C_{\text{RKC}}}{C_{\text{mRKC}}} = \frac{\sqrt{\rho_F + \rho_S}}{(1 - c_F)\sqrt{\rho_S} + c_F\sqrt{\rho_S + 3\rho_F}} = \frac{\sqrt{1 + r_\rho}}{1 + c_F (\sqrt{1 + 3r_\rho} - 1)}, \quad (3.43)$$

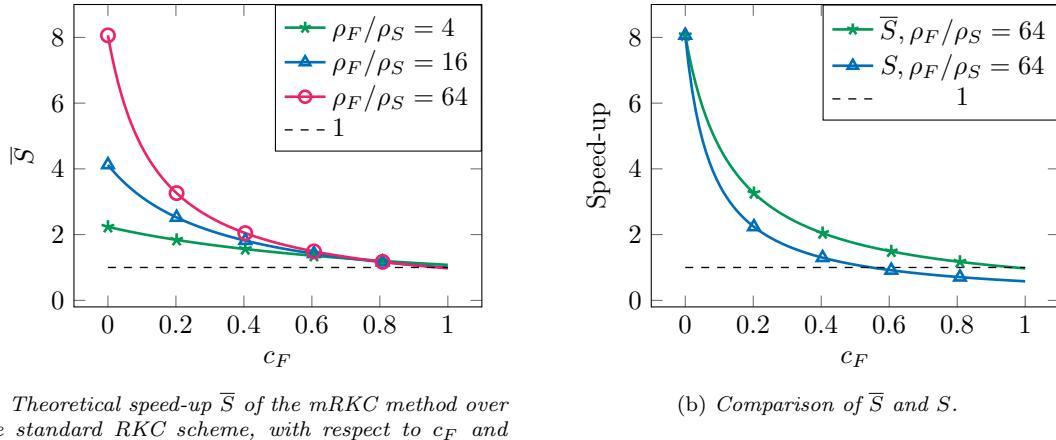
with $r_\rho = \rho_F/\rho_S \in [0, \infty)$. In Figure 3.9(a), we show the speed-up S as a function of c_F for different values of r_ρ . We observe that for c_F sufficiently small, the mRKC scheme is faster than RKC ($S > 1$). When $c_F \approx 1$, however, the mRKC scheme is slower than RKC ($S < 1$), though this case is somewhat irrelevant since by assumption f_F is cheap to evaluate. Nevertheless, we solve the inequality $S > 1$, with S as in (3.43), for varying c_F to determine the maximal value of c_F that still leads to a reduced cost in using mRKC. We find that $S > 1$ if, and only if,

$$c_F < c_F^{\max} = \frac{\sqrt{1 + r_\rho} - 1}{\sqrt{1 + 3r_\rho} - 1}.$$

In Figure 3.9(b), we monitor c_F^{\max} as a function of r_ρ . For small $r_\rho = \rho_F/\rho_S$ we observe that the evaluation of f_F must be quite cheap. As ρ_F/ρ_S increases, however, the mRKC method is faster than RKC, even if f_F is relatively expensive to evaluate ($c_F^{\max} > 0.5$ for $\rho_F/\rho_S > 8$).

Efficiency of the mRKC scheme for problems with well separated scales or spatially discretized parabolic problems

The stability conditions (3.34) are necessary when solving a general problem (3.1). However, we will see in Sections 3.4.5 and 3.5.4 that for problems with well separated scales and for systems stemming from the spatial discretization of parabolic equations, conditions (3.34) can in fact be



(a) Theoretical speed-up \bar{S} of the mRKC method over the standard RKC scheme, with respect to c_F and ρ_F/ρ_S .

(b) Comparison of \bar{S} and S .

Figure 3.10. The relative speed-up \bar{S} obtained using (3.44) compared to S , obtained with (3.34).

replaced by the weaker ones

$$\tau\rho_S \leq \beta s^2, \quad \eta\rho_F \leq \bar{\beta}m^2 \quad \text{with} \quad \eta = \frac{2\tau}{\beta s^2}, \quad (3.44)$$

$\bar{\beta} = 2 - 4\bar{\varepsilon}/3 \approx 1.86$ and $\bar{\varepsilon} = 0.1$. Since the value for η in (3.44) is smaller than that in (3.34), m can also be smaller, which results in fewer evaluations of f_F in (3.37) and improved efficiency. Hence, let us estimate the efficiency of the mRKC scheme under conditions (3.44) instead of (3.34).

Taking $\beta = 2$, the number of external stages is again $s = \sqrt{\tau\rho_S/2}$. On the other hand, $\bar{\beta}$ cannot be taken without damping and the internal stages are given by

$$m = \sqrt{\frac{\eta\rho_F}{\bar{\beta}}} = \sqrt{\frac{2\rho_F}{\bar{\beta}\rho_S}},$$

hence the cost of the mRKC method under conditions (3.44) is

$$\bar{C}_{\text{mRKC}} = s c_S + s m c_F = (1 - c_F) \sqrt{\frac{\tau\rho_S}{2}} + c_F \sqrt{\frac{\tau\rho_F}{\bar{\beta}}} \quad (3.45)$$

and the relative speed-up is given by

$$\bar{S} = \frac{C_{\text{RKC}}}{\bar{C}_{\text{mRKC}}} = \frac{\sqrt{\rho_F + \rho_S}}{(1 - c_F)\sqrt{\rho_S} + c_F \sqrt{\frac{2\rho_F}{\bar{\beta}}}} = \frac{\sqrt{1 + r_\rho}}{1 + c_F \left(\sqrt{\frac{2}{\bar{\beta}}} r_\rho - 1 \right)}, \quad (3.46)$$

where $2/\bar{\beta} \approx 1.07$. Note that if in (3.46) we had 1 instead of $2/\bar{\beta}$ then we would have $\bar{S} > 1$ for all $c_F \in [0, 1]$.

In Figure 3.10(a), we plot \bar{S} as a function of c_F for different values of $r_\rho = \rho_F/\rho_S$, as in Figure 3.9(a) for S . We observe that $\bar{S} > 1$ for all $c_F \in [0, 1 - \epsilon]$, for $\epsilon > 0$ very small. In Figure 3.10(b), we compare S and \bar{S} and observe that $\bar{S} > S$ for all values of c_F .

Observe that if in (3.45) we had 2 instead of $\bar{\beta} \approx 1.86$ then \bar{C}_{mRKC} would be the cost of solving an uncoupled system of equations of the type $y' = f_F(y)$, $z' = f_S(z)$ using two RKC methods, one for each equation. Hence, since $\bar{\beta}$ is very close to 2, the cost added by the coupling is very low.

3.4.3 Stability and convergence analysis

In this section, we perform the stability and convergence analysis of the multirate RKC method introduced in Section 3.4.1. First, we prove that the mRKC method is stable when it is applied to the multirate test equation (3.12), which is a good model for problems where the Jacobians of f_F and f_S are simultaneously triangularizable. Then, we also show stability for a 2×2 model problem (the same as in Section 2.2.1) where the Jacobians of f_F and f_S are not simultaneously triangularizable and hence the stability analysis cannot be reduced to (3.12). Finally, we prove its first-order accuracy.

Stability analysis on the multirate test equation

Since (3.3) is approximated numerically, the stability analysis performed in Section 3.1 or in Section 3.3 is no longer valid; indeed, $\varphi(z)$ is now replaced by a numerical approximation with different stability properties, as for the mEE method in Section 3.2.

Hence, we now compute a closed expression for u_η given u_0 , as in (3.13) for $u(\eta)$. We denote by

$$P_m(z) = a_m T_m(v_0 + v_1 z) \quad (3.47)$$

the stability polynomial of the m -stage RKC scheme, with a_m, v_0, v_1 from (3.38). The next lemma computes the solution u_η of (3.37) in the case of the multirate test equation (3.12).

Lemma 3.16. *Let $\lambda, \zeta \leq 0$, $f_S(y) = \zeta y$, $f_F(y) = \lambda y$, $\eta > 0$, $m \in \mathbb{N}$ and $u_0 \in \mathbb{R}$. Then, the solution u_η of (3.37), is given by*

$$u_\eta = (P_m(\eta\lambda) + \Phi_m(\eta\lambda)\eta\zeta)u_0, \quad (3.48)$$

where $P_m(z)$ is given in (3.47),

$$\Phi_m(z) = \frac{P_m(z) - 1}{z} \quad \text{for } z \neq 0 \quad (3.49)$$

and $\Phi_m(0) = 1$ is defined by continuous extension.

Proof. Plugging $f_S(y) = \zeta y$ and $f_F(y) = \lambda y$ into (3.37) and letting $r_j = \alpha_j \eta \zeta u_0$, we obtain

$$\begin{aligned} u_1 &= u_0 + \alpha_1 \eta \lambda u_0 + r_1, \\ u_j &= \beta_j u_{j-1} + \gamma_j u_{j-2} + \alpha_j \eta \lambda u_{j-1} + r_j, \quad j = 2, \dots, m. \end{aligned} \quad (3.50)$$

We note that (3.50) corresponds to an RKC scheme where each stage is perturbed by r_j , as (1.24). From (1.25) follows

$$u_j = a_j T_j(v_0 + v_1 \eta \lambda) u_0 + \sum_{k=1}^j \frac{a_j}{a_k} U_{j-k}(v_0 + v_1 \eta \lambda) r_k,$$

with a_j, v_0, v_1 as in (3.38) and $U_j(x)$ is the Chebyshev polynomial of the second kind of degree j defined in (1.26). Thus, as $u_\eta = u_m$ and $r_k = \alpha_k \eta \zeta u_0$,

$$u_\eta = P_m(\eta\lambda)u_0 + \sum_{k=1}^m \frac{a_m}{a_k} \alpha_k U_{m-k}(v_0 + v_1 \eta \lambda) \eta \zeta u_0. \quad (3.51)$$

In Lemma 3.17 below we show the identity

$$\Phi_m(\eta\lambda) = \sum_{k=1}^m \frac{a_m}{a_k} \alpha_k U_{m-k}(v_0 + v_1 \eta \lambda), \quad (3.52)$$

which, in combination with (3.51) implies (3.48). ■

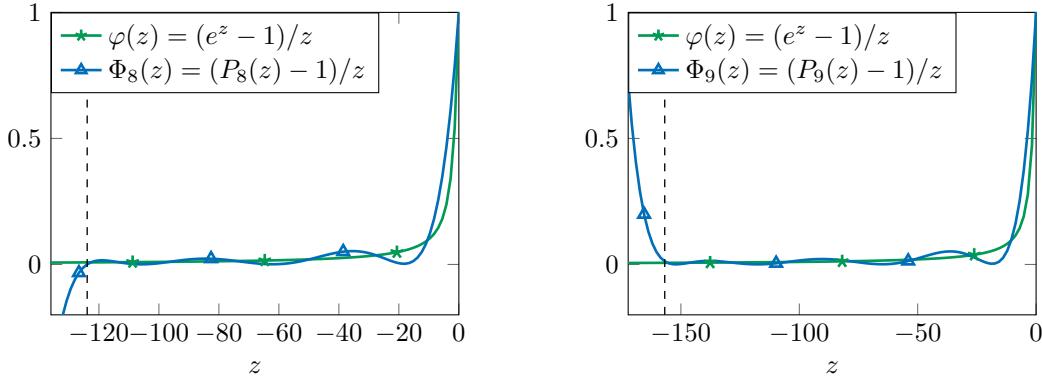


Figure 3.11. Illustration of $\varphi(z)$ and $\Phi_m(z)$ for $m = 8$ (left) and $m = 9$ (right). The dashed line indicates the end of the stability domain.

Note the similarity between (3.13) and (3.48), with $e^z, \varphi(z)$ replaced by $P_m(z), \Phi_m(z)$, respectively. In Figure 3.11, we also observe that $\Phi_m(z)$ and $\varphi(z)$ share similar stability properties. Indeed, $\Phi_m(z)$ is the numerical counterpart of $\varphi(z)$ as it can be written as $\varphi(z)$ yet with the exponential replaced by the stability polynomial, compare (3.7) and (3.49). The next lemma shows the identity (3.52), the proof is rather technical and therefore postponed to Section 3.6.2.

Lemma 3.17. Let $\Phi_m(z)$ be as in (3.49), then

$$\Phi_m(z) = \sum_{k=1}^m \frac{a_m}{a_k} \alpha_k U_{m-k}(v_0 + v_1 z). \quad (3.53)$$

Thanks to (3.49), we can also compute the stability polynomial of the mRKC scheme. From (3.36) and (3.48), we get

$$\bar{f}_\eta(u_0) = \frac{1}{\eta} (P_m(\eta\lambda) + \Phi_m(\eta\lambda)\eta\zeta - 1)u_0 = \Phi_m(\eta\lambda)(\lambda + \zeta)u_0. \quad (3.54)$$

Now, we introduce (3.54) into (3.35), which leads to

$$y_{n+1} = R_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))y_n, \quad (3.55)$$

where $R_s(z)$ is the stability polynomial of the s -stage RKC scheme, given in Section 1.1. Relation (3.55) motivates the following definition.

Definition 3.18. Let $s, m \in \mathbb{N}$, $\tau > 0$ be a step size, $\eta > 0$ and $\lambda, \zeta \leq 0$. The stability polynomial of the (s, m) -stage mRKC scheme (3.35) to (3.37) is defined as

$$R_{s,m}(\lambda, \zeta, \tau, \eta) = R_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta)).$$

The following lemma is the discrete version of Lemma 3.6 and is needed to prove stability of the mRKC scheme in Theorem 3.20 below. Its proof is purely technical and also postponed to Section 3.6.2.

Lemma 3.19. Let $m \in \mathbb{N}$ and $w \leq 0$. There is $\bar{\varepsilon}_m > 0$ such that for $\varepsilon \leq \bar{\varepsilon}_m$ it holds $\Phi_m(z)(z + w) \in [w, 0]$ for all $z \in [-\ell_m^\varepsilon, 0]$ if, and only if, $\Phi'_m(0)|w| \geq 1$. As $\Phi'_m(0) = P''_m(0)/2$, then $|w| \geq 2/P''_m(0)$.

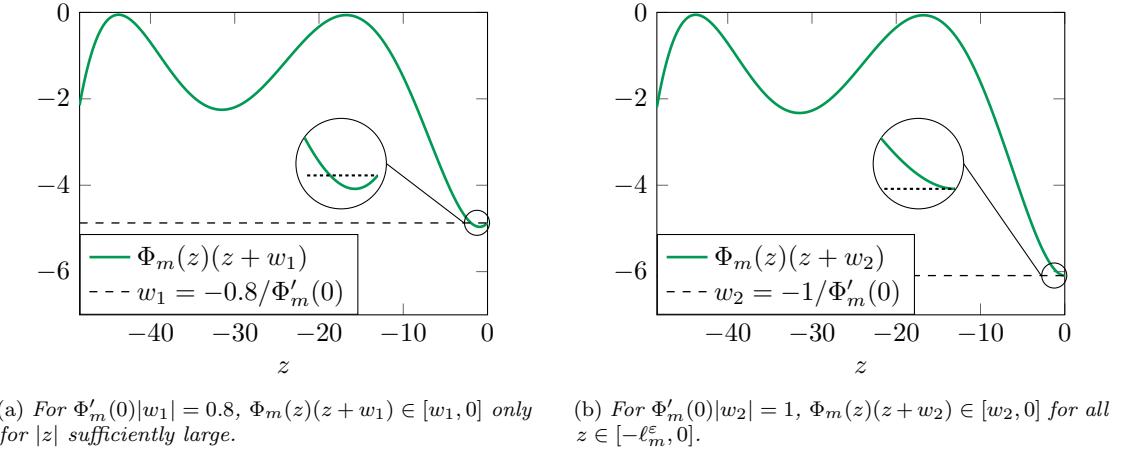


Figure 3.12. Illustration of the condition $\Phi_m(z)(z + w) \in [w, 0]$ for all $z \in [-\ell_m^\varepsilon, 0]$, with $m = 5$, $\varepsilon = 0.05$ and two different values for w .

For $\varepsilon = 0$, it holds $2/P''_m(0) = 6m^2/(m^2 - 1) > 6$. In the continuous setting, the condition on w in Lemma 3.6 was $|w| \geq 2$. For the discrete mRKC scheme, however, $|w| > 6$ is necessary because of the milder slope of $\Phi_m(z)$ at the origin, see Figure 3.11. We illustrate in Figure 3.12 the condition $\Phi_m(z)(z + w) \in [w, 0]$ for all $z \in [-\ell_m^\varepsilon, 0]$, for $m = 5$, $\varepsilon = 0.05$ and considering two different values of w . We see in Figure 3.12(a) that if $\Phi'_m(0)|w| < 1$ then $\Phi_m(z)(z + w) \in [w, 0]$ holds only for z away from zero. In contrast, for $\Phi'_m(0)|w| \geq 1$, as in Figure 3.12(b), we note that $\Phi_m(z)(z + w) \in [w, 0]$ for all z in the stability domain.

Theorem 3.20. Let $\bar{\varepsilon}_m$ be as in Lemma 3.19 and, for $\varepsilon \geq 0$, let $\varepsilon_m = \min\{\varepsilon, \bar{\varepsilon}_m\}$. Let $\lambda \leq 0$ and $\zeta < 0$. Then, for all $\tau > 0$, s, m and η such that

$$\tau|\zeta| \leq \ell_s^\varepsilon, \quad \eta|\lambda| \leq \ell_m^{\varepsilon_m} \quad \text{with} \quad \eta \geq \frac{6\tau}{\ell_s^\varepsilon} \frac{m^2}{m^2 - 1}, \quad (3.56)$$

$|R_{s,m}(\lambda, \zeta, \tau, \eta)| \leq 1$, i.e. the mRKC scheme is stable.

Proof. If $\tau\Phi_m(\eta\lambda)(\lambda + \zeta) \in [-\ell_s^\varepsilon, 0]$ then $|R_{s,m}(\lambda, \zeta, \tau, \eta)| = |R_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))| \leq 1$. Hence, it is sufficient to prove the equivalent condition:

$$\Phi_m(\eta\lambda)(\eta\lambda + \eta\zeta) \in [w(\eta), 0], \quad \text{with} \quad w(\eta) = -\frac{\eta}{\tau} \ell_s^\varepsilon.$$

Since $\eta\lambda \in [-\ell_m^{\varepsilon_m}, 0]$, it holds $|P_m(\eta\lambda)| \leq 1$ and from (3.49) we thus deduce that $\Phi_m(\eta\lambda) \geq 0$. Furthermore, (3.56) yields $\eta\zeta \geq w(\eta)$ which implies

$$0 \geq \Phi_m(\eta\lambda)(\eta\lambda + \eta\zeta) \geq \Phi_m(z(\eta))(z(\eta) + w(\eta)),$$

with $z(\eta) = \eta\lambda$. Hence, it is sufficient to show that $\Phi_m(z(\eta))(z(\eta) + w(\eta)) \geq w(\eta)$ for all $z(\eta) \in [-\ell_m^{\varepsilon_m}, 0]$. From Lemma 3.19, we know that

$$|w(\eta)| \geq \frac{2}{P''_m(0)}$$

is necessary and sufficient. In Lemma 3.23 in Section 3.6.2 we show that $T'_m(v_0)^2/(T_m(v_0)T''_m(v_0))$ is decreasing for $v_0 \geq 1$, we therefore infer from the definition of η in (3.56) that

$$|w(\eta)| \geq 6 \frac{m^2}{m^2 - 1} = \frac{2 T'_m(1)^2}{T_m(1)T''_m(1)} \geq \frac{2 T'_m(v_0)^2}{T_m(v_0)T''_m(v_0)} = \frac{2}{P''_m(0)}. \quad \blacksquare$$

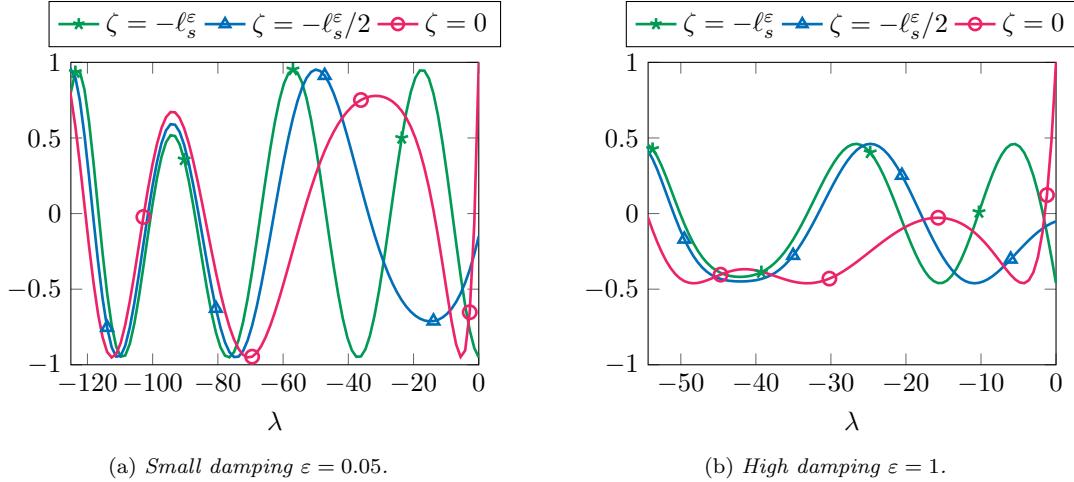


Figure 3.13. Stability polynomial $R_{s,m}(\lambda, \zeta, \tau, \eta)$ of the mRKC method vs. λ for $\zeta = -\ell_s^\varepsilon, -\ell_s^\varepsilon/2$ or 0 and $s = 5$, $m = 3$, $\tau = 1$, η as in (3.56) and damping $\varepsilon = 0.05$ (left) or $\varepsilon = 1$ (right).

In the continuous setting in Section 3.1, η was directly dependent on f_S ; indeed we required $|\varphi(\eta\lambda)(\lambda + \zeta)| \leq |\zeta|$, which implies $\eta \geq 2/|\zeta|$ (see Theorem 3.7). Therefore, η could rapidly grow as $\zeta \rightarrow 0$. In contrast, we see in (3.56) that for the mRKC method η remains bounded as ζ decreases, indeed $\ell_s^\varepsilon \geq 2$ for all $s \in \mathbb{N}$. The reason for this is that for stability of the mRKC scheme it is sufficient that $\tau\Phi_m(\eta\lambda)(\lambda + \zeta) \in [-\ell_s^\varepsilon, 0]$, thus that the spectrum of $\tau\bar{f}_\eta$ falls inside the stability domain of the s -stage RKC method. Hence, as in Section 3.3, η depends only indirectly on the slow term f_S : η depends on the s -stage RKC method, which in turn depends on f_S . This indirect dependence of η on f_S creates a protective ‘‘buffer’’ which prevents the explosion of η as $\zeta \rightarrow 0$.

The restriction $\varepsilon \leq \bar{\varepsilon}_m$ is necessary for proving Lemma 3.19, but probably not needed in practice. Indeed, we have verified numerically that for any $\varepsilon \geq 0$, $\Phi_m(z)(z + w) \in [w, 0]$ for all $z \in [-\ell_m^\varepsilon, 0]$ if, and only if, $|w| \geq 2/P_m''(0)$. Hence, we can suppose $\varepsilon_m = \varepsilon$ in (3.56) and replace $\ell_s^\varepsilon, \ell_m^\varepsilon$ by $\beta s^2, \beta m^2$, respectively, which yields (3.34). In Figure 3.13, we display the stability polynomial $R_{s,m}(\lambda, \zeta, \tau, \eta)$ for $s = 5$ and $m = 3$ as a function of λ for $\varepsilon = 0.05$ or $\varepsilon = 1$. Here, we set $\tau = 1$, η to its lower bound in (3.56), and $\zeta = -\ell_s^\varepsilon, -\ell_s^\varepsilon/2$ or 0. Since $|R_{s,m}(\lambda, \zeta, \tau, \eta)| \leq 1$, the mRKC method is always stable.

Stability analysis on a 2×2 model

Here, we consider a 2×2 linear problem where the Jacobians of f_F and f_S are not simultaneously triangularizable and therefore cannot be reduced to the scalar multirate test equation (3.12); we will show that the same stability conditions nonetheless hold. Moreover, we introduce a coupling term between the fast and slow variables and show that the same stability conditions are necessary even when the coupling is weak.

Thus, we consider the system of differential equations

$$y' = Ay, \quad \text{with} \quad A = \begin{pmatrix} \zeta & \sigma \\ \sigma & \lambda \end{pmatrix} \quad (3.57)$$

and $y(0) = y_0 \in \mathbb{R}^2$. We let $\lambda, \zeta < 0$, $\sigma \in \mathbb{R}$ the coupling term, and assume that $\sigma^2 \leq \lambda\zeta$, to ensure that both eigenvalues of A are negative or zero. We note $D \in \mathbb{R}^{2 \times 2}$ the diagonal

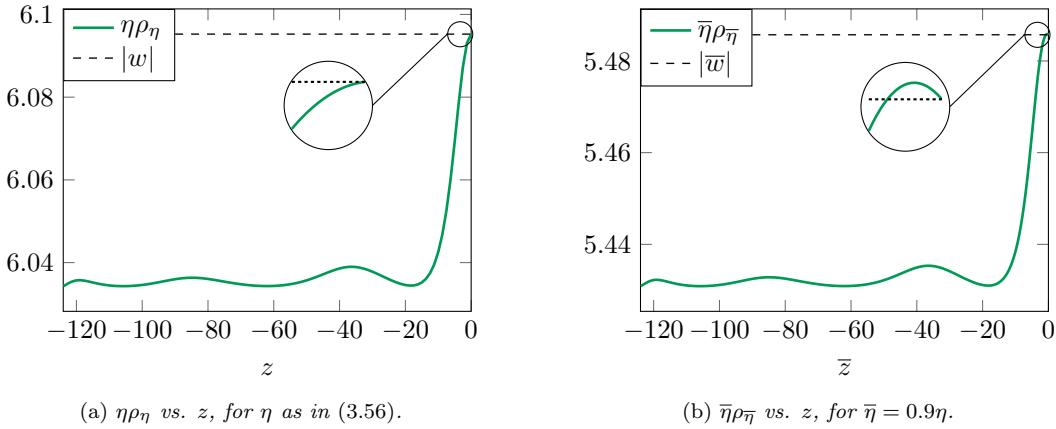


Figure 3.14. Showing that stability conditions (3.56) are sufficient and necessary when mRKC is applied to systems of equations as (3.57).

matrix satisfying $D_{11} = 0$ and $D_{22} = 1$ and consider the splitting defined by $f_F(y) = A_F y$ and $f_S(y) = A_S y$, where

$$A_F := DA = \begin{pmatrix} 0 & 0 \\ \sigma & \lambda \end{pmatrix}, \quad A_S := (I - D)A = \begin{pmatrix} \zeta & \sigma \\ 0 & 0 \end{pmatrix}. \quad (3.58)$$

Observe that $\rho_F = |\lambda|$ and $\rho_S = |\zeta|$. The matrices A_F, A_S are simultaneously triangularizable if, and only if, they have a common eigenvector. This happens if $\sigma = 0$ or if $\sigma^2 = \lambda\zeta$. We will choose $\sigma = 0.1\sqrt{\lambda\zeta}$, so that the present stability analysis cannot be reduced to the previous multirate test equation. Furthermore, as the eigenvalues of A are negative or zero for all $|\sigma| \leq \sqrt{\lambda\zeta}$, the current coupling $\sigma = 0.1\sqrt{\lambda\zeta}$ can be considered to be weak when compared to the maximal coupling $\sqrt{\lambda\zeta}$.

Given $u_0 \in \mathbb{R}^2$, we obtain $\bar{f}_\eta(u_0)$ by replacing λ, ζ in (3.54) by A_F, A_S , respectively. This yields

$$\bar{f}_\eta(u_0) = A_\eta u_0, \quad \text{with} \quad A_\eta = \Phi_m(\eta A_F) A u_0, \quad (3.59)$$

and since Φ_m is a polynomial, A_η is well-defined. From (3.55) it follows $y_{n+1} = R_s(\tau A_\eta) y_n$. If the eigenvalues of τA_η are in the interval $[-\ell_s^\varepsilon, 0]$, then the mRKC method is stable. For convenience, we set $\tau = 1$, $|\zeta| = \ell_s^\varepsilon$ with $s = 10$ and also fix $m = 8$ and $\eta = \frac{6\tau}{\ell_s^\varepsilon} \frac{m^2}{m^2 - 1}$ (as in (3.56)). Then, the mRKC method is stable if the spectral radius ρ_η of A_η satisfies $\rho_\eta \leq |\zeta|$ for all $\eta\lambda \in [-\ell_m^\varepsilon, 0]$, or equivalently $\eta\rho_\eta \leq \eta|\zeta| = |w|$.

In Figure 3.14(a), we display $\eta\rho_\eta$ as a function of $z = \eta\lambda \in [-\ell_m^\varepsilon, 0]$ and observe that $\eta\rho_\eta \leq |w|$; thus, the mRKC scheme is stable. Hence, the stability conditions (3.34) guarantee stability of the scheme even though the Jacobians of f_F, f_S are not simultaneously triangularizable.

In Figure 3.14(b), we consider a value for η smaller than that dictated by (3.56). For $\bar{\eta} = 0.9\eta$, we again display $\bar{\eta}\rho_{\bar{\eta}}$ as a function of $\bar{z} = \bar{\eta}\lambda \in [-\ell_m^\varepsilon, 0]$. Then, a small region of instability appears for \bar{z} close to zero, where $\bar{\eta}\rho_{\bar{\eta}} > |\bar{w}|$. Hence, conditions (3.56) are necessary even for systems of equations with a weak coupling $\sigma = 0.1\sqrt{\lambda\zeta}$, where $\sqrt{\lambda\zeta}$ corresponds to the maximal coupling strength.

Convergence analysis

We end this section by proving that the mRKC scheme is first-order accurate. Although we only consider the test equation here, this is sufficient to prove first-order accuracy for general nonlinear problems [69].

Theorem 3.21. *The mRKC scheme is first-order accurate.*

Proof. Let $y_1(\tau, \eta) = R_{s,m}(\lambda, \zeta, \tau, \eta)y_0 = R_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))y_0$, hence

$$\nabla y_1(\tau, \eta) = \begin{pmatrix} R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\Phi_m(\eta\lambda)(\lambda + \zeta) \\ R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\tau\lambda\Phi'_m(\eta\lambda)(\lambda + \zeta) \end{pmatrix} y_0 \quad \text{and} \quad \nabla y_1(0, 0) = \begin{pmatrix} \lambda + \zeta \\ 0 \end{pmatrix} y_0.$$

By Taylor expansion of $y_1(\tau, \eta)$ in $(\tau, \eta) = (0, 0)$, we obtain

$$y_1(\tau, \eta) = y_1(0, 0) + \nabla y_1(0, 0) \cdot (\tau, \eta) + \mathcal{O}(\tau^2 + \eta^2) = y_0 + \tau(\lambda + \zeta)y_0 + \mathcal{O}(\tau^2 + \eta^2).$$

From the definition of m and η , we deduce that $m \geq 2$ (see (3.40)) and hence $\eta \leq 8\tau/(\beta s^2)$. It follows $y_1(\tau, \eta) - y(\tau) = \mathcal{O}(\tau^2)$, which implies first-order accuracy. ■

Typically $s \gg 1$, i.e. $\eta \ll \tau$, and the error made when approximating f by the averaged force f_η is negligible. In fact, we observe that the difference between the RKC and the mRKC solutions in our numerical experiments in Section 3.5 is always very small.

3.4.4 A step size control strategy

In this section, we first derive an error estimator for the RKC and mRKC scheme and then we recall the PID step size control strategy already introduced in [66], see also [69, 109]. For the standard RKC scheme the error estimator introduced here is asymptotically exact. In contrast, for the mRKC method one of the error terms is not estimated; however, because of the small size of η with respect to τ the remaining term is usually negligible. Unfortunately, for the multirate test equation (3.12) the error estimator for the RKC scheme cannot be bounded by a stiffness independent constant but grows linearly with the fast term λ , in contrast the error estimator for the mRKC scheme grows linearly with the slow term ζ ; which still is undesirable. We will see in Section 3.5.5 that despite these two issues the error control strategy is able to efficiently estimate the error and adapt the step size.

Error estimator

Here we derive an error estimator for the mRKC and RKC scheme. We first consider the mRKC method applied to the multirate test equation (3.12), then we extend the estimator to general equations (3.1) and to the RKC method.

Let $e_{n+1}(\lambda, \zeta, \tau)$ be the local error committed at time $t_{n+1} = t_n + \tau$ by the mRKC method applied to the multirate test equation (3.12); hence, denoting y_n the solution at time t_n , it holds

$$e_{n+1}(\lambda, \zeta, \tau) = R_{s,m}(\lambda, \zeta, \tau, \eta)y_n - e^{(\lambda+\zeta)\tau}y_n, \quad (3.60)$$

with s, m, η given by (3.56). The next lemma describes the asymptotic behavior of (3.60).

Lemma 3.22. *The local error (3.60) satisfies $|e_{n+1}(\lambda, \zeta, \tau)| \leq 2|y_n|$ and*

$$e_{n+1}(\lambda, \zeta, \tau) = \frac{1}{2}(R''_s(0) - 1)\tau^2(\lambda + \zeta)^2y_n + \frac{1}{2}P''_m(0)\tau\eta\lambda(\lambda + \zeta)y_n + \mathcal{O}(\tau^3(\lambda + \zeta)^3) \quad (3.61)$$

as $\tau(\lambda + \zeta) \rightarrow 0$.

Proof. The first result is obvious since $|R_{s,m}(\lambda, \zeta, \tau, \eta)| \leq 1$ and $|e^{(\lambda+\zeta)\tau}| \leq 1$. For the second, we compute the derivatives of

$$R_{s,m}(\lambda, \zeta, \tau, \eta) = R_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta)),$$

obtaining

$$\begin{aligned} \partial_\tau R_{s,m}(\lambda, \zeta, \tau, \eta) &= R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\Phi_m(\eta\lambda)(\lambda + \zeta), \\ \partial_\eta R_{s,m}(\lambda, \zeta, \tau, \eta) &= R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\tau\Phi'_m(\eta\lambda)\lambda(\lambda + \zeta), \\ \partial_\tau^2 R_{s,m}(\lambda, \zeta, \tau, \eta) &= R''_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))(\Phi_m(\eta\lambda)(\lambda + \zeta))^2, \\ \partial_\eta^2 R_{s,m}(\lambda, \zeta, \tau, \eta) &= R''_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))(\tau\Phi'_m(\eta\lambda)\lambda(\lambda + \zeta))^2 \\ &\quad + R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\tau\Phi''_m(\eta\lambda)\lambda^2(\lambda + \zeta), \\ \partial_{\tau\eta} R_{s,m}(\lambda, \zeta, \tau, \eta) &= R''_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\tau\Phi_m(\eta\lambda)\Phi'_m(\eta\lambda)\lambda(\lambda + \zeta)^2 \\ &\quad + R'_s(\tau\Phi_m(\eta\lambda)(\lambda + \zeta))\Phi'_m(\eta\lambda)\lambda(\lambda + \zeta). \end{aligned}$$

Thus, the gradient and the Hessian of $R_{s,m}(\lambda, \zeta, \tau, \eta)$ in $(\tau, \eta) = (0, 0)$ are

$$\nabla_{\tau,\eta} R_{s,m}(\lambda, \zeta, 0, 0) = \begin{pmatrix} \lambda + \zeta \\ 0 \end{pmatrix}, \quad H_{\tau,\eta} R_{s,m}(\lambda, \zeta, 0, 0) = \begin{pmatrix} R''_s(0)(\lambda + \zeta)^2 & \Phi'_m(0)\lambda(\lambda + \zeta) \\ \Phi'_m(0)\lambda(\lambda + \zeta) & 0 \end{pmatrix},$$

respectively, and therefore Taylor expanding $R_{s,m}(\lambda, \zeta, \tau, \eta)$ yields

$$R_{s,m}(\lambda, \zeta, \tau, \eta) = 1 + \tau(\lambda + \zeta) + \frac{1}{2}R''_s(0)\tau^2(\lambda + \zeta)^2 + \Phi'_m(0)\tau\eta\lambda(\lambda + \zeta) + \mathcal{O}(\tau^3(\lambda + \zeta)^3).$$

Using

$$\begin{aligned} \Phi'_m(z) &= \frac{P'_m(z)z - (P_m(z) - 1)}{z^2} \\ &= \frac{(P'_m(0) + P''_m(0)z)z - P'_m(0)z - \frac{1}{2}P''_m(0)z^2 + \mathcal{O}(z^3)}{z^2} = \frac{1}{2}P''_m(0) + \mathcal{O}(z) \end{aligned} \tag{3.62}$$

we have, taking the limit, $\Phi'_m(0) = \frac{1}{2}P''_m(0)$ and it follows

$$R_{s,m}(\lambda, \zeta, \tau, \eta) = 1 + \tau(\lambda + \zeta) + \frac{1}{2}R''_s(0)\tau^2(\lambda + \zeta)^2 + \frac{1}{2}P''_m(0)\tau\eta\lambda(\lambda + \zeta) + \mathcal{O}(\tau^3(\lambda + \zeta)^3).$$

Since

$$e^{(\lambda+\zeta)\tau} = 1 + \tau(\lambda + \zeta) + \frac{1}{2}\tau^2(\lambda + \zeta)^2 + \mathcal{O}(\tau^3(\lambda + \zeta)^3),$$

(3.61) follows from (3.60). ■

Now, we search for a computable error estimator $\bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau)$ having the same asymptotic behavior of $e_{n+1}(\lambda, \zeta, \tau)$, i.e.

$$\bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau) = e_{n+1}(\lambda, \zeta, \tau) + \mathcal{O}(\tau^3(\lambda + \zeta)^3), \tag{3.63}$$

and which $|\bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau)| \leq C|y_n|$, with C a stiffness independent constant.

Our strategy is to use a linear combination of the internal stages, hence

$$\bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau) = \sum_{j=1}^s r_j R_{j,m}(\lambda, \zeta, \tau, \eta) y_n, \tag{3.64}$$

where $r_j \in \mathbb{R}$ and $R_{j,m}(\lambda, \zeta, \tau, \eta)y_n$ are the internal stages of the mRKC scheme, i.e.

$$R_{j,m}(\lambda, \zeta, \tau, \eta) = R_j(\tau\Phi_m(\eta\lambda)(\lambda + \zeta)), \quad (3.65)$$

with $R_j(z)$ for $j = 0, \dots, s$ the internal stability polynomial of the RKC scheme, defined in (1.18). Taylor expanding $R_{j,m}(\lambda, \zeta, \tau, \eta)$ we find

$$\begin{aligned} R_{j,m}(\lambda, \zeta, \tau, \eta) &= 1 + \tau(\lambda + \zeta)R'_j(0) + \frac{1}{2}\tau\eta\lambda(\lambda + \zeta)P''_m(0)R'_j(0) \\ &\quad + \frac{1}{2}\tau^2(\lambda + \zeta)^2R''_j(0) + \mathcal{O}(\tau^3(\lambda + \zeta)^3), \end{aligned}$$

and thus, from (3.64),

$$\begin{aligned} \bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau) &= y_n \sum_{j=1}^s r_j + \tau(\lambda + \zeta)y_n \sum_{j=1}^s r_j R'_j(0) + \frac{1}{2}\tau\eta\lambda(\lambda + \zeta)P''_m(0)y_n \sum_{j=1}^s r_j R'_j(0) \\ &\quad + \frac{1}{2}\tau^2(\lambda + \zeta)^2y_n \sum_{j=1}^s r_j R''_j(0) + \mathcal{O}(\tau^3(\lambda + \zeta)^3), \end{aligned} \quad (3.66)$$

which implies that (3.63) is not solvable. Indeed, the terms $\tau(\lambda + \zeta)$ and $\frac{1}{2}\tau\eta\lambda(\lambda + \zeta)P''_m(0)$ appearing in (3.66) have the same weight $\sum_{j=1}^s r_j R'_j(0)$, but only one of them appears in the right hand side of (3.63), see (3.61). Therefore, we solve

$$\bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau) = \frac{1}{2}\tau^2(R''_s(0) - 1)(\lambda + \zeta)^2y_n + \mathcal{O}(\tau^3(\lambda + \zeta)^3), \quad (3.67)$$

which is equivalent to (3.63) with $\eta = 0$. Observe that, usually, $\eta \ll \tau$ and thus the missing term is of minor importance.

Choosing $r_j = 0$ for $j < s - 2$ a solution to (3.67) is given by

$$\begin{aligned} \bar{r} &= \frac{R''_s(0) - 1}{(R'_{s-2}(0) - 1)(R''_{s-1}(0) - R''_s(0)) - (R'_{s-1}(0) - 1)(R''_{s-2}(0) - R''_s(0))}, \\ r_{s-2} &= \bar{r}(1 - R'_{s-1}(0)), \\ r_{s-1} &= \bar{r}(R'_{s-2}(0) - 1), \\ r_s &= \bar{r}(R'_{s-1}(0) - R'_{s-2}(0)), \end{aligned}$$

thus we define the error estimator as

$$\begin{aligned} \bar{e}_{n+1}^{\text{mRKC}}(\lambda, \zeta, \tau) &= r_{s-2}R_{s-2,m}(\lambda, \zeta, \tau, \eta)y_n + r_{s-1}R_{s-1,m}(\lambda, \zeta, \tau, \eta)y_n \\ &\quad + r_sR_{s,m}(\lambda, \zeta, \tau, \eta)y_n. \end{aligned} \quad (3.68)$$

For general equations (3.1), we simply identify $R_{j,m}(\lambda, \zeta, \tau, \eta)y_n$ with the stage k_j and define the error estimator as

$$\bar{e}_{n+1}(\tau) = r_{s-2}k_{s-2} + r_{s-1}k_{s-1} + r_sk_s. \quad (3.69)$$

Next, we observe that the standard RKC scheme is equivalent to the mRKC method with $m = 1$, indeed $\Phi_1(z) = 1$ and thus, from (3.65),

$$R_{j,1}(\lambda, \zeta, \tau, \eta) = R_j(\tau(\lambda + \zeta)).$$

Hence, for the test equation (3.12) the error estimator for the RKC scheme is given by (3.68) with $m = 1$, i.e.

$$\begin{aligned} \bar{e}_{n+1}^{\text{RKC}}(\tau(\lambda + \zeta)) &= r_{s-2}R_{s-2}(\tau(\lambda + \zeta))y_n + r_{s-1}R_{s-1}(\tau(\lambda + \zeta))y_n \\ &\quad + r_sR_s(\tau(\lambda + \zeta))y_n. \end{aligned} \quad (3.70)$$

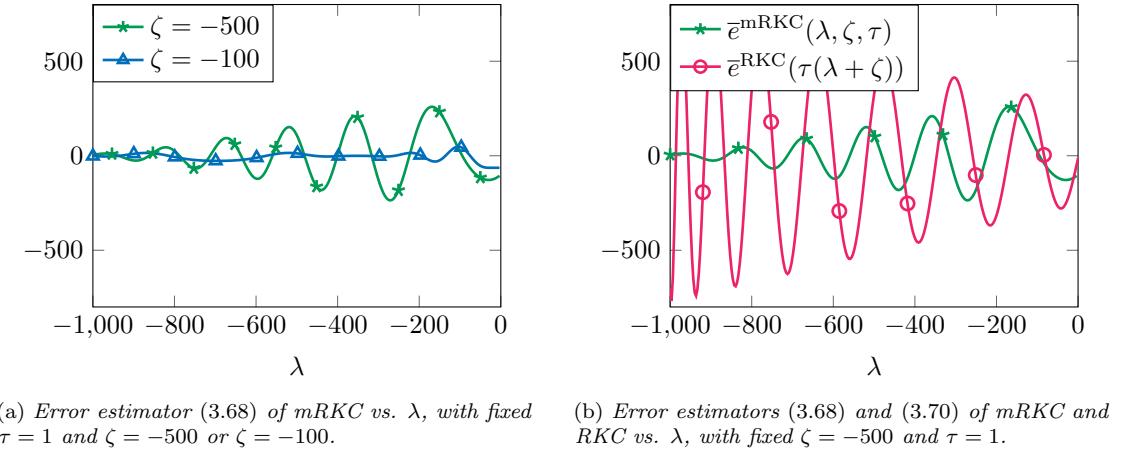


Figure 3.15. Dependence of error estimators on problem's stiffness.

Since $P_1''(0) = 0$ the second term in (3.61) disappears and for RKC relations (3.63) and (3.67) are equivalent; for the RKC scheme the error estimator (3.70) is therefore asymptotically exact. For general equations (3.1) the error estimator is again obtained identifying $R_j(\tau(\lambda + \zeta))y_n$ with k_j , which gives (3.69).

We end this section pointing out that, unfortunately, the error estimators (3.68) and (3.70) cannot be bounded by a stiffness independent constant; the coefficients r_{s-2}, r_{s-1}, r_s in (3.68) grow with s . In Figure 3.15(a) we plot the error estimator (3.68) for the mRKC method with fixed $\tau = 1$. First, we set $\zeta = -100$ and let λ vary in $[\lambda_{\min}, 0]$, with $\lambda_{\min} = -1000$. The number of stages s, m is chosen as in (3.56) with λ replaced by λ_{\min} , hence the mRKC scheme is stable. Then we do the same but setting $\zeta = -500$, hence increasing the stiffness of the slow components. We see that the error estimator oscillates with higher amplitude. This is due to the fact that the coefficients r_j in (3.68) increase with s , hence with $|\zeta|$. On the other hand, they are independent of the fast term λ . In Figure 3.15(b) we plot the error estimators (3.68) and (3.70) of the mRKC and the RKC scheme, with $\tau = 1$, fixed $\zeta = -500$ and we let again λ vary in $[\lambda_{\min}, 0]$ with $\lambda_{\min} = -1000$. For mRKC the number of stages is chosen again as in (3.56) and for RKC it is chosen as in (1.10) but with ρ replaced by $|\lambda_{\min} + \zeta|$. Since for the RKC scheme the number of stages s is higher and depends on the fast terms, the error estimator's oscillations are even stronger than for the mRKC method. We remind that the same issue is present in the error estimator proposed in [110] for the second-order RKC scheme.

Step size selection

Here we recall the step size selection strategy introduced in [66].

We denote $t_{n+1} = t_n + \tau_n$ and the local error committed at time t_{n+1} by $\bar{e}_{n+1} := \bar{e}_{n+1}(\tau_n)$, with $\bar{e}_{n+1}(\tau)$ defined in (3.69). Given an error tolerance tol , the errors \bar{e}_{n-1}, \bar{e}_n committed at time t_{n-1}, t_n using time steps τ_{n-2}, τ_{n-1} , respectively, we choose the next step size τ_n used to compute y_{n+1} as

$$\tau_n = \left(\frac{\text{tol}}{\bar{e}_n} \right)^{1/2} \left(\frac{\bar{e}_{n-1}}{\bar{e}_n} \right)^{1/2} \frac{\tau_{n-1}}{\tau_{n-2}} \tau_{n-1}. \quad (3.71)$$

The above formula is obtained using control theory and is well analyzed in [66]. It is not the only

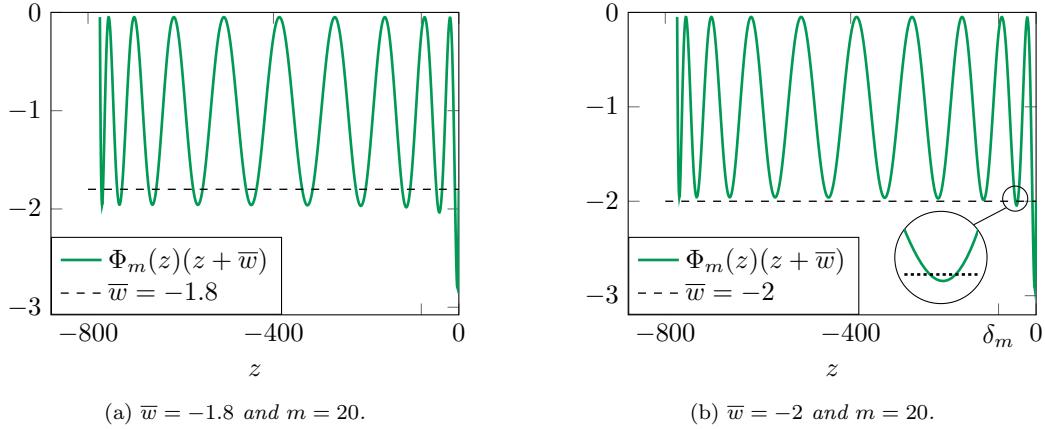


Figure 3.16. $\Phi_m(z)(z + \bar{w})$ vs. z , for $m = 20$ and $\bar{w} = -1.8$ (left) or $\bar{w} = -2$ (right).

possible choice, an alternative being, for instance,

$$\tau_n = \left(\frac{\text{tol}}{\bar{e}_n} \right)^{1/2} \tau_{n-1}, \quad (3.72)$$

which is employable also for $n = 1$.

The goal in (3.71) is to choose τ_n such that the next error (not known yet) satisfies $\bar{e}_{n+1} = \text{tol}$. To do so, it is assumed that $\bar{e}_{n+1} = \phi_n \tau_n^2$, where ϕ_n represents the differentials of f at time t_n , and that $\log(\phi_n)$ is an affine function, i.e. $\log(\phi_n) = \log(\phi_{n-1}) + \Delta \log(\phi_{n-1})$ with $\Delta \log(\phi_{n-1}) = \log(\phi_{n-1}) - \log(\phi_{n-2})$. Using the Z-transform one can then obtain the new step size τ_n , see [66] for more details. Formula (3.72) is obtained employing similar techniques but assuming constant $\phi_n = \phi_{n-1}$ and is therefore less robust. In control theory, (3.72) corresponds to an Integral (I) controller and (3.71) to a Proportional Integral Derivative (PID) controller, see [109] for a review on error controllers for initial value problems.

3.4.5 The mRKC method for problems with well separated scales

In this section we analyze the mRKC method under the additional assumption that (3.1) has well separated scales, as for multiscale and projective methods in [42, 57]. The difference, however, is that here we do not derive an effective equation nor perform a sequence of relaxation steps; indeed, we use the mRKC method (3.35) to (3.37) but with weaker stability conditions. In Figure 3.12(a) we saw that if $\Phi'_m(0)|w| \geq 1$ is not satisfied, then condition $\Phi_m(z)(z + w) \in [w, 0]$ is not satisfied in a neighborhood of $z = 0$ but it is still fulfilled elsewhere. In this section, we assume that scales are well separated and therefore z is never close to zero, it follows that the stability condition $\Phi'_m(0)|w| \geq 1$ can be alleviated. We warn that the weaker stability conditions, given in (3.74), are obtained experimentally and confirmed by the asymptotic behavior of stability polynomials, but are not proven analytically.

We know from Lemma 3.19 that choosing m such that $z \in [-\ell_m^\epsilon, 0]$ and $|w| = 2/P_m''(0) \approx 6$ gives a stable scheme. Let us fix m and consider $\bar{w} \in (w, 0)$, from Lemma 3.19 we also know that $\Phi_m(z)(z + \bar{w}) \notin [\bar{w}, 0]$ for some $z \in [-\ell_m^\epsilon, 0]$. We set $m = 20$ and display $\Phi_m(z)(z + \bar{w})$ for $z \in [-\ell_m^\epsilon, 0]$ with $\bar{w} = -1.8$ and $\bar{w} = -2$ in Figure 3.16. We observe that for $\bar{w} = -1.8$ the condition $\Phi_m(z)(z + \bar{w}) \in [\bar{w}, 0]$ is violated along the whole interval $[-\ell_m^\epsilon, 0]$, while for $\bar{w} = -2$ there is a $\delta_m > 0$ such that $\Phi_m(z)(z + \bar{w}) \in [\bar{w}, 0]$ for all $z \in [-\ell_m^\epsilon, -\delta_m]$ and for $z \in [-\delta_m, 0]$ it

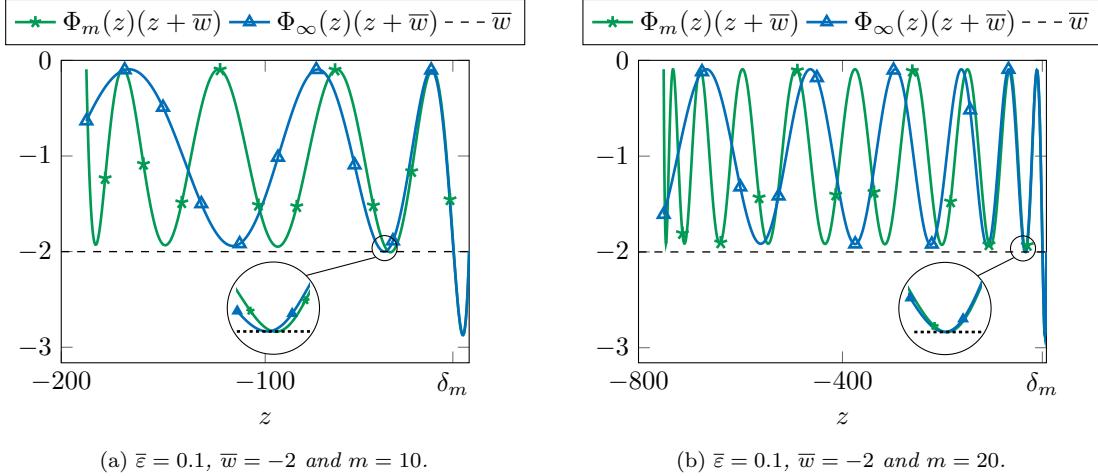


Figure 3.17. Illustration of $\Phi_\infty(z)(z + \bar{w})$ and $\Phi_m(z)(z + \bar{w})$ vs. z , for $\bar{\varepsilon} = 0.1$, $\bar{w} = -2$ and $m = 10$ (left) or $m = 20$ (right).

holds $\Phi_m(z)(z + \bar{w}) \notin [\bar{w}, 0]$ only two situations: when z is close to a local maximum of $\Phi_m(z)$ (see Figure 3.16(b)) or in the neighborhood of $z = 0$. Hence, the method is stable in most of the spectrum. Using $P_m(z) = a_m T_m(v_0 + v_1 z)$ we deduce

$$\Phi_m(z)(z + \bar{w}) = (P_m(z) - 1)(1 + \bar{w}/z) \geq (-1 - a_m)(1 + \bar{w}/z).$$

Since $-1 - a_m > -2$ if \bar{w}/z is small enough, that is when $z < -\delta_m$, then $\Phi_m(z)(z + \bar{w}) \geq -2$. Hence, if $\bar{w} \leq -2$ there is $\delta_m > 0$ such that $\Phi_m(z)(z + \bar{w}) \geq \bar{w}$ for $z \leq -\delta_m$. Henceforth, we consider $\bar{w} = -2$.

We see in Figure 3.16(b) that if the oscillations of $\Phi_m(z)$ had a slightly smaller amplitude then $\Phi_m(z)(z + \bar{w}) \in [\bar{w}, 0]$ would be violated only in the neighborhood of the origin. Indeed, we plot $\Phi_m(z)(z + \bar{w})$ for a damping $\bar{\varepsilon} = 0.1$ and $m = 10, 20$ in Figure 3.17. We find that $\Phi_m(z)(z + \bar{w}) \in [\bar{w}, 0]$ for all $z \in [\ell_m^{\bar{\varepsilon}}, -\delta_m]$ with $\delta_m = 8$. The same happens for all m large enough. Indeed, $P_m(z)$ converges uniformly as $m \rightarrow \infty$, for all z in a bounded set, to

$$\lim_{m \rightarrow \infty} P_m(z) = P_\infty(z) := \frac{\cosh(\sqrt{2(\varepsilon + \Omega(\varepsilon)^{-1}z)})}{\cosh(\sqrt{2\varepsilon})}, \quad \text{with} \quad \Omega(\varepsilon) = \frac{\tanh(\sqrt{2\varepsilon})}{\sqrt{2\varepsilon}},$$

see [2, 124]. Hence we can define $\Phi_\infty(z) = (P_\infty(z) - 1)/z$ and plot $\Phi_\infty(z)(z + \bar{w})$ in Figure 3.17, we see that $\Phi_\infty(z)(z + \bar{w}) \in [\bar{w}, 0]$ for all $z \leq -\delta_\infty$ with $\delta_\infty = 8$.

From the above considerations it follows that if $z \leq -8$ and $|\bar{w}| \geq 2$ then $\Phi_m(z)(z + \bar{w}) \in [\bar{w}, 0]$. Hence, for a step size $\tau > 0$ and $s, m, \bar{\eta}$ satisfying

$$\tau|\zeta| \leq \ell_s^\varepsilon, \quad \bar{\eta}|\lambda| \leq \ell_m^{\bar{\varepsilon}}, \quad \bar{\eta} \geq \max \left\{ \frac{8}{|\lambda|}, \frac{2\tau}{\ell_s^\varepsilon} \right\} \quad (3.73)$$

then the mRKC method applied to the multirate test equation (3.12) is stable. Indeed, the mRKC method is stable if its stability polynomial $R_{s,m}(\lambda, \zeta, \tau, \bar{\eta}) = R_s(\tau \Phi_m(\bar{\eta}\lambda)(\lambda + \zeta))$ is in $[-1, 1]$. As for Theorem 3.20, stability is implied by $\tau \Phi_m(\bar{\eta}\lambda)(\lambda + \zeta) \in [-\ell_s^\varepsilon, 0]$, which is equivalent to $\Phi_m(\bar{\eta}\lambda)(\bar{\eta}\lambda + \bar{\eta}\zeta) \in [w(\bar{\eta}), 0]$ with $w(\bar{\eta}) = -\bar{\eta}\ell_s^\varepsilon/\tau$. Since $\bar{\eta}\zeta \geq w(\bar{\eta})$, then

$$0 \geq \Phi_m(\bar{\eta}\lambda)(\bar{\eta}\lambda + \bar{\eta}\zeta) \geq \Phi_m(\bar{\eta}\lambda)(\bar{\eta}\lambda + w(\bar{\eta})) \geq w(\bar{\eta}).$$

The last inequality follows from the condition on $\bar{\eta}$ in (3.73), which implies both

$$\bar{\eta}\lambda \leq -8 \quad \text{and} \quad |w(\bar{\eta})| \geq 2,$$

therefore $\Phi_m(\bar{\eta}\lambda)(\bar{\eta}\lambda + w(\bar{\eta})) \geq w(\bar{\eta})$. Observe that for a problem without scale separation some eigenvalues λ of the Jacobian of f_F might be small, in magnitude, and thus $\bar{\eta}$ large. Assuming that all λ are large, in magnitude, we obtain the stability conditions for problems with well separated scales, given by

$$\tau\rho_S \leq \ell_s^\varepsilon, \quad \bar{\eta}\rho_F \leq \ell_m^{\bar{\varepsilon}}, \quad \text{with} \quad \bar{\eta} = \frac{2\tau}{\ell_s^\varepsilon}. \quad (3.74)$$

In practice we can replace ℓ_s^ε by βs^2 and $\ell_m^{\bar{\varepsilon}}$ by $\bar{\beta}m^2$, where $\bar{\beta} = 2 - 4\bar{\varepsilon}/2$ and $\bar{\varepsilon} = 0.1$. Hence $\bar{\beta} \approx 1.867$. The efficiency gain of replacing conditions (3.34) by the weaker (3.74) has already been discussed in Section 3.4.2.

3.5 Numerical Experiments

In this section we compare the mRKC scheme from Section 3.4.1 against the classical RKC method of Section 1.2.4 through a series of experiments. First, we apply mRKC to a stiff nonlinear dynamical system to verify convergence in the standard ‘‘ODE sense’’ and do a first efficiency comparison in the ODE context. Then, we apply mRKC to the heat equation to verify convergence in the ‘‘PDE sense’’, i.e. when both the mesh size H and the time step τ decrease simultaneously. In the third experiment, we compare the performance and efficiency of the mRKC and RKC schemes when applied to a PDE. In the fourth experiment, we study numerically the stability of mRKC when it is applied to various advection-diffusion-reaction problems. In the last experiment we verify the effectiveness of the step size adaptivity strategy derived in Section 3.4.4.

Both the RKC and mRKC methods need bounds on the spectral radii of the Jacobians of f_F and f_S in order to choose the number of stages s, m . In our experiments, we estimate them with a cheap nonlinear power method [86, 123]. The numerical experiments in Sections 3.5.2 to 3.5.4 have been performed using the C++ library `libMesh` [79].

3.5.1 Robertson’s stiff test problem

First, we study the convergence of the mRKC scheme on a very popular test problem for stiff numerical integrators: the Robertson’s nonlinear chemical reaction model [45, 69]

$$\begin{aligned} y'_1 &= -0.04 y_1 + 10^4 y_2 y_3, & y_1(0) &= 1, \\ y'_2 &= 0.04 y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2, & y_2(0) &= 2 \cdot 10^{-5}, \\ y'_3 &= 3 \cdot 10^7 y_2^2, & y_3(0) &= 10^{-1}, \end{aligned} \quad (3.75)$$

where $t \in [0, 100]$. With this set of parameters and initial conditions, the only term inducing severe stiffness is $-10^4 y_2 y_3$. Thus, we let

$$f_F(y) = \begin{pmatrix} 0 \\ -10^4 y_2 y_3 \\ 0 \end{pmatrix}, \quad f_S(y) = \begin{pmatrix} -0.04 y_1 + 10^4 y_2 y_3 \\ 0.04 y_1 - 3 \cdot 10^7 y_2^2 \\ 3 \cdot 10^7 y_2^2 \end{pmatrix}, \quad f(y) = f_F(y) + f_S(y).$$

Now, we solve (3.75) either with the RKC or the mRKC scheme using step sizes $\tau = 1/2^k$, $k = 0, \dots, 7$. For comparison, we use a reference solution obtained with the standard fourth order Runge–Kutta scheme using $\tau = 10^{-4}$. In Figure 3.18(a), we observe that both the RKC and the

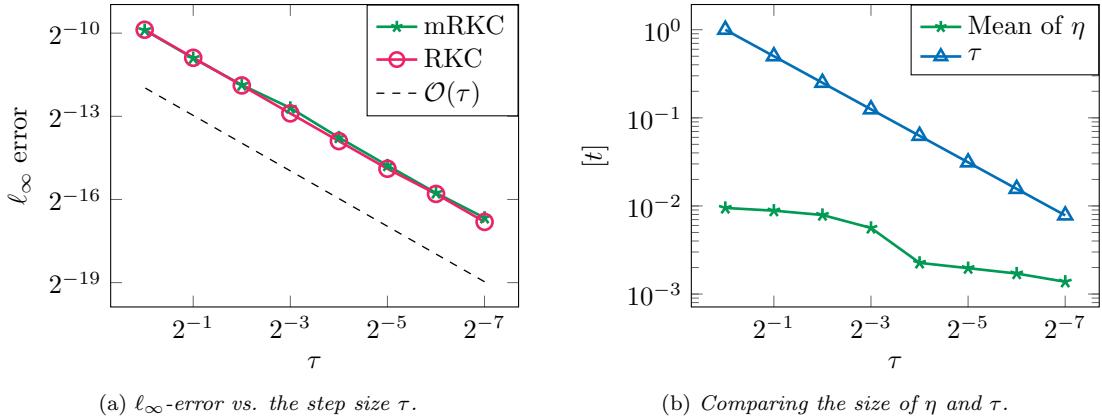
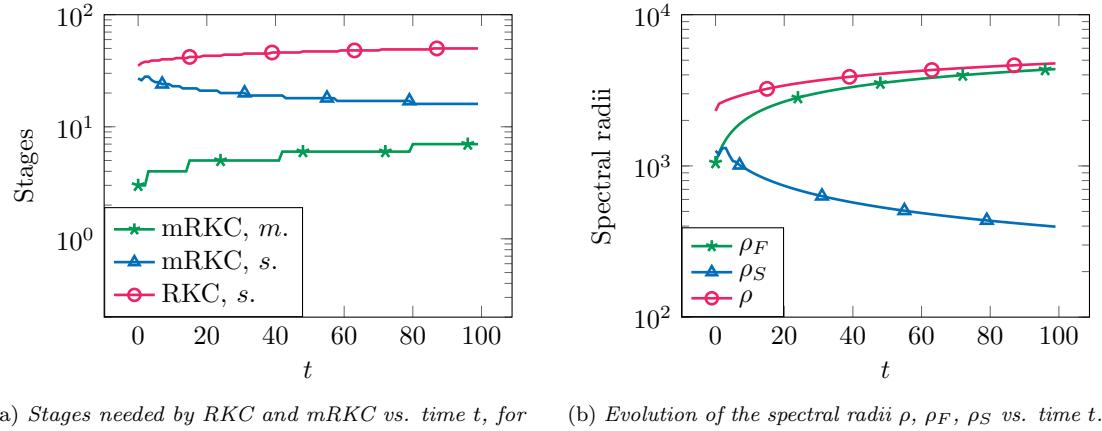
Figure 3.18. Robertson's stiff test problem. Convergence and comparison of η against τ .

Figure 3.19. Robertson's stiff test problem. Comparison of spectral radii and number of stages taken by the mRKC and the RKC scheme.

mRKC method achieve first-order convergence. In fact, both errors are hardly distinguishable, indicating that the error introduced by the approximation of f by f_η is negligible. We observe in Figure 3.18(b) that the mean value of η during integration is indeed considerably smaller than τ .

Now, we compare the two schemes for a fixed step size $\tau = 1$. In Figure 3.19(a), we display the number of stages taken by the mRKC and the RKC method at each time step with respect to $t \in [0, 100]$. Moreover, Figure 3.19(b) depicts the evolution of the spectral radii ρ, ρ_F, ρ_S of the Jacobians of f, f_F, f_S , respectively. We observe that ρ_S decreases with time and consequently the mRKC scheme decreases the number s of expensive function evaluations f_S per step. In contrast, ρ increases and thus the RKC scheme must increase the number s of f_S function evaluations, although this term does not introduce any stiffness; indeed, ρ increases only because of the term contained in f_F . Finally, we notice in Figure 3.19(a) that the mRKC scheme increases the number m of (cheap) function evaluations f_F because of the increase in ρ_F and η . Indeed, η increases too, due to the decrease in s and (3.34). This added cost, however, is much smaller than that from the many additional (expensive) evaluations of f_S required by the RKC method.

3.5.2 Heat equation in the unit square

Next, we verify the space-time convergence properties of the mRKC method. To do so, we consider the heat equation in the unit square $\Omega = [0, 1] \times [0, 1]$,

$$\begin{aligned}\partial_t u - \Delta u &= g && \text{in } \Omega \times [0, T], \\ u &= 0 && \text{in } \partial\Omega \times [0, T], \\ u &= 0 && \text{in } \Omega \times \{0\},\end{aligned}\tag{3.76}$$

where $T = 1/2$ and g is chosen such that $u(\mathbf{x}, t) = \sin(\pi x_1)^2 \sin(\pi x_2)^2 \sin(\pi t)^2$ is the exact solution.

Starting from a mesh of $2^j \times 2^j$ simplicial elements with $j = 2, \dots, 5$, we locally refine twice all the elements inside the square $\Omega_F = (1/4, 3/4) \times (1/4, 3/4)$. Each refinement step is performed by splitting all edges of any simplex, i.e. every triangle is split into four self-similar children. Let \mathcal{M} be the set of elements in the mesh and $\mathcal{M}_F = \{T \in \mathcal{M} : \bar{T} \cap \bar{\Omega}_F \neq \emptyset\}$ the set of refined elements or their direct neighbors. Then $h = H/4$ is the diameter of the elements inside of Ω_F , with H the diameter of the elements outside of Ω_F .

Next, we discretize (3.76) in space with first-order DG-FE [39] on the mesh \mathcal{M} . After inverting the block-diagonal mass matrix, the resulting system is

$$y' = Ay + G, \quad y(0) = y_0,$$

where $A \in \mathbb{R}^{N \times N}$ and $G \in C([0, T], \mathbb{R}^N)$ corresponds to the spatial discretization of $g(\cdot, t)$. Let $D \in \mathbb{R}^{N \times N}$ be a diagonal matrix with $D_{ii} = 1$ if the i th degree of freedom belongs to an element in \mathcal{M}_F and $D_{ii} = 0$ otherwise. We also introduce

$$A_F = DA, \quad A_S = (I - D)A \quad \text{and} \quad f_F(y) = A_F y, \quad f_S(t, y) = A_S y + G(t), \tag{3.77}$$

with I the identity. It is well-known that the spectral radii ρ_S and ρ_F of A_S and A_F behave as $\mathcal{O}(1/H^2)$ and $\mathcal{O}(1/h^2) = \mathcal{O}(16/H^2)$, respectively.

We now consider a sequence of meshes with $j = 2, \dots, 5$ and run the mRKC and the RKC scheme with the same time step $\tau = 1/2^j$. The parameters s and m for mRKC are chosen according to (3.44). In Figure 3.20(a), we display the $H^1(\Omega)$ errors at final time for mRKC and RKC. Both methods yield space-time first-order convergence and result in similar errors. In Figure 3.20(b), we show the number of stages needed by RKC and mRKC. For both schemes, s increases as the mesh size H decreases, but for mRKC it is smaller, since it depends on the coarse elements only. On the other hand, m remains constant, since the ratio between ρ_F and ρ_S is constant.

3.5.3 Diffusion across a narrow channel

To illustrate the efficiency of the mRKC method in a situation where geometry constraints require local mesh refinement, we consider the heat equation

$$\begin{aligned}\partial_t u - \Delta u &= g && \text{in } \Omega_\delta \times [0, T], \\ \nabla u \cdot \mathbf{n} &= 0 && \text{in } \partial\Omega_\delta \times [0, T], \\ u &= 0 && \text{in } \Omega_\delta \times \{0\},\end{aligned}\tag{3.78}$$

with $T = 0.1$ inside Ω_δ , which consists of two 10×5 rectangles linked by a narrow $\delta \times 0.05$ channel of width $\delta > 0$, see Figure 3.21. The right-hand side $g(\mathbf{x}, t) = \sin(10\pi t)^2 e^{-5\|\mathbf{x}-\mathbf{c}\|^2}$ corresponds to a smoothed Gaussian point source centered at \mathbf{c} in the middle of the upper rectangle.

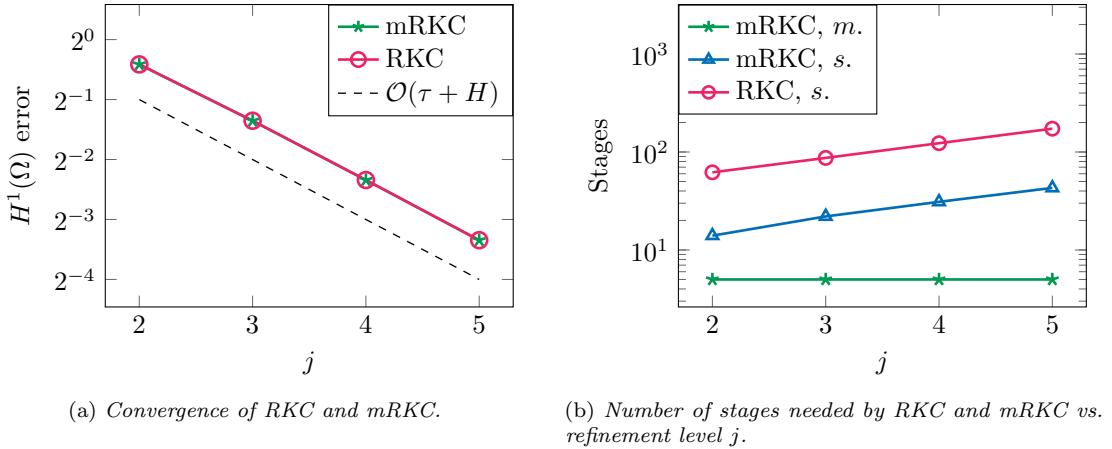
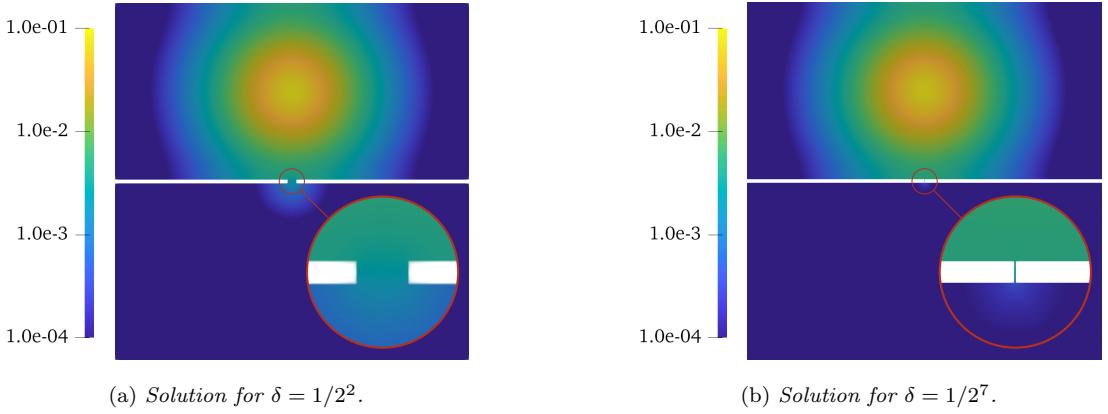


Figure 3.20. Heat equation in the unit square. Space-time convergence and number of stages.

Figure 3.21. Narrow channel. Numerical solutions of (3.78) at $t = 1$ using mRKC for a channel width $\delta = 1/2^2$ or $\delta = 1/2^7$.

Inside Ω_δ , we use a Delaunay triangulation with maximal element size $H \approx 0.015$. As δ approaches zero, the elements inside the channel become increasingly smaller and the system stiffer. For each $\delta > 0$, we define a neighborhood $\Omega_{F,\delta} \subset \Omega_\delta$ of the channel and \mathcal{M}_F , A , A_F , A_S , f_F , f_S as in Section 3.5.2. Here, $\Omega_{F,\delta}$ is chosen such that the spectral radius of A_S is almost independent of δ and only that of A_F increases with decreasing δ . Hence, $\Omega_{F,\delta}$ contains the channel together with all neighboring elements of mesh size smaller than H , see Figures 3.22(a) and 3.22(b).

For varying channel width $\delta = 1/2^k$, $k = 0, \dots, 15$, we now solve (3.78) with the RKC and mRKC method using the choice of parameters (3.44) with $\tau = 0.01$. In Figure 3.23(a), the relative speed-up defined as the ratio between the computational times of RKC and mRKC always exceeds one and reaches a value as high as 60. Note that the relative error between the two solutions in the $H^1(\Omega_\delta)$ norm is at most $3 \cdot 10^{-4}$, as shown in Figure 3.23(c).

In Figure 3.23(d), we display for varying δ also the spectral radii ρ, ρ_F, ρ_S of A, A_F, A_S , respectively; note that ρ_F and ρ essentially coincide. For large δ , we also have $\rho_F \approx \rho_S$ since the typical element size is sufficiently small to resolve the channel (Figure 3.22(a)). For δ small, we observe that ρ, ρ_F increase as $1/\delta^2$ while ρ_S remains almost constant. Figure 3.23(e) shows that

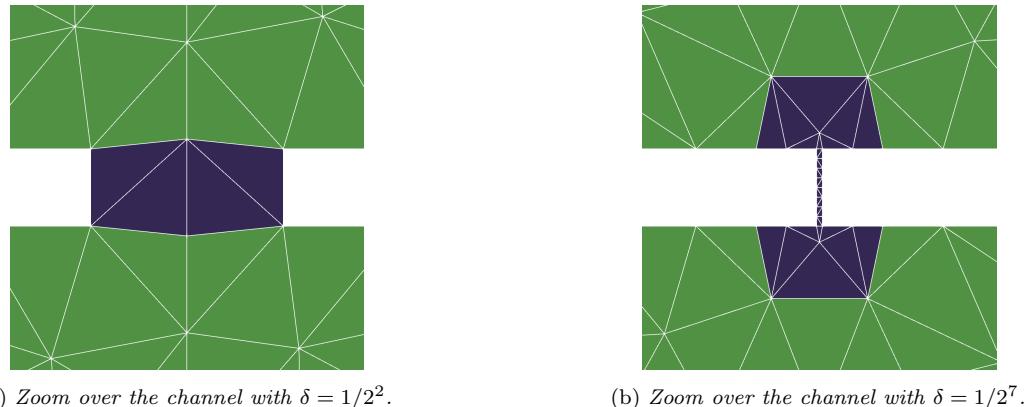


Figure 3.22. Narrow channel. Zoom of the FE mesh for a channel width $\delta = 1/2^2$ or $1/2^7$, with the subdomain $\Omega_{F,\delta}$ (in blue).

the number of stages s in the mRKC scheme remains constant, as does ρ_S in Figure 3.23(d), while m increases (as ρ_F). For large δ , we have $\rho_F \approx \rho_S$ and thus $m = 1$; then, the RKC and mRKC schemes are equivalent. Indeed, as is shown in Figure 3.23(c), for $m = 1$ the relative error between the RKC and mRKC solutions in $H^1(\Omega_\delta)$ norm is of the order of machine precision. As $m > 1$ the two schemes differ, the relative error increases and the jump in Figure 3.23(c) appears; nonetheless, the error remains very small.

In Figure 3.23(b), we observe that for δ large the CPU times of the two methods are similar; thus, despite $\rho_F \approx \rho_S$, there is no loss in efficiency and the speed-up is at least one (Figure 3.23(a)). For moderate values of δ , the cost of RKC increases proportionally to $1/\delta$, while the cost of mRKC is hardly affected. For even smaller δ , the number of evaluations of f_F increases and so does its cost with respect to f_S (see Figure 3.23(f)), since the number of elements in \mathcal{M}_F increases (Figure 3.22). In this regime, evaluation of f_F dominates the computational cost of mRKC, which increases linearly in $1/\delta$, too. Still, the mRKC method remains about sixty times faster than the classical RKC method for this particular discretization inside Ω_δ , see Figure 3.23(a).

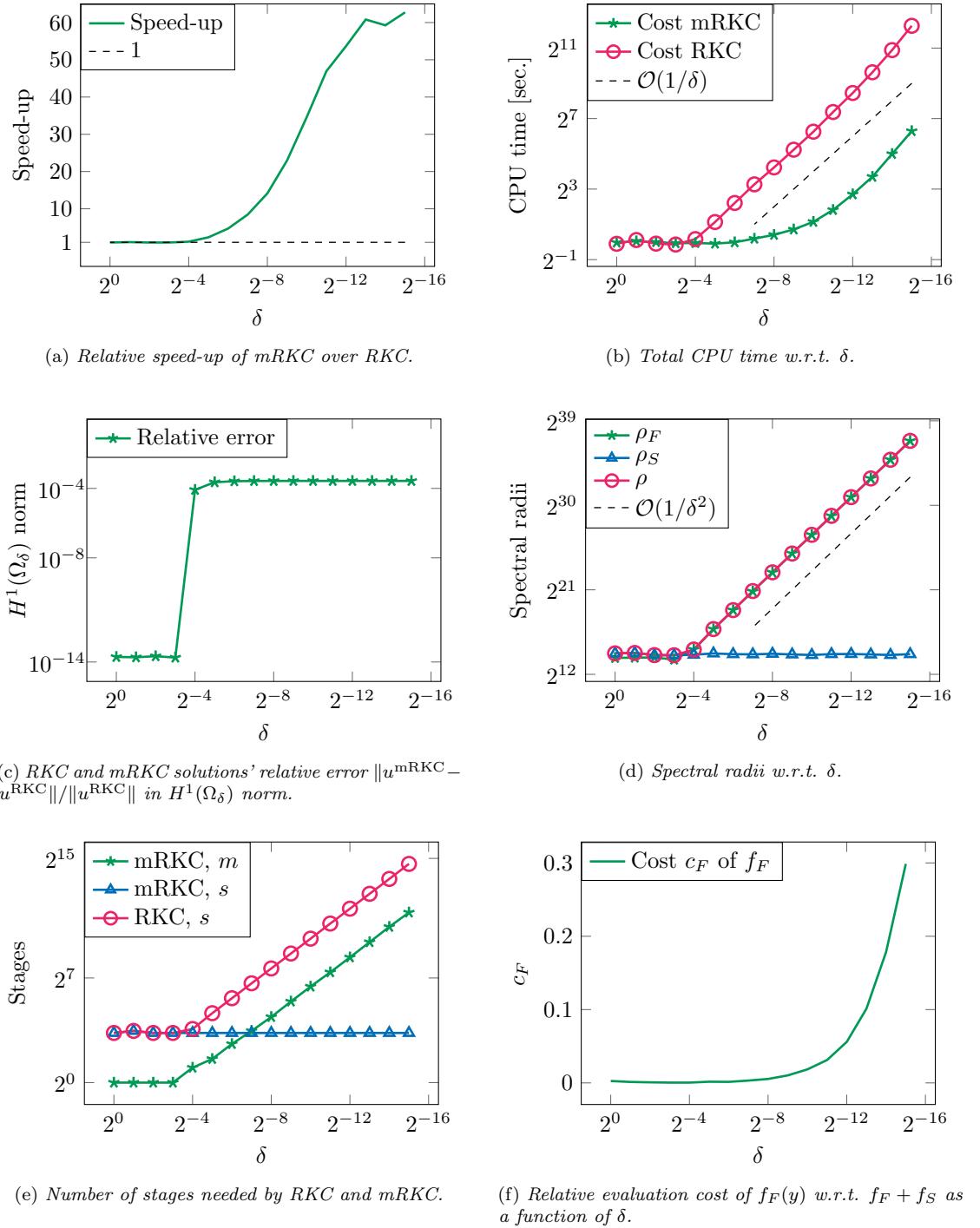
3.5.4 Reaction-convection-diffusion problem

In Section 3.4.3 we proved that the stability conditions of the mRKC method are the same for the 2×2 model problem (3.57) and for the scalar multirate test equation (3.12). The splitting of the discrete Laplace operator in (3.77) in fact is similar to that in (3.58) for the 2×2 model problem. Thus, one could expect that the stability conditions (3.34) are also necessary for more general parabolic problems. However, spatial discretizations of parabolic problems are much more complex than (3.57). Here we shall demonstrate via numerical experiment that the weaker stability conditions (3.44) in fact are also necessary and sufficient for general parabolic reaction-convection-diffusion problems, such as

$$\begin{aligned} \partial_t u - \nabla \cdot (K \nabla u) + \beta \cdot \nabla u + \mu u &= g && \text{in } \Omega \times [0, T], \\ u &= 0 && \text{in } \partial\Omega \times [0, T], \\ u &= u_0 && \text{in } \Omega \times \{0\}. \end{aligned}$$

We shall also demonstrate numerically that the mRKC does not need any scale separation assumption.

We will consider three parameter regimes. First, we let $\Omega = [0, 2] \times [0, 1]$, $K = I_{2 \times 2}$, $\beta = \mathbf{0}$

Figure 3.23. Narrow channel. Speed-up, error, spectral radii and stages number w.r.t. channel width δ .

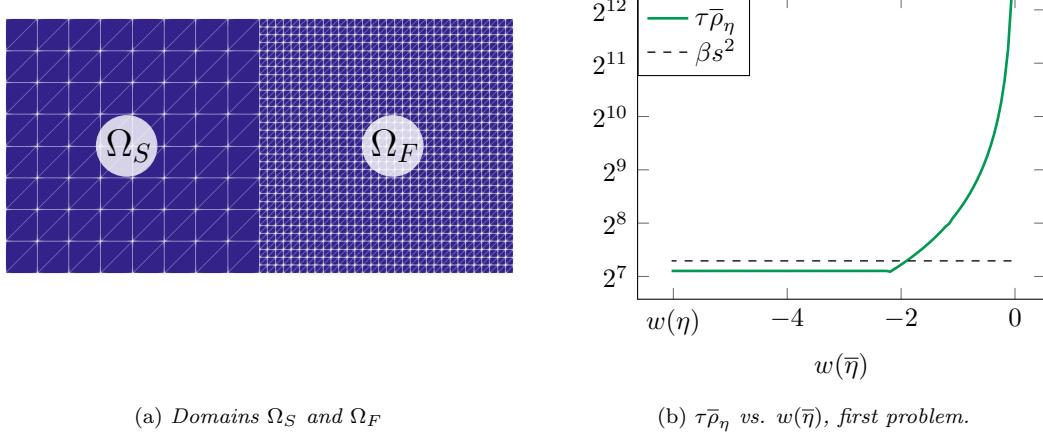


Figure 3.24. Reaction-convection-diffusion problem. Mesh and spectral radius $\tau\bar{\rho}_\eta$ of the first problem setting.

and $\mu = 0$. We build a 16×8 uniform mesh in Ω and refine twice the elements inside of $\Omega_F = (1, 2) \times (0, 1)$ (see Figure 3.24(a)). We use DG-FE and get the two matrices A_F and A_S as described in Section 3.5.2. Next, we set $\tau = 1$, s, m, η as in (3.34) and A_η as in (3.59). One step of the mRKC scheme is given by $y_1 = R_s(\tau A_\eta)y_0$. We recall that a necessary condition for stability of the scheme (at least for linear problems) is $\tau\rho_\eta \leq \beta s^2$, where ρ_η is the spectral radius of A_η .

Let $\bar{\beta}$ be as in (3.44), $\bar{\eta} \in [0, \eta]$, \bar{m} such that $\bar{\eta}\rho_F \leq \bar{\beta}\bar{m}^2$,

$$\bar{A}_\eta = \Phi_{\bar{m}}(\bar{\eta}A_F)A$$

and $\bar{\rho}_\eta$ be the spectral radius of \bar{A}_η . We wish to study for which $\bar{\eta}$ it holds $\tau\bar{\rho}_\eta \leq \beta s^2$. In Figure 3.24(b), we display $\tau\bar{\rho}_\eta$ for $\bar{\eta} \in (0, \eta)$ with respect to $w(\bar{\eta}) = -\bar{\eta}\beta s^2/\tau$: for $|w(\bar{\eta})| \geq 2$, it holds $\tau\bar{\rho}_\eta \leq \beta s^2$ and thus the scheme is stable. Observe that $|w(\bar{\eta})| \geq 2$ is equivalent to $\bar{\eta} \geq 2\tau/(\beta s^2)$, as in (3.44).

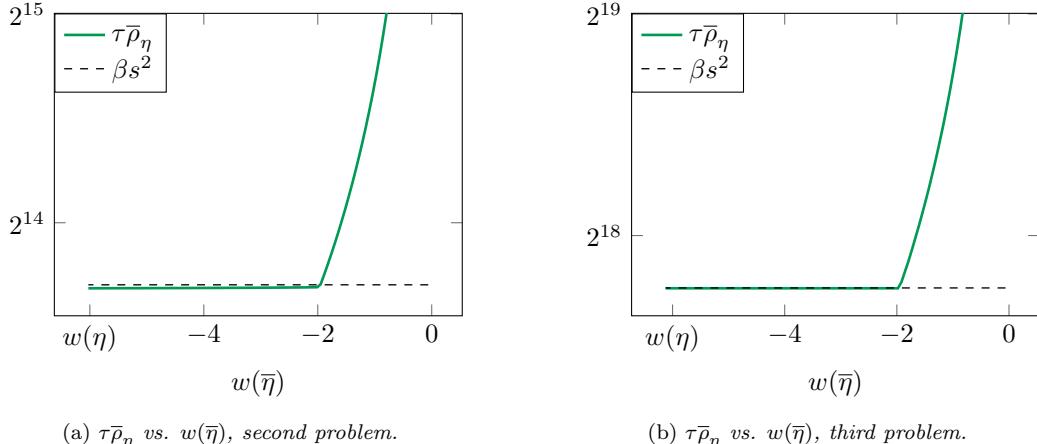


Figure 3.25. Stability experiment. Illustration of $\tau\bar{\rho}_\eta$ versus $w(\bar{\eta})$, second and third problem setting.

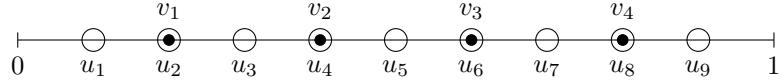


Figure 3.26. Brusselator problem. Illustration of the mesh for u and v with $N_u = 9$, $N_v = 4$.

Since the smallest eigenvalue (in magnitude) of the discrete Laplacian is almost independent of the mesh size, but depends on the size of the domain, the smallest (in magnitude) nonzero eigenvalues of A_F and A_S are approximately the same. As a consequence, this problem indeed does not satisfy any scale separation assumption. Still, as expected from the theory, the mRKC scheme remains stable.

Finally, we consider two additional cases that further corroborate the previous findings. First, we set $\Omega = [0, 1] \times [0, 1]$, $K = I_{2 \times 2}$, $\beta = (1, 1)^\top$ and $\mu = 1$. We build a 8×8 uniform mesh in Ω and refine three times the elements inside of $\Omega_F = (1/4, 3/4) \times (1/4, 3/4)$. In Figure 3.25(a), we plot again $\tau\bar{\rho}_\eta$ for $\bar{\eta} \in (0, \eta)$ with respect to $w(\bar{\eta})$: for $|w(\bar{\eta})| \geq 2$, $\tau\bar{\rho}_\eta \leq \beta s^2$ holds. Next, we use a uniform 32×32 mesh in $\Omega = [0, 1] \times [0, 1]$ which is refined twice in $\Omega_F = (0, 1/32) \times (0, 1/32)$. We also set $\beta = 0$, $\mu = 0$, $K(\mathbf{x}) = 1$ for $x_1 \geq x_2$ and $K(\mathbf{x}) = 0.1$ elsewhere. The results, shown in Figure 3.25(b), again confirm the stability of the mRKC with parameters chosen according to (3.44).

3.5.5 The Brusselator reaction-diffusion problem

In this section we explore the numerical properties of the error estimator \bar{e}_n given in (3.69). We consider the Brusselator problem [69, Section IV.1], defined by

$$\begin{aligned} \partial_t u - \alpha \Delta u + 4.4u - u^2v - 1 &= 0 && \text{in } [0, 1] \times [0, T], \\ \partial_t v - \alpha \Delta v - 3.4u + u^2v &= 0 && \text{in } [0, 1] \times [0, T], \\ u(t, 0) = u(t, 1) &= 1 && \text{for } t \in [0, T], \\ v(t, 0) = v(t, 1) &= 3 && \text{for } t \in [0, T], \\ u(0, x) &= 1 + \sin(2\pi x) && \text{for } x \in [0, 1], \\ v(0, x) &= 3 && \text{for } x \in [0, 1], \end{aligned} \tag{3.79}$$

with $T = 15$ and $\alpha = 1/50$. We discretize (3.79) using the finite difference scheme and different mesh sizes for the unknowns u and v : we use N_v mesh nodes for v and $N_u = 2N_v + 1$ nodes for u , see Figure 3.26 for an illustration. The discretized system can be written as

$$\begin{pmatrix} \mathbf{u}' \\ \mathbf{v}' \end{pmatrix} = f_F(\mathbf{u}) + f_S(\mathbf{u}, \mathbf{v}),$$

where $\mathbf{u} = (u_1, \dots, u_{N_u})^\top$, $\mathbf{v} = (v_1, \dots, v_{N_v})^\top$,

$$f_F(\mathbf{u}) = \begin{pmatrix} A_u \mathbf{u} \\ \mathbf{0} \end{pmatrix}, \quad f_S(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} \mathbf{0} \\ A_v \mathbf{v} \end{pmatrix} + F(\mathbf{u}, \mathbf{v}),$$

A_u , A_v are the discrete Laplacians for u , v , respectively, and $F(\mathbf{u}, \mathbf{v})$ contains the reaction and boundary condition terms.

We solve (3.79) with the RKC and mRKC schemes using step size adaptivity with a tolerance $\text{tol} = 10^{-5}$ and $N_v = 128$. In Figures 3.27(a) and 3.27(b) we show the step sizes taken by RKC and mRKC and observe that they are similar, indeed in Figures 3.27(c) and 3.27(d) we notice

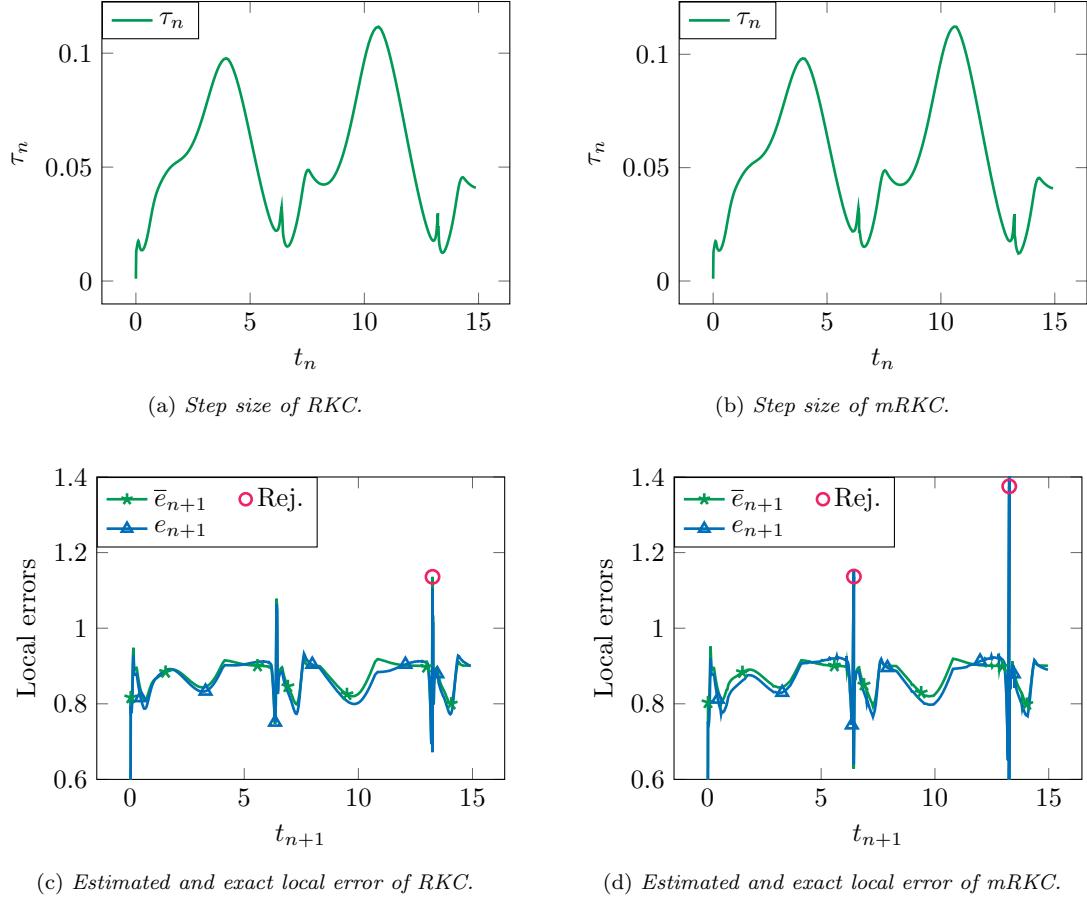


Figure 3.27. Brusselator problem. Evolution of the step size and the error estimators, for $N_v = 128$ and $\text{tol} = 10^{-5}$.

that the estimated errors \bar{e}_{n+1} of the two schemes are similar as well (note that the errors are scaled by tol). In the same figures we display the exact local error e_{n+1} , approximated doing explicit Euler steps of size $\delta\tau = \min\{\tau/10^4, 0.1 \cdot 2/\rho_F\}$. We see that in both cases the exact and estimated local errors behave similarly.

We performed the same experiment but with $N_v = 32$ and plotted the local errors in Figures 3.28(a) and 3.28(b). In this case, we observe that the RKC scheme is again well estimating the local error e_{n+1} while instead the mRKC scheme is less precise. The reason is that with the choice $N_v = 32$ the slow term f_S is less stiff and hence η is larger, as a consequence the missing term in the error estimator becomes important. We display in Figure 3.29 the values of τ and η for $N_v = 128$ or $N_v = 32$. We note that for $N_v = 128$ η is much smaller than τ while for $N_v = 32$ they are similar.

3.6 Proofs of technical results

In this section we prove all the technical results needed in Sections 3.2 and 3.4 to show the stability of the mEE and mRKC methods.

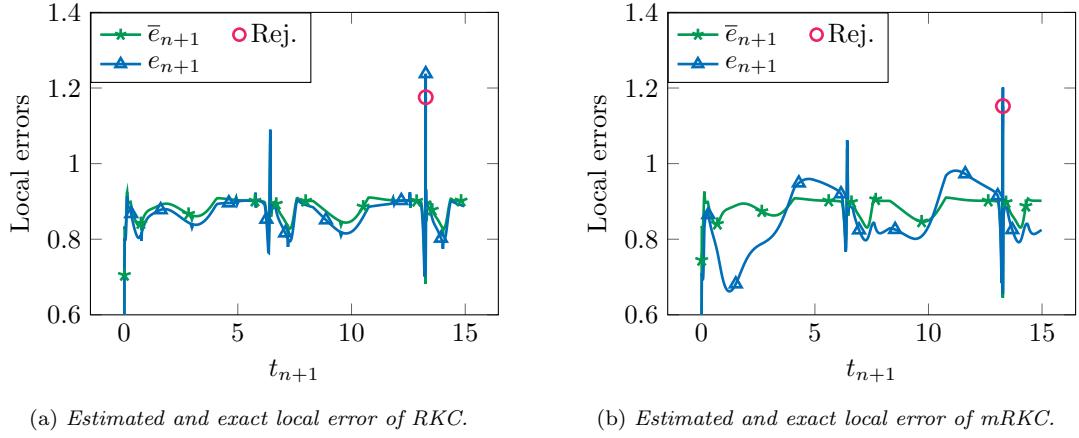


Figure 3.28. Brusselator problem. Evolution of the error estimators, for $N_v = 32$ and $\text{tol} = 10^{-5}$.

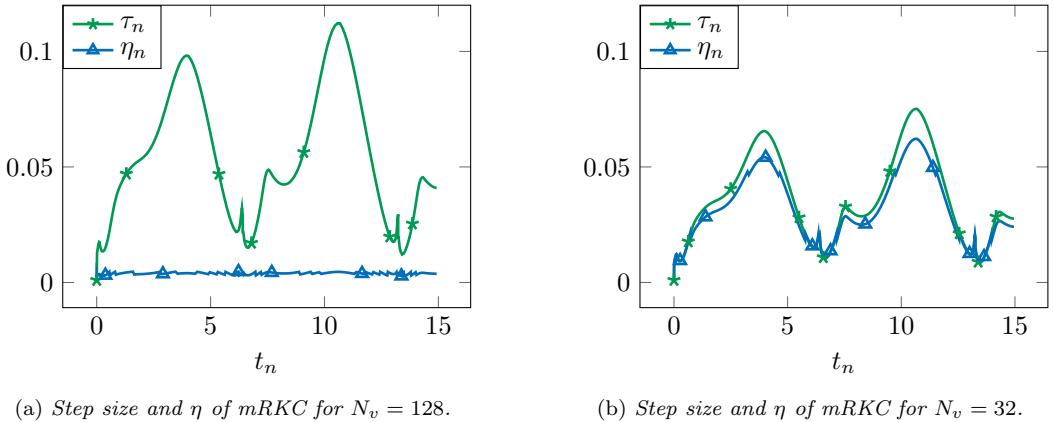


Figure 3.29. Brusselator problem. Evolution of the step size τ and η for different systems' stiffness.

3.6.1 Technical results for the mEE method

Here we prove Lemmas 3.10 and 3.11, needed in the proof of Theorem 3.12.

Proof of Lemma 3.10. We have that $(1+z)^N \leq 1$ for all N if and only if $z \in [-2, 0]$ and that $1+Nz \leq (1+z)^N$ for all z by the Bernoulli inequality. Rearranging inequalities $1+Nz \leq (1+z)^N \leq 1$ we find $0 \leq \Phi_N^{EE}(z) \leq 1$. \blacksquare

Proof of Lemma 3.11. We start supposing $\Phi_N^{EE}(z)(z+w) \in [w, 0]$ for all $z \in [-2, 0]$, hence $w \leq \Phi_N^{EE}(z)(z+w)$ holds. Writing

$$\Phi_N^{EE}(z) = \frac{1}{N} \sum_{k=0}^{N-1} (1+z)^k$$

we easily compute $\Phi_N^{EE}'(0) = (N-1)/2$ and following the lines of the first part of the proof of Lemma 3.6 we obtain $\Phi_N^{EE}'(0)|w| \geq 1$.

Now let us suppose $|w| \geq 2/(N-1)$ and show $\Phi_N^{EE}(z)(z+w) \in [w, 0]$ for all $z \in [-2, 0]$. Since $\Phi_N^{EE}(z) \geq 0$ we need only to show $w \leq \Phi_N^{EE}(z)(z+w)$, or equivalently

$$\Phi_N^{EE}(z)z \geq w(1 - \Phi_N^{EE}(z)),$$

which is implied by

$$(N-1)\Phi_N^{EE}(z)z \geq -2(1 - \Phi_N^{EE}(z)),$$

as $-2/(N-1) \geq w$. Rearranging the terms we obtain the equivalent condition

$$(1+z)^N((N-1)z-2) + 2 + (N+1)z \leq 0. \quad (3.80)$$

For $N = 1$ (3.80) holds clearly true. Let us suppose that (3.80) holds for N and $z \in [-1, 0]$, for $N+1$ we have

$$\begin{aligned} (1+z)^{N+1}(Nz-2) + 2 + (N+2)z &= (1+z)^N((N-1)z-2) + 2 + (N+1)z \\ &\quad + (1+z)^N(Nz^2-z) + z \\ &\leq (1+z)^N(Nz^2-z) + z \end{aligned}$$

and $(1+z)^N(Nz^2-z) + z \leq 0$ if and only if $(1+z)^N(1-Nz) \leq 1$. As $z \in [-1, 0]$ then $0 \leq (1+z)^N \leq e^{Nz}$ and $(1-Nz) \leq e^{-Nz}$, which implies $(1+z)^N(1-Nz) \leq 1$. Let us consider now the case $z \in [-2, -1]$ and N even. Since $1+Nz \leq (1+z)^N$ then

$$\begin{aligned} (1+z)^N((N-1)z-2) + 2 + (N+1)z &\leq (1+z)^N((N-1)z-2) + (1+z)^N + 1 + z \\ &\leq (1+z)^N((N-1)z-1) + 1 + z \leq 0 \end{aligned}$$

and (3.80) holds. For $z \in [-2, -1]$ and $N = \hat{N} + 1$ odd we have

$$\begin{aligned} (1+z)^N((N-1)z-2) + 2 + (N+1)z &= (1+z)^{\hat{N}}((\hat{N}-1)z-2) + 2 + (\hat{N}+1)z \\ &\quad + (1+z)^{\hat{N}}(\hat{N}z^2-z) + z \leq 0 \end{aligned}$$

as \hat{N} is even and $(1+z)^{\hat{N}}(\hat{N}z^2-z) + z \leq 0$, indeed $0 \leq (1+z)^{\hat{N}} \leq e^{\hat{N}z}$ and $(1-\hat{N}z) \leq e^{-\hat{N}z}$, thus $(1+z)^{\hat{N}}(1-\hat{N}z) \leq 1$. \blacksquare

3.6.2 Technical results for the mRKC method

Here we prove Lemmas 3.17, 3.19 and 3.23, all needed to prove Theorem 3.20.

Proof of Lemma 3.17. We start observing that $\Phi_m(z)$ defined in (3.49) is a polynomial. Indeed, since $P_m(0) = 1$ then $P_m(z) - 1$ is a polynomial with no zero order term. Hence, both sides of (3.53) are polynomials and if we can show that (3.53) holds in an interval then it holds everywhere. We will consider the values of z such that $v_0 + v_1 z \in]-1, 1[$, hence there exists $\theta \in]0, \pi[$ such that $v_0 + v_1 z = \cos(\theta)$. For such z , the right-hand side of (3.53) is

$$\sum_{j=1}^m \frac{a_m}{a_j} \alpha_j U_{m-j}(\cos(\theta)) = \frac{v_1}{T_m(v_0)} \left(U_{m-1}(\cos(\theta)) + 2 \sum_{j=1}^{m-1} T_{m-j}(v_0) U_{j-1}(\cos(\theta)) \right),$$

where we used $\alpha_1 = v_1/v_0$ and $\alpha_j = 2v_1 a_j / a_{j-1}$ for $j = 2, \dots, m$. Hence, (3.53) is equivalent to

$$\Phi_m(z) = \frac{v_1}{T_m(v_0)} \bar{\Phi}_m(\theta), \quad (3.81)$$

with

$$\bar{\Phi}_m(\theta) := U_{m-1}(\cos(\theta)) + 2 \sum_{j=1}^{m-1} T_{m-j}(v_0) U_{j-1}(\cos(\theta)).$$

Since

$$U_{j-1}(\cos(\theta)) = \frac{\sin(j\theta)}{\sin(\theta)}$$

we have

$$\bar{\Phi}_m(\theta) = \frac{\sin(m\theta)}{\sin(\theta)} + 2 \sum_{j=1}^{m-1} T_{m-j}(v_0) \frac{\sin(j\theta)}{\sin(\theta)}.$$

For $x \geq 1$ we have the identity $T_n(x) = \cosh(n \operatorname{acosh}(x))$, letting $\sigma = \operatorname{acosh}(v_0)$ it holds

$$T_{m-j}(v_0) = \cosh((m-j)\sigma) = \frac{1}{2} \left(e^{(m-j)\sigma} + e^{-(m-j)\sigma} \right). \quad (3.82)$$

Let $u = e^{-\sigma+i\theta}$ and $v = e^{\sigma+i\theta}$, from (3.81) and (3.82) we deduce

$$\begin{aligned} \bar{\Phi}_m(\theta) &= \frac{1}{\sin(\theta)} \left(\sin(m\theta) + \sum_{j=1}^{m-1} \left(e^{(m-j)\sigma} + e^{-(m-j)\sigma} \right) \sin(j\theta) \right) \\ &= \frac{1}{\sin(\theta)} \left(\sin(m\theta) + e^{m\sigma} \sum_{j=1}^{m-1} \operatorname{Im}(u^j) + e^{-m\sigma} \sum_{j=1}^{m-1} \operatorname{Im}(v^j) \right). \end{aligned}$$

We have

$$\begin{aligned} e^{m\sigma} \sum_{j=1}^{m-1} \operatorname{Im}(u^j) &= e^{m\sigma} \operatorname{Im} \left(\sum_{j=0}^{m-1} u^j \right) = e^{m\sigma} \operatorname{Im} \left(\frac{u^m - 1}{u - 1} \right) \\ &= \frac{\sin((m-1)\theta) - e^\sigma \sin(m\theta) + e^{m\sigma} \sin(\theta)}{2(\cosh(\sigma) - \cos(\theta))} \end{aligned}$$

and similarly

$$e^{-m\sigma} \sum_{j=1}^{m-1} \operatorname{Im}(v^j) = \frac{\sin((m-1)\theta) - e^{-\sigma} \sin(m\theta) + e^{-m\sigma} \sin(\theta)}{2(\cosh(\sigma) - \cos(\theta))}.$$

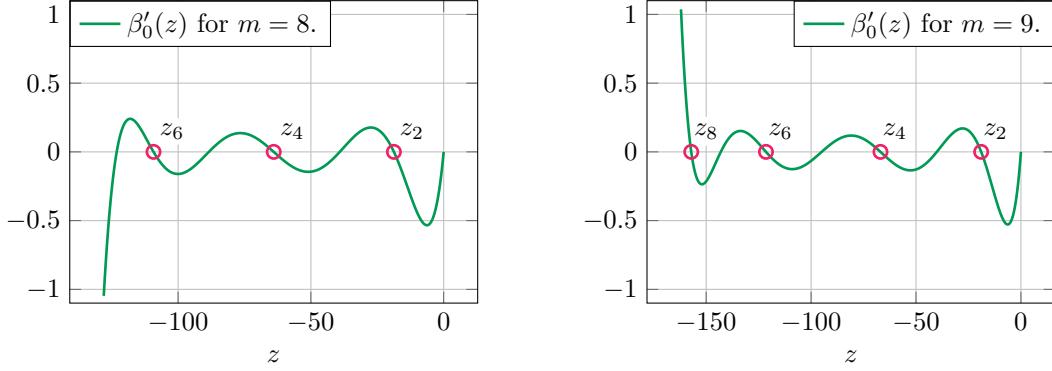


Figure 3.30. Plot of $\beta_0'(z)$ for $m = 8, 9$ and $|w| = 2/P_m''(0)$.

Using $\cosh(\sigma) = v_0$ and (3.82) for $k = 0$ we obtain

$$\bar{\Phi}_m(\theta) = \frac{\sin((m-1)\theta) - \sin(m\theta)\cos(\theta) + T_m(v_0)\sin(\theta)}{\sin(\theta)(v_0 - \cos(\theta))} = \frac{T_m(v_0) - \cos(m\theta)}{v_0 - \cos(\theta)},$$

where we used $\sin((m-1)\theta) = \sin(m\theta)\cos(\theta) - \sin(\theta)\cos(m\theta)$. Using the identity $T_m(\cos(\theta)) = \cos(m\theta)$, $\cos(\theta) = v_0 + v_1 z$ we get

$$\frac{v_1}{T_m(v_0)} \bar{\Phi}_m(\theta) = \frac{v_1}{T_m(v_0)} \frac{T_m(v_0 + v_1 z) - T_m(v_0)}{v_1 z} = \frac{P_m(z) - 1}{z}$$

and thus (3.81) holds. ■

Proof of Lemma 3.19. For the only if part we follow the lines of the proof of Lemma 3.6 and find that $\Phi'_m(0)|w| \geq 1$ is a necessary condition. We already computed $\Phi'(0) = P_m''(0)/2$ in (3.62).

Now, let us suppose $2/P_m''(0) \leq |w|$ and show $\beta_\varepsilon(z) = \Phi_m(z)(z+w) \geq w$, where ε is the damping. For $z = 0$ it is clear, independently of ε . We will show $\beta_0(z) > w$ for all $z < 0$, since $\beta_\varepsilon(z)$ depends continuously on ε there exists $\varepsilon_m > 0$ such that $\beta_\varepsilon(z) \geq w$ for all $\varepsilon \leq \varepsilon_m$. We have

$$\begin{aligned} \beta_0'(z) &= \frac{P'_m(z)}{z}(z+w) - \frac{P_m(z)-1}{z^2}w, \\ \beta_0''(z) &= \frac{P''_m(z)}{z}(z+w) - 2\frac{P'_m(z)}{z^2}w + 2\frac{P_m(z)-1}{z^3}w \end{aligned}$$

and since $\beta_0(z)$ is a polynomial of degree m then $\beta_0'(z)$ has at most $m-1$ zeros. We are going to locate the zeros $z_{m-1} < \dots < z_3 < z_2$ of $\beta_0'(z)$. Then we will use the fact that $\beta_0'(z)$ has at most one zero on the right of z_2 . In order to help the understanding of the proof we plot $\beta_0'(z)$ in Figure 3.30 for two values of m .

Since $P_m(z) = T_m(1+z/m^2)$ and $T_m(\cos(\theta)) = \cos(m\theta)$, choosing z_{2k} such that

$$1 + \frac{z_{2k}}{m^2} = \cos(\theta_{2k}) \quad \text{with} \quad \theta_{2k} = \frac{2k\pi}{m}$$

it yields

$$\beta_0'(z_{2k}) = 0 \quad \text{for } 2k = 2, 4, \dots, 2\lfloor \frac{m-1}{2} \rfloor.$$

Since z_{2k} is a local maximum of $P_m(z)$ then $P''_m(z_{2k}) < 0$ and $\beta''_0(z_{2k}) < 0$. In a neighborhood of z_{2k} we have

$$\beta'_0(z) = \beta''_0(z_{2k})(z - z_{2k}) + \mathcal{O}((z - z_{2k})^2),$$

hence for $\delta > 0$ small and $2k = 2, 4, \dots, 2(\lfloor \frac{m-1}{2} \rfloor - 1)$ we have $\beta'_0(z_{2k+2} + \delta) < 0$ and $\beta'_0(z_{2k} - \delta) > 0$, implying that there exists $z_{2k+1} \in [z_{2k+2}, z_{2k}]$ such that $\beta'_0(z_{2k+1}) = 0$. If m is odd then $2\lfloor \frac{m-1}{2} \rfloor = m - 1$ and we located the zeros z_j for $j = 2, 3, \dots, m - 1$. If m is even then $2\lfloor \frac{m-1}{2} \rfloor = m - 2$, but $P_m(-2m^2) = 1$ and $P'_m(-2m^2) = -1$ and hence $\beta'_0(-2m^2) < 0$. Thus, since $\beta'_0(z_{m-2} - \delta) > 0$ there exists $z_{m-1} \in]-2m^2, z_{m-2}[$ such that $\beta'_0(z_{m-1}) = 0$. Finally, we located z_j for $j = 2, 3, \dots, m - 1$ for m even and odd. We will show $\beta_0(z) > w$ for $z \in [z_2, 0[$ and then for $z \in [-2m^2, z_2]$.

Let $z \in [z_2, 0[$, if $z = z_1$ then $\beta'_0(z) = 0$ and else $\beta'_0(z) < 0$. Indeed, for z close to zero we have

$$\beta'_0(z) = 1 + \frac{1}{2}P''_m(0)w + (P''_m(0) + \frac{1}{3}P'''_m(0)w)z + \mathcal{O}(z^2).$$

If $2/P''_m(0) < |w|$ then $1 + \frac{1}{2}P''_m(0)w < 0$ and $\beta'_0(z) < 0$ in the neighborhood of zero. If $2/P''_m(0) = |w|$ then

$$\beta'_0(z) = \left(P''_m(0) - \frac{2}{3} \frac{P'''_m(0)}{P''_m(0)} \right) z + \mathcal{O}(z^2) = \frac{m^2 + 1}{5m^2} z + \mathcal{O}(z^2),$$

and $\beta'_0(z) < 0$ in the neighborhood of zero as well. If there exists $\bar{z} \in]z_2, 0[$ such that $\beta'_0(\bar{z}) > 0$ we can take $\delta > 0$ small enough to have $\bar{z} \in]z_2 + \delta, -\delta[$ and $\beta'_0(z_2 + \delta) < 0$ and $\beta'_0(-\delta) < 0$. Hence, β'_0 would change sign twice in the interval $]z_2 + \delta, -\delta[$, which is impossible since β'_0 has at most one zero on the right of z_2 . Hence, $\beta'_0(z) < 0$ for all $z \in [z_2, 0[$ except at most one point, since $\beta_0(0) = w$ it follows $\beta_0(z) > w$ for all $z \in [z_2, 0[$.

We consider now $z \leq z_2$. Using $1 - \cos(\theta) = 2 \sin(\theta/2)^2$ it holds $z_2 = -2m^2 \sin(\pi/m)$ and

$$\begin{aligned} \beta_0(z) &= \Phi_m(z)(z + w) = \frac{P_m(z) - 1}{z}(z + w) \geq -2 \frac{z + w}{z} \geq -2 - 2 \frac{w}{z} \geq wP''_m(0) - 2 \frac{w}{z_2} \\ &= \left(\frac{m^2 - 1}{3m^2} + \frac{1}{m^2 \sin(\pi/m)^2} \right) w > \left(\frac{1}{3} + \frac{1}{m^2 \sin(\pi/m)^2} \right) w. \end{aligned}$$

Thus, if $m^2 \sin(\pi/m)^2 \geq 3/2$ then $\beta_0(z) > w$. For $m = 2$ it is clearly true. We let $g(x) = x^2 \sin(\pi/x)^2$ and show that $g(x)$ is strictly increasing in $x \in [2, \infty[$, which implies $m^2 \sin(\pi/m)^2 \geq 3/2$ for all $m \geq 2$. We have

$$g'(x) = 2 \sin(\pi/x)(x \sin(\pi/x) - \pi \cos(\pi/x)) \geq 0$$

if and only if $x \sin(\pi/x) - \pi \cos(\pi/x) \geq 0$, which is equivalent to $\tan(\pi/x) \geq \pi/x$. The latter holds true since $\tan(\theta) \geq \theta$ for $\theta \in [0, \pi/2]$. \blacksquare

The next lemma has been used in the proof of Theorem 3.20.

Lemma 3.23. $P''_m(0) = T_m(v_0)T''_m(v_0)/T'_m(v_0)^2$ is increasing for $v_0 \geq 1$.

Proof. For $v_0 \geq 1$ we have the relation $T_m(v_0) = \cosh(m \operatorname{arccosh}(v_0))$. Let $\theta(v_0) = \operatorname{arccosh}(v_0)$, it holds

$$\frac{T_m(v_0)T''_m(v_0)}{T'_m(v_0)^2} = \coth(m\theta)(\coth(m\theta) - \frac{1}{m} \coth(\theta)) = c(\theta).$$

Since $\theta(v_0)$ is increasing for $v_0 \geq 1$, if $c(\theta)$ is increasing for $\theta \geq 0$ then $T_m(v_0)T_m''(v_0)/T_m'(v_0)^2$ is increasing for $v_0 \geq 1$. We have

$$c'(\theta) \geq 0 \iff \tanh(\theta) \tanh(m\theta) - 2m \tanh(\theta)^2 + \frac{1}{m} \frac{\sinh(m\theta)^2}{\cosh(\theta)^2} \geq 0.$$

From the convexity of $\sinh(\theta)$ we deduce $\sinh(m\theta)^2/m \geq m \sinh(\theta)^2$ and thus

$$c'(\theta) \geq 0 \iff \tanh(\theta) \tanh(m\theta) - m \tanh(\theta)^2 \geq 0.$$

The latter holds true since $\tanh(\theta)$ is convex and hence $\tanh(m\theta) \geq m \tanh(\theta)$. ■

4 Stabilized explicit multirate methods for stiff stochastic differential equations

In this chapter we extend the modified equation framework and the multirate methods of Chapter 3 to stiff stochastic differential equations

$$dX(t) = f(X(t)) dt + g(X(t)) dW(t), \quad X(0) = X_0, \quad (4.1a)$$

where

$$f(X) = f_F(X) + f_S(X), \quad (4.1b)$$

$X(t)$ is a stochastic process in \mathbb{R}^n , $f_F, f_S : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are drift terms, $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ is the diffusion term and $W(t)$ is an m -dimensional Wiener process. As in Chapter 3, we suppose that the drift term (4.1b) is composed by a cheap but severely stiff part f_F and an expensive but only mildly stiff part f_S and the eigenvalues of their Jacobian are distributed in a narrow strip along the negative real axis.

Standard numerical methods for (4.1) face the same computational challenges than those for (3.1): due to the stiffness of f_F standard methods suffer from a stringent step size restriction and implicit methods must solve large nonlinear systems. Stochastic stabilized explicit methods (the S-ROCK family) [2, 3, 4, 16] are a good compromise, as they enjoy an extended stability domain growing quadratically with the number of stages s . The scheme presented in [2] (see also Section 1.4.1), called SK-ROCK for second kind Runge–Kutta orthogonal Chebyshev, attains an optimal mean-square stability domain of size $L_s \approx 2s^2$. It is based on the deterministic Runge–Kutta–Chebyshev (RKC) method [110, 118, 125] (see also Section 1.2.4) and employs Chebyshev polynomials of the second kind for the stabilization of the stochastic integral. However, the number of stages s taken by the SK-ROCK method is dictated by the stiffness of f , as for the RKC scheme. Therefore, even if stiffness is induced by a few degrees of freedom only, the cost of numerical integration is high; indeed, the nonstiff expensive term f_S is evaluated concurrently to the stiff term f_F .

The goal of this chapter is twofold. First, we extend the modified equation (3.2) framework for stiff deterministic multirate equations (3.1) to stiff stochastic multirate equations (4.1). To do so, we replace $f = f_F + f_S$ by the averaged force f_η (see Definition 3.1) and then introduce an approximation g_η of the diffusion term g , called *damped diffusion*, where the stiffness is decreased thanks to the solutions to two fast deterministic auxiliary problems. Hence, we replace (4.1) by

$$dX_\eta(t) = f_\eta(X(t)) dt + g_\eta(X(t)) dW(t), \quad X_\eta(0) = X_0. \quad (4.2)$$

In Chapter 3 we defined f_η such that $\rho_\eta \leq \rho_S$, where ρ_η, ρ_S are the spectral radii of the Jacobians of f_η, f_S , respectively; indeed, our goal is to obtain a modified equation whose stiffness depends only on the slow dynamics. Here, we define g_η such that the mean-square stability properties of (4.1) are inherited by (4.2) albeit the drift has weaker dissipative properties. Second, departing

from (4.2) we will derive explicit multirate methods defined as the time discretization of (4.2) and of the auxiliary problems by explicit schemes. In particular, we will define the multirate Euler–Maruyama (mEM) method, where (4.2) is discretized by the Euler–Maruyama method and the auxiliary problems are solved by the explicit Euler scheme. Then we derive the multirate SK-ROCK (mSK-ROCK) method as a time discretization of the modified equation using the SK-ROCK method of Section 1.4.1, while the auxiliary problems are solved by the first-order RKC schemes of Section 1.2.4.

We would like to note that the multirate mEM method of Section 4.2 is mean-square stable but does not converge, in contrast the mSK-ROCK method of Section 4.3 is mean-square stable and converges with strong order 1/2 and weak order 1. The reason for the lack of convergence in the mEM method is that the Euler–Maruyama scheme has too strong stability conditions, in the sense that the step size restriction is stronger than for the explicit Euler scheme. As a consequence, the stability conditions for the mEM method are stronger than for the multirate explicit Euler method (see Section 3.2), to the point that η cannot decrease with τ and convergence is precluded. In contrast, the SK-ROCK scheme has the same stability conditions as the RKC scheme, thus the mSK-ROCK method has the same stability conditions as the multirate RKC method (see Section 3.4) and η behaves as τ , allowing for convergence. Because of this severe drawback in the mEM method, Section 4.2 must be regarded only as an exercise to understand the strategy to follow to derive stable multirate methods departing from the modified equation (4.2). Furthermore, it is useful to learn which stability properties a Runge–Kutta scheme for stochastic differential equations must satisfy in order to be employed in the derivation of a multirate method. The stability analysis derived in Sections 4.2.3 and 4.3.3 suggests that a Runge–Kutta scheme for stochastic differential equations can be used to derive a multirate method only if its stability conditions for stochastic and deterministic equations are equivalent, as for the SK-ROCK method and not for the Euler–Maruyama scheme. However, further research in this direction is needed.

The remainder of this chapter is organized as follows. In Section 4.1 we define the damping procedure for the diffusion term g and the stochastic modified equation (4.2). Next, in Section 4.2 we introduce the mEM method, study its efficiency, stability and accuracy. In Section 4.3 we introduce the mSK-ROCK method, we analyze its computational efficiency, prove its stability on a model problem, and prove that it has strong order 1/2 and weak order 1. Finally, Section 4.4 is devoted to numerical experiments, where we verify the theoretical results and confirm the efficiency of the mSK-ROCK method comparing it to the standard SK-ROCK scheme. The chapter is closed in Section 4.5, where we group all the technical results needed to study stability and accuracy of the mEM and mSK-ROCK methods.

4.1 The stochastic modified equation

We introduce here the stochastic modified equation (4.2) and study its mean-square stability properties. We begin providing some motivations for the definition of g_η , which is given in Section 4.1.2. In Section 4.1.2 we indeed define and analyze the modified equation.

4.1.1 Preliminary motivations

Before introducing the stabilization procedure of g for a general nonlinear equation (4.1) we motivate its definition considering the *stochastic multirate test equation*

$$dX(t) = (\lambda + \zeta)X(t) dt + \mu X(t) dW(t), \quad X(0) = X_0, \quad (4.3)$$

with $\lambda, \zeta \leq 0$ and $\mu \in \mathbb{R}$. As in Section 3.1.2, we identify $f_F(X) = \lambda X$ and $f_S(X) = \zeta X$, while we let $g(X) = \mu X$.

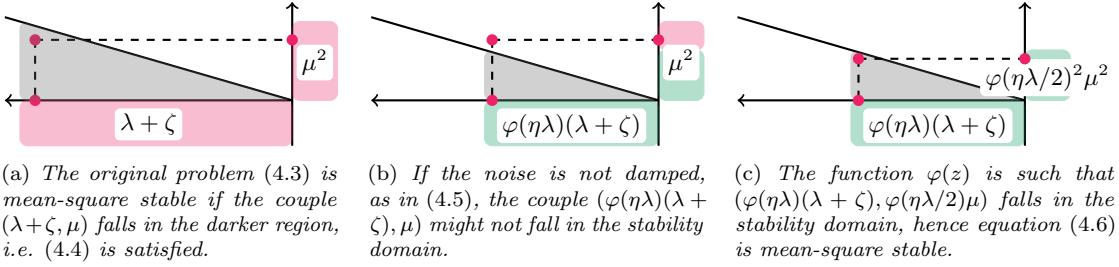


Figure 4.1. Illustration of the mean-square stability properties of the stochastic multirate test equation (4.3) (left), an unstable case where only the drift is damped (center) and the mean-square stable modified equation (4.6) (right), where both the drift and the noise are damped.

From Section 1.4.1 we deduce that (4.3) is mean-square stable if, and only if, $(\lambda + \zeta, \mu) \in \mathcal{S}^{MS}$, that is

$$\lambda + \zeta + \frac{1}{2}|\mu|^2 < 0. \quad (4.4)$$

In Figure 4.1(a) we depict the stability condition (4.4).

If $f(X) = (\lambda + \zeta)X$ is replaced by $f_\eta(X) = \varphi(\eta\lambda)(\lambda + \zeta)X$, see (3.14), then it yields

$$d\tilde{X}_\eta(t) = \varphi(\eta\lambda)(\lambda + \zeta)\tilde{X}_\eta(t) dt + \mu\tilde{X}_\eta(t) dW(t), \quad \tilde{X}_\eta(0) = X_0 \quad (4.5)$$

and the stability condition

$$\varphi(\eta\lambda)(\lambda + \zeta) + \frac{1}{2}|\mu|^2 < 0$$

of (4.5) is not guaranteed to hold, as is depicted in Figure 4.1(b). Indeed, η is chosen such that $\varphi(\eta\lambda)(\lambda + \zeta) \geq \zeta$, see Theorem 3.7, and usually $\zeta \gg \lambda + \zeta$.

Thus, the noise term μ must be damped; to do so, the next result is crucial.

Lemma 4.1. Let $z \in \mathbb{R}$, then $\varphi(z/2)^2 \leq \varphi(z)$.

Proof. Since $\varphi(z) = \int_0^1 e^{zs} ds$ the result follows from Jensen's inequality, as

$$\varphi\left(\frac{z}{2}\right)^2 = \left(\int_0^1 e^{\frac{z}{2}s} ds\right)^2 \leq \int_0^1 e^{zs} ds = \varphi(z). \quad \blacksquare$$

Hence, if in (4.5) we replace μ by $\varphi(\eta\lambda/2)\mu$, yielding

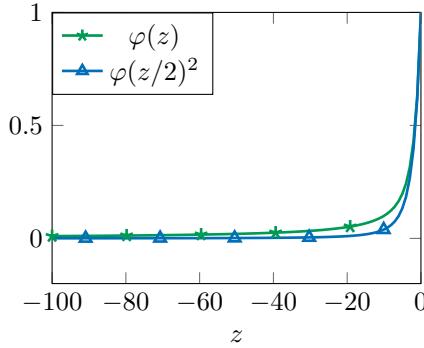
$$dX_\eta(t) = \varphi(\eta\lambda)(\lambda + \zeta)X_\eta(t) dt + \varphi\left(\frac{\eta}{2}\lambda\right)\mu X_\eta(t) dW(t), \quad X_\eta(0) = X_0, \quad (4.6)$$

the equation is stable. Indeed, the stability condition $(\varphi(\eta\lambda)(\lambda + \zeta), \varphi(\eta\lambda/2)\mu) \in \mathcal{S}^{MS}$, i.e.

$$\varphi(\eta\lambda)(\lambda + \zeta) + \frac{1}{2}|\varphi(\eta\lambda/2)\mu|^2 < 0, \quad (4.7)$$

is satisfied thanks to Lemma 4.1 and the stability properties of (4.3):

$$\varphi(\eta\lambda)(\lambda + \zeta) + \frac{1}{2}|\varphi(\eta\lambda/2)\mu|^2 \leq \varphi(\eta\lambda)(\lambda + \zeta + \frac{1}{2}|\mu|^2) < 0.$$

Figure 4.2. Illustration of $\varphi(z)$ and $\varphi(z/2)^2$.

We depict (4.7) in Figure 4.1(c) for completeness. In Figure 4.2 we notice that the inequality $\varphi(z/2)^2 \leq \varphi(z)$ is very tight and therefore the replacement of μ by $\varphi(\eta\lambda/2)\mu$ guarantees stability of (4.6) without over damping the noise term.

The goal of the next section is to define an approximation g_η of a general nonlinear diffusion term g such that when $g(X) = \mu X$ and $f_F(X) = \lambda X$ then $g_\eta(X) = \varphi(\eta\lambda/2)\mu X$.

4.1.2 The damped diffusion

Here, we define the damped diffusion g_η of (4.2) and study its properties.

Definition 4.2. Let $\eta \geq 0$, $v_0 \in \mathbb{R}^n$ and $v, \bar{v} : [0, \eta] \rightarrow \mathbb{R}^n$ defined by $v(0) = \bar{v}(0) = v_0$ and

$$v' = \frac{1}{2}f_F(v) + g(v_0), \quad \bar{v}' = \frac{1}{2}f_F(\bar{v}), \quad (4.8)$$

for $t \in [0, \eta]$. For $\eta > 0$, we define the *damped diffusion* g_η as

$$g_\eta(v_0) = \frac{1}{\eta}(v(\eta) - \bar{v}(\eta)) \quad (4.9)$$

and for $\eta = 0$ we define $g_0 = g$.

From (4.8) and (4.9) we obtain

$$g_\eta(v_0) = \frac{1}{\eta} \int_0^\eta (v'(s) - \bar{v}'(s)) \, ds = g(v_0) + \frac{1}{2\eta} \int_0^\eta (f_F(v(s)) - f_F(\bar{v}(s))) \, ds, \quad (4.10)$$

hence g_η is composed by g plus additional higher order terms. The role of $f_F(v)$ is to stabilize g , while $f_F(\bar{v})$ is used to remove the low order polluting terms introduced by $f_F(v)$. This is better seen in the next lemma.

Lemma 4.3. Let $\mu_F \in \mathbb{R}$ and f_F satisfy the one-sided Lipschitz condition (3.6). Then

$$\|g_\eta(v_0)\| \leq \varphi\left(\frac{\eta}{2}\mu_F\right) \|g(v_0)\|. \quad (4.11)$$

If, moreover, $f_F(y) = A_F y$ with $A_F \in \mathbb{R}^{n \times n}$, then

$$g_\eta(v_0) = \varphi\left(\frac{\eta}{2}A_F\right) g(v_0). \quad (4.12)$$

Proof. We let $\delta = \|v'(s) - \frac{1}{2}f_F(v(s))\| = \|g(v_0)\|$. Observing that the logarithmic norm of the Jacobian of $\frac{1}{2}f_F$ is bounded by $\frac{1}{2}\mu_F$, from a classical result on differential inequalities [68, Chapter I.10, Theorem 10.6] we obtain

$$\|v(s) - \bar{v}(s)\| \leq e^{\frac{s}{2}\mu_F} \int_0^s e^{-\frac{r}{2}\mu_F} \delta dr = s\varphi\left(\frac{s}{2}\mu_F\right) \|g(v_0)\|,$$

which yields (4.11) using (4.9). For the second result, replacing $f_F(y) = A_F y$ in (4.8) and using the variation-of-constants formula we deduce

$$v(\eta) = e^{\frac{\eta}{2}A_F} v_0 + \eta\varphi\left(\frac{\eta}{2}A_F\right) g(v_0), \quad \bar{v}(\eta) = e^{\frac{\eta}{2}A_F} v_0$$

and (4.12) follows from (4.9). \blacksquare

In (4.12) we observe the smoothing effect of f_F on g (as in (3.8) for f). A similar result holds for nonlinear f_F in (4.11). Furthermore, since in (4.10) it holds $f_F(v(s)) - f_F(\bar{v}(s)) = \mathcal{O}(\eta)$ as $\eta \rightarrow 0$, then $g_\eta(v_0) = g(v_0) + \mathcal{O}(\eta)$; the same is seen in (4.12), since $\varphi(z) = 1 + \mathcal{O}(z)$ as $z \rightarrow 0$.

4.2 The multirate Euler–Maruyama method

We introduce here the multirate Euler–Maruyama (mEM) method, which is a generalization of the multirate explicit Euler (mEE) method of Section 3.2. It is obtained by integrating (4.2) with the Euler–Maruyama method and (3.3) and (4.8) by explicit Euler methods. However, for v in (4.8) we need a modified explicit Euler scheme. Indeed, we saw in Section 4.1.1 that a crucial point for mean-square stability of the modified multirate test equation (4.6) is Lemma 4.1, i.e. $\varphi(z/2)^2 \leq \varphi(z)$, and it turns out that a numerical counterpart of this condition must also hold for the mean-square stability of the mEM method. We recall that for the mEE method $\Phi_N^{EE}(z)$ is the numerical version of $\varphi(z)$ and unfortunately relation $\Phi_N^{EE}(z/2)^2 \leq \Phi_N^{EE}(z)$ is not satisfied. Therefore, a modified explicit Euler scheme for v in (4.8) must be used, such that a similar condition is satisfied.

We would like to notice that the mEM method, albeit mean-square stable, does not converge. This deficiency follows from the fact that the stability domain of the EM scheme does not contain a portion of the mean-square stability domain \mathcal{S}^{MS} (1.42) of the stochastic test equation (1.41), see Section 1.4.1 and Figure 1.4(b) for more details about this. The main consequence of this downside is that the stability conditions for the EM scheme are stronger than for the explicit Euler scheme, thus the stability conditions for the mEM method are stronger than for the mEE method of Section 3.2; to the point that they prevent η to decrease with τ and thus inhibit convergence. See Remarks 4.4 and 4.9 for further details. In contrast, the stability domain of the SK-ROCK scheme contains a portion of the mean-square stability domain \mathcal{S}^{MS} , see Figure 1.4(a). Thus, the SK-ROCK scheme enjoys the same stability conditions as the RKC scheme and as a consequence the mSK-ROCK method derived in Section 4.3 enjoys the same stability conditions as the mRKC scheme of Section 3.4; therefore, η behaves as τ and the method genuinely converges with strong order 1/2 and weak order 1. Although the mEM method cannot be used in practice, its stability analysis is useful to understand the methodology to follow in order to derive multirate methods departing from the modified equation (4.2). Indeed, we will see that the mEM method and its stability analysis have similarities with the mSK-ROCK method of Section 4.3 below.

4.2.1 The mEM algorithm

We will present the algorithm for a vector valued diffusion $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$, the same algorithm for a matrix valued diffusion $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ is obtained following the strategy explained in

Section 4.3.1.

Let ρ_F, ρ_S, ρ_g be the spectral radii of the Jacobians of f_F, f_S, g , respectively, $\tau > 0$ the step size used to integrate (4.2), $\Delta\tau$ the micro step size used to integrate (3.3) and $N \in \mathbb{N}$ the number of micro step sizes. We suppose that $\tau, \Delta\tau, N, \eta$ satisfy

$$\tau \leq \frac{2(\rho_F + \rho_S) - \rho_g^2}{\rho_S(\rho_F + \rho_S)}, \quad \Delta\tau\rho_F \leq 2, \quad N \geq 1 + \frac{2}{\Delta\tau\rho_S}, \quad \eta = N\Delta\tau, \quad (4.13)$$

with the additional requirement that N is even with $2M = N$.

One step of the mEM method is given by

$$X_{n+1} = X_n + \tau \bar{f}_\eta(X_n) + \bar{g}_\eta(X_n)\Delta W_n, \quad (4.14)$$

with $\Delta W_n = W(t_{n+1}) - W(t_n)$, \bar{f}_η is the same discretization of f_η used by the mEE method in (3.20) and (3.21) and

$$\bar{g}_\eta(v_0) = \frac{1}{\eta}(v_M - \bar{v}_M) \quad (4.15)$$

is the numerical counterpart of g_η in (4.9). We compute v_M , the numerical approximation to $v(\eta)$ in (4.8), using M steps of a modified explicit Euler method, given by

$$\begin{aligned} v_1 &= v_0 + \Delta\tau f_F \left(v_0 + \frac{1}{2}\eta g(v_0) \right) + \eta g(v_0), \\ v_{j+1} &= v_j + \Delta\tau f_F(v_j), \quad j = 1, \dots, M-1. \end{aligned} \quad (4.16)$$

For the approximation \bar{v}_M of $\bar{v}(\eta)$ we use the explicit Euler method

$$\bar{v}_{j+1} = \bar{v}_j + \Delta\tau f_F(\bar{v}_j), \quad j = 0, \dots, M-1, \quad (4.17)$$

with initial condition $\bar{v}_0 = v_0$.

Stability of the mEM method is proved in Theorem 4.8 while accuracy is studied in Theorem 4.11.

Remark 4.4. We note in (4.13) that $\eta = N\Delta\tau \geq \Delta\tau + 2/\rho_S$, hence η is bounded from below by $2/\rho_S$. Furthermore, $2/\rho_S \geq \tau$. See also Remark 4.9.

4.2.2 Efficiency of the mEM method

Given ρ_F, ρ_S, ρ_g the spectral radii of the Jacobians of f_F, f_S, g , respectively, we compare here the computational costs of the standard Euler–Maruyama (EM) scheme and the mEM method, in terms of function evaluations. Hence, we denote by c_F, c_S, c_g the relative costs of evaluating f_F, f_S, g , respectively, with respect to the total evaluation cost of $f_F + f_S + g$, hence we set $c_F, c_S, c_g \in [0, 1]$ with $c_F + c_S + c_g = 1$.

Let us estimate the cost of the EM scheme, whose stability condition is

$$\tau \leq \frac{2\rho - \rho_g^2}{\rho^2},$$

with ρ the spectral radius of the Jacobian of f . Letting $\rho = \rho_F + \rho_S$, as we already did in Section 3.2.2, for the EM scheme we set

$$\tau = \frac{2(\rho_F + \rho_S) - \rho_g^2}{(\rho_F + \rho_S)^2}.$$

Supposing that we are integrating (4.1) in $[0, 1]$, the number of time steps needed by EM is $N_{\text{EM}} = 1/\tau$ and since at each time step it evaluates once f_F, f_S and g the integration cost is

$$C_{\text{EM}} = N_{\text{EM}}(c_F + c_S + c_g) = \frac{(\rho_F + \rho_S)^2}{2(\rho_F + \rho_S) - \rho_g^2}.$$

For the mEM method, following (4.13), we set

$$\tau = \frac{2(\rho_F + \rho_S) - \rho_g^2}{\rho_S(\rho_F + \rho_S)}, \quad \Delta\tau\rho_F = 2, \quad N \geq 1 + \frac{2}{\Delta\tau\rho_S} = 1 + \frac{\rho_F}{\rho_S},$$

with N even. Hence, denoting $r_\rho = \rho_F/\rho_S$, we set

$$N = 1 + \lceil r_\rho \rceil,$$

where $\lceil x \rceil$ denotes the smallest odd integer greater or equal to x .

Per time step, the mEM method needs N evaluations of f_F and one of f_S to compute \bar{f}_η and $2M = N$ evaluations of f_F and one of g for computing \bar{g}_η . Since $N_{\text{mEM}} = 1/\tau$ is the number of time steps, the cost of mEM is

$$\begin{aligned} C_{\text{mEM}} &= N_{\text{mEM}}(2Nc_F + c_S + c_g) = \frac{\rho_S(\rho_F + \rho_S)}{2(\rho_F + \rho_S) - \rho_g^2} (2(1 + \lceil r_\rho \rceil)c_F + 1 - c_F) \\ &= \frac{\rho_S(\rho_F + \rho_S)}{2(\rho_F + \rho_S) - \rho_g^2} (1 + (1 + 2\lceil r_\rho \rceil)c_F). \end{aligned}$$

The relative speed-up is defined as the ratio between the two costs, hence

$$S = \frac{C_{\text{EM}}}{C_{\text{mEM}}} = \frac{\rho_F + \rho_S}{\rho_S} \frac{1}{1 + (1 + 2\lceil r_\rho \rceil)c_F} = \frac{1 + r_\rho}{1 + (1 + 2\lceil r_\rho \rceil)c_F}.$$

As we already did in Section 3.2.2, we display in Figure 4.3(a) the relative speed-up S as a function of c_F for some values of $r_\rho = \rho_F/\rho_S$. We observe as S increases for $c_F \rightarrow 0$ but does not satisfy $S > 1$ for all $c_F \in [0, 1]$, in contrast to the mEE method in Figure 3.5(a). The main reason for this is that the mEM method needs to evaluate f_F to compute \bar{g}_η .

Now, we are interested on the maximal cost for c_F which still leads to a speed-up $S > 1$ and hence the mEM method is advantageous over the standard EM scheme. To do so, we solve the inequality $S > 1$ for varying c_F and find that $S > 1$ if, and only if,

$$c_F < c_F^{\max} = \frac{r_\rho}{1 + 2\lceil r_\rho \rceil}.$$

We monitor c_F^{\max} with respect to $r_\rho = \rho_F/\rho_S$ in Figure 4.3(b), we note that for $r_\rho > 1$, i.e. $\rho_F > \rho_S$, c_F^{\max} rapidly approaches 0.5 as r_ρ grows; with $c_F^{\max} > 0.4$ already for $r_\rho > 10$. For the less relevant case $r_\rho < 1$, the evaluation cost of f_F must be low but not necessarily zero.

4.2.3 Mean-square stability analysis

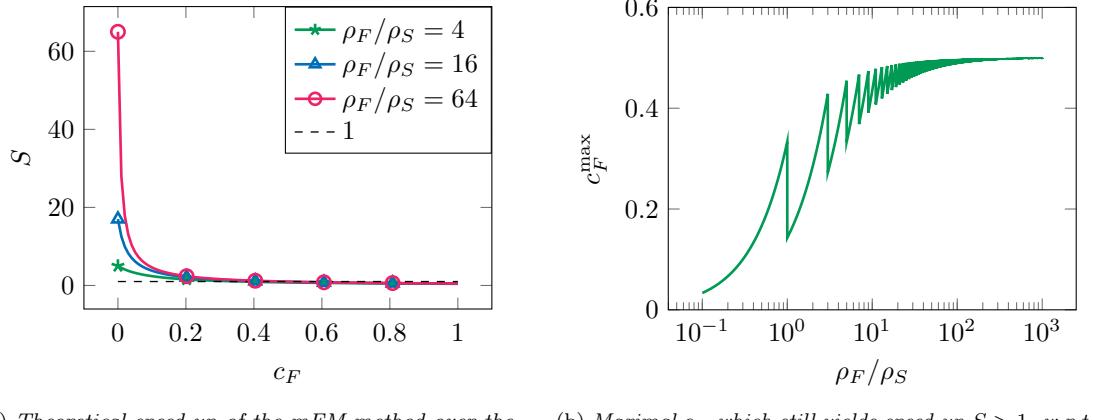
We prove now that the mEM method applied to the stochastic multirate test equation (4.3) is mean-square stable. We start by giving a closed expression for \bar{g}_η in (4.15).

Lemma 4.5. *Let $\lambda, \mu \in \mathbb{R}$, $f_F(x) = \lambda x$, $g(x) = \mu x$ and $v_0 \in \mathbb{R}$. Then, the modified diffusion \bar{g}_η (4.15), is given by*

$$\bar{g}_\eta(v_0) = \Psi_M^{EE}(\Delta\tau\lambda)\mu v_0, \tag{4.18}$$

where

$$\Psi_M^{EE}(z) = (1 + z)^{M-1} \left(1 + \frac{1}{2}z\right). \tag{4.19}$$



(a) Theoretical speed-up of the mEM method over the standard EM scheme, for varying c_F and fixed $\rho_F/\rho_S \in \mathbb{N}$.
 (b) Maximal c_F which still yields speed-up $S > 1$, w.r.t. $\rho_F/\rho_S \in \mathbb{R}$.

Figure 4.3. The relative speed-up S of the mEM method over the EM scheme with respect to c_F and the maximal value for c_F which still leads to an efficiency gain.

Proof. Replacing $f_F(x) = \lambda x$ and $g(v_0) = \mu v_0$ in (4.16) we obtain

$$v_1 = (1 + \Delta\tau\lambda)v_0 + \left(1 + \frac{1}{2}\Delta\tau\lambda\right)\eta\mu v_0,$$

$$v_{j+1} = (1 + \Delta\tau\lambda)v_j, \quad j = 1, \dots, M-1$$

and hence, by recursion,

$$v_M = (1 + \Delta\tau\lambda)^M v_0 + (1 + \Delta\tau\lambda)^{M-1} \left(1 + \frac{1}{2}\Delta\tau\lambda\right) \eta\mu v_0.$$

Similarly, we have $\bar{v}_M = (1 + \Delta\tau\lambda)^M v_0$ and we obtain (4.18) using (4.15). \blacksquare

Let $\xi_n \sim \mathcal{N}(0, 1)$ and $\Delta W_n = \tau^{1/2}\xi_n$, replacing (3.25) and (4.18) into (4.14) yields

$$X_{n+1} = X_n + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta)X_n + \tau^{1/2}\xi_n\Psi_M^{EE}(\Delta\tau\lambda)\mu X_n,$$

from which follows the next definition.

Definition 4.6. Let $\tau, \Delta\tau > 0$, $M, N \in \mathbb{N}$, $\lambda, \zeta \leq 0$, $\mu \in \mathbb{R}$ and $\xi \sim \mathcal{N}(0, 1)$ a Gaussian random variable. The stability polynomial of the mEM method is defined as

$$R_{N,M}(\tau, \Delta\tau, \lambda, \zeta, \mu, \xi) = 1 + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) + \tau^{1/2}\xi\Psi_M^{EE}(\Delta\tau\lambda)\mu, \quad (4.20)$$

with $2M = N$ and Φ_N^{EE} , Ψ_M^{EE} as in (3.24) and (4.19).

The next lemma is the numerical counterpart of Lemma 4.1, its proof is postponed to Section 4.5.1.

Lemma 4.7. Let $M, N \in \mathbb{N}$ with $2M = N$. Then $\Psi_M^{EE}(z)^2 \leq \Phi_N^{EE}(z)$ for all $z \in [-2, 0]$.

We display in Figure 4.4 the polynomials $\Psi_M^{EE}(z)^2$ and $\Phi_N^{EE}(z)$; note that $\Psi_M^{EE}(z)^2 \leq \Phi_N^{EE}(z)$ indeed holds for $2M = N$.

Thanks to Lemma 4.7 we can now show the stability of the mEM method.

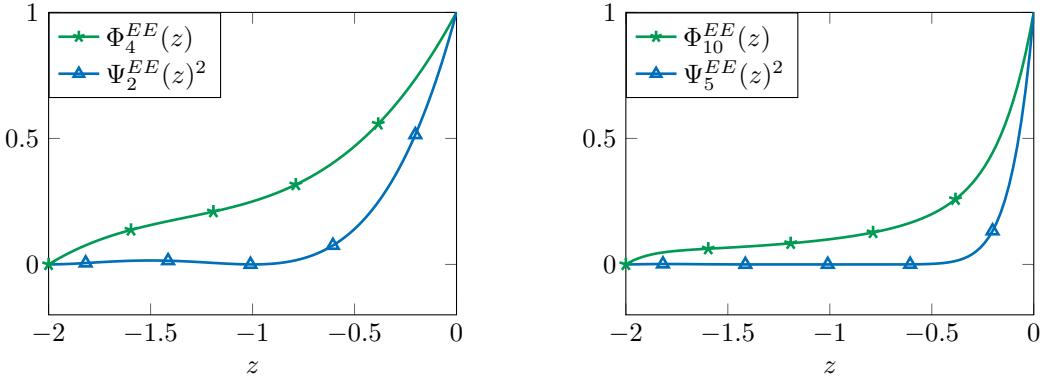


Figure 4.4. Illustration of $\Phi_N^{EE}(z)$ and $\Psi_M^{EE}(z)^2$ for $2M = N$, with $M = 2$ (left) and $M = 5$ (right).

Theorem 4.8. Let $(\lambda + \zeta, \mu) \in \mathcal{S}^{MS}$ with $\lambda \leq 0$ and $\zeta < 0$. Then, for all $\tau, \Delta\tau, N$ and η with $N = 2M$ satisfying

$$\tau \leq \frac{|2(\lambda + \zeta) + |\mu|^2|}{|\zeta(\lambda + \zeta)|}, \quad \Delta\tau|\lambda| \leq 2, \quad N \geq 1 + \frac{2}{\Delta\tau|\zeta|}, \quad \eta = N\Delta\tau, \quad (4.21)$$

it holds $\mathbb{E}(|R_{N,M}(\tau, \Delta\tau, \lambda, \zeta, \mu, \xi)|^2) \leq 1$, i.e. the mEM method is stable.

Proof. From (4.20) follows that $\mathbb{E}(|R_{N,M}(\tau, \Delta\tau, \lambda, \zeta, \mu, \xi)|^2) \leq 1$ is equivalent to

$$\tau\Phi_N^{EE}(\Delta\tau\lambda)^2(\lambda + \zeta)^2 + 2\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) + \Psi_M^{EE}(\Delta\tau\lambda)^2|\mu|^2 \leq 0. \quad (4.22)$$

From Lemma 4.7 it holds $\Psi_M^{EE}(\Delta\tau\lambda)^2 \leq \Phi_N^{EE}(\Delta\tau\lambda)$ and thus

$$\begin{aligned} \tau\Phi_N^{EE}(\Delta\tau\lambda)^2(\lambda + \zeta)^2 + 2\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) + \Psi_M^{EE}(\Delta\tau\lambda)^2|\mu|^2 \\ \leq \Phi_N^{EE}(\Delta\tau\lambda)(\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta)^2 + 2(\lambda + \zeta) + |\mu|^2), \end{aligned}$$

we recall that $\Phi_N^{EE}(\Delta\tau\lambda) \geq 0$, from (4.21) and Lemma 3.10. As $\Delta\tau|\zeta| \geq 2/(N-1)$, Lemma 3.11 implies $\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq \zeta$, hence $\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta)^2 \leq \zeta(\lambda + \zeta)$ and

$$\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta)^2 + 2(\lambda + \zeta) + |\mu|^2 \leq \tau\zeta(\lambda + \zeta) + 2(\lambda + \zeta) + |\mu|^2 \leq 0,$$

by choice of τ in (4.21). ■

Remark 4.9. We remind that, for proving stability of the mEE method in Theorem 3.12 it was sufficient to show $1 + \tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \in [-1, 1]$ and thus $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq -2$, which is an explicit Euler-like condition. However, the Euler–Maruyama scheme has stronger stability conditions than the explicit Euler scheme. We see in Figure 1.4(b) that it is not sufficient that $p \in [-2, 0]$ and $(p, q) \in \mathcal{S}^{MS}$, i.e. (p, q) in-between the dashed lines, for stability of the scheme. Instead, it is necessary that (p, q) enters in the ellipse. In the multirate framework, this condition translates in inequality (4.22), which in turn imposes $\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq \zeta$ and therefore $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq \tau\zeta$. From (4.21) follows that if $\mu \neq 0$ then $\tau\zeta > -2$, hence $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq \tau\zeta > -2$ is a stronger condition than that needed for stability of the mEE method, i.e. $\tau\Phi_N^{EE}(\Delta\tau\lambda)(\lambda + \zeta) \geq -2$. The consequence of this is a stronger and ζ -dependent condition on N which prevents η to decrease with τ , see Remark 4.4. As the numerical error depends on the size of η , see Theorem 4.11, the mEM method does not converge. In contrast, we see in Figure 1.4(a) that for stability of the SK-ROCK scheme it is sufficient that $p \in [-\ell_s^\varepsilon, 0]$ and $(p, q) \in \mathcal{S}^{MS}$, where $p \in [-\ell_s^\varepsilon, 0]$ is the stability condition for the RKC scheme. As a consequence, the mSK-ROCK method of Section 4.3 enjoys the same stability conditions of the mRKC scheme and η is allowed to decrease with τ .

4.2.4 Convergence analysis

Here we analyze the strong and weak errors of the multirate Euler–Maruyama method (3.20), (3.21) and (4.14) to (4.17). We will see that the errors depend on η and since η is bounded from below, see Remark 4.4, convergence is lost. We suppose that f_F, f_S, g are Lipschitz continuous, hence that there exist a constant $C > 0$ such that for all $x, y \in \mathbb{R}^n$ it holds

$$|f_F(x) - f_F(y)| + |f_S(x) - f_S(y)| + |g(x) - g(y)| \leq C|x - y|. \quad (4.23)$$

Therefore, f_F, f_S, g have linear growth and satisfy

$$|f_F(x)| + |f_S(x)| + |g(x)| \leq C(1 + |x|) \quad (4.24)$$

for some $C > 0$. Indeed, we denote by C a positive constant which may change from line to line and depends on the data, but is always independent from the step sizes $\tau, \Delta\tau$ and the step indices n, m . Furthermore, we will denote by $C_p^4(\mathbb{R}^n, \mathbb{R}^n)$ the space of functions from \mathbb{R}^n to \mathbb{R}^n four times continuously differentiable having derivatives with at most polynomial growth. The next convergence analysis is purely technical and therefore most of the proofs are postponed to Section 4.5.1.

Lemma 4.10. *The solution X_n of the mEM method (4.14) satisfies*

$$X_{n+1} = X_n + \tau f(X_n) + g(X_n)\Delta W_n + R, \quad (4.25)$$

with $|\mathbb{E}(R|X_n)| \leq C(1 + |X_n|)\eta\tau$ and $\mathbb{E}(|R|^2|X_n)^{1/2} \leq C(1 + |X_n|)\eta\tau^{1/2}$. Moreover, for $\psi \in C_p^4(\mathbb{R})$ it holds

$$\mathbb{E}(\psi(X_{n+1})|X_n) = \psi(X_n) + \tau\psi'(X_n)f(X_n) + \frac{1}{2}\tau\psi''(X_n)g(X_n)^2 + R_w, \quad (4.26)$$

with $|\mathbb{E}(R_w|X_n)| \leq C(1 + |X_n|^q)(\eta + \eta^2 + \tau)\tau$.

The next theorem bounds the error committed by the mEM method.

Theorem 4.11. *The mEM method has mean-square error*

$$\mathbb{E}(|X(t_n) - X_n|^2)^{1/2} \leq C(1 + \mathbb{E}(|X_0|^2))^{1/2}(\eta + \tau^{1/2}) \quad (4.27)$$

and if $f_F, f_S, g \in C_p^4(\mathbb{R})$ then the following weak convergence error holds

$$|\mathbb{E}(\psi(X(t_n)) - \mathbb{E}(\psi(X_n)))| \leq C(1 + \mathbb{E}(|X_0|^q))(\eta + \eta^2 + \tau)$$

for all $\psi \in C_p^4(\mathbb{R})$.

Proof. As f_F, f_S, g are Lipschitz continuous, by the Itô formula applied to (4.1) with initial value $X(t_n) = X_n$ we obtain

$$X(t_{n+1}) = X_n + \tau f(X_n) + g(X_n)\Delta W_n + \bar{R},$$

with $|\mathbb{E}(\bar{R}|X_0)| \leq C(1 + |X_0|)\tau^{3/2}$ and $\mathbb{E}(|\bar{R}|^2|X_0)^{1/2} \leq C(1 + |X_0|)\tau$. Therefore, it follows from Lemma 4.19 that the local errors satisfy

$$\begin{aligned} |\mathbb{E}(X_{n+1} - X(t_{n+1})|X_n)| &\leq C(1 + |X_n|)(\eta + \tau^{1/2})\tau, \\ \mathbb{E}(|X_{n+1} - X(t_{n+1})|^2|X_n)^{1/2} &\leq C(1 + |X_n|)(\eta + \tau^{1/2})\tau^{1/2}. \end{aligned}$$

Estimate (4.27) follows from the classical result [93, Theorem 1.1], which asserts the global order of convergence from the local errors.

Now, let f_F, f_S, g and $\psi \in C_p^4(\mathbb{R})$, applying Itô formula to $\psi(X(t_{n+1}))$ yields

$$\mathbb{E}(\psi(X(t_{n+1}))|X_n) = \psi(X_n) + \tau\psi'(X_n)f(X_n) + \frac{1}{2}\tau\psi''(X_n)g(X_n)^2 + \bar{R}_w,$$

with $|\mathbb{E}(\bar{R}_w|X_n)| \leq C(1 + |X_n|^q)\tau^2$. Hence, from (4.26) follows

$$|\mathbb{E}(\psi(X(t_{n+1})) - \psi(X_{n+1})|X_n)| \leq C(1 + |X_n|^q)(\eta + \eta^2 + \tau)\tau.$$

In Lemmas 4.22 and 4.23 we show $|\bar{f}_\eta(u_0)| \leq C(1 + |u_0|)$ and $|\bar{g}_\eta(v_0)| \leq C(1 + |v_0|)$, therefore

$$\begin{aligned} |\mathbb{E}(X_{n+1} - X_n|X_n)| &\leq C(1 + |X_n|)\tau, \\ |X_{n+1} - X_n| &\leq M_n(1 + |X_n|)\tau^{1/2}, \end{aligned}$$

with M_n having moments of all orders. Thus the mEM method satisfies [93, Lemma 2.2] and all the conditions of [93, Theorem 2.1] are met, from which follows the weak convergence result. ■

Therefore, we see from Theorem 4.11 that the mEM method converges, strongly or weakly, only if $\eta \rightarrow 0$. However, we saw in Remark 4.4 that η is bounded from below by $2/\rho_S$, as further explained in Remark 4.9. We will see that the mSK-ROCK method does not suffer from this deficiency and converges with strong order 1/2 and weak order 1.

4.3 The multirate SK-ROCK method

In this section we finally introduce the mSK-ROCK scheme and study its efficiency, mean-square stability and accuracy properties. The mSK-ROCK method is obtained by discretizing the modified equation (4.2) with the SK-ROCK scheme of Section 1.4.1 and as for the mRKC method f_η is replaced by the numerical approximation \bar{f}_η given in (3.36) and (3.37), furthermore we also replace g_η by a numerical approximation \bar{g}_η computed by discretizing (4.8) with a modified RKC scheme.

4.3.1 The mSK-ROCK algorithm

We define here the mSK-ROCK method, for the time being we restrict ourselves to a vector valued diffusion term $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and explain at the end of this section how to extend the method to a matrix valued diffusion $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$.

Let $s, m \in \mathbb{N}$ and $\eta > 0$ be as for the mRKC method in (3.34), i.e.

$$\tau\rho_S \leq \beta s^2, \quad \eta\rho_F \leq \beta m^2, \quad \text{with} \quad \eta = \frac{6\tau}{\beta s^2} \frac{m^2}{m^2 - 1}, \quad (4.28)$$

but with the additional constraint that m must be even, we will denote $m = 2r$ with $r \in \mathbb{N}$. One step of the mSK-ROCK method is given by

$$\begin{aligned} K_0 &= X_n, \\ K_1 &= K_0 + \mu_1\tau\bar{f}_\eta(K_0 + \nu_1\bar{Q}_\eta) + \kappa_1\bar{Q}_\eta, \\ K_j &= \nu_j K_{j-1} + \kappa_j K_{j-2} + \mu_j\tau\bar{f}_\eta(K_{j-1}), \quad j = 2, \dots, s, \\ X_{n+1} &= K_s, \end{aligned} \quad (4.29)$$

with $\bar{Q}_\eta = \bar{g}_\eta(K_0)\Delta W_n$, \bar{f}_η as in (3.36) and (3.37) and μ_j, ν_j, κ_j as in (1.20) and (1.39). We define \bar{g}_η as a numerical counterpart of g_η , hence from (4.9) follows

$$\bar{g}_\eta(v_0) = \frac{1}{\eta}(v_\eta - \bar{v}_\eta), \quad (4.30)$$

where v_η and \bar{v}_η are approximations of $v(\eta)$ and $\bar{v}(\eta)$, respectively. We compute v_η using a modified r -stage RKC scheme

$$\begin{aligned} v_1 &= v_0 + \alpha_1 \eta f_F(v_0 + \beta_1 \theta_1 \eta g(v_0)) + \gamma_1 \theta_1 \eta g(v_0), \\ v_j &= \beta_j v_{j-1} + \gamma_j v_{j-2} + \alpha_j \eta f_F(v_{j-1}), \quad j = 2, \dots, r, \\ v_\eta &= v_r, \end{aligned} \quad (4.31)$$

while \bar{v}_η is given by

$$\begin{aligned} \bar{v}_1 &= \bar{v}_0 + \alpha_1 \eta f_F(\bar{v}_0), \\ \bar{v}_j &= \beta_j \bar{v}_{j-1} + \gamma_j \bar{v}_{j-2} + \alpha_j \eta f_F(\bar{v}_{j-1}), \quad j = 2, \dots, r, \\ \bar{v}_\eta &= \bar{v}_r, \end{aligned} \quad (4.32)$$

with initial condition $\bar{v}_0 = v_0$.

In (4.31) and (4.32) the parameters α_1 and $\alpha_j, \beta_j, \gamma_j$ for $j = 2, \dots, r$ are the parameters of the m -stage RKC scheme given in (3.39), with $m = 2r$. Hence, the same parameters used to compute u_η in (3.37) are also used to compute v_η, \bar{v}_η ; with the difference that for v_η, \bar{v}_η the algorithm stops at stage r , and not at stage $m = 2r$ as for u_η . The additional parameters $\beta_1, \gamma_1, \theta_1$ needed in (4.31) are given by

$$\beta_1 = m \frac{v_1}{2}, \quad \gamma_1 = m \frac{v_1}{v_0}, \quad \theta_1 = \frac{1}{2v_1} \frac{T_r(v_0)}{T'_r(v_0)}. \quad (4.33)$$

Now, we discuss the case where $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ is a matrix valued function and therefore (4.31) is not well-defined. One possible approach is to compute (4.31) for each column of g and build a modified matrix g_η column-wise. However, this way of proceeding entails the computation of (4.31) for each column of g , which can rapidly become expensive. A better solution is to replace $\eta g(v_0)$ in (4.31) by $\eta g(v_0)\Delta W_n$, which is vector valued and therefore (4.30) to (4.32) can be computed. Then we replace \bar{Q}_η in (4.29) by $\bar{Q}_\eta = \bar{g}_\eta(v_0)$, as ΔW_n is already contained in $\bar{g}_\eta(v_0)$. With this second approach, (4.31) is computed only once and therefore the cost of stabilizing a vector or a matrix valued diffusion term g is equivalent. Note that when f_F is linear the two approaches give exactly the same result \bar{g}_η . When f_F is nonlinear we obtain two slightly different methods, nonetheless we can show that both have the same accuracy and mean-square stability properties.

Stability of the mSK-ROCK method is proved in Theorem 4.15, in Theorem 4.20 we prove that it has stronger order 1/2 and weak order 1.

4.3.2 Efficiency of the mSK-ROCK method

Given the spectral radii ρ_F, ρ_S of the Jacobians of f_F, f_S , respectively, we want to compare the theoretical efficiency, in terms of function evaluations, of the mSK-ROCK and SK-ROCK method. We set $\varepsilon = 0$ and let s, m vary in \mathbb{R} . The cost of evaluating f_F, f_S, g relatively to the cost of evaluating $f_F + f_S + g$ is denoted $c_F, c_S, c_g \in [0, 1]$, respectively, with $c_F + c_S + c_g = 1$.

One step of the SK-ROCK scheme (1.40) needs s evaluations of $f_F + f_S$ and one of g , with $s = \sqrt{\tau(\rho_F + \rho_S)/2}$. Hence, the cost of one step of SK-ROCK is

$$C_{\text{SK-ROCK}} = s(c_F + c_S) + c_g = (s-1)(c_F + c_S) + 1 = \left(\sqrt{\frac{\tau(\rho_F + \rho_S)}{2}} - 1 \right) (c_F + c_S) + 1,$$

where we used $c_g = 1 - c_F - c_S$. In contrast, one step of mSK-ROCK requires s evaluations of \bar{f}_η and one of \bar{g}_η . Each evaluation of \bar{f}_η needs m evaluations of f_F and one of f_S , an evaluation of \bar{g}_η requires $2r = m$ evaluations of f_F and one of g . Hence the cost of one step of mSK-ROCK is given by

$$C_{\text{mSK-ROCK}} = s(mc_F + c_S) + mc_F + c_g = ((s+1)m-1)c_F + (s-1)c_S + 1.$$

Conditions (4.28) with $\beta = 2$ yield $s = \sqrt{\tau\rho_S/2}$ and $m = \sqrt{3\rho_F/\rho_S + 1}$, thus

$$C_{\text{mSK-ROCK}} = \left(\left(\sqrt{\frac{\tau\rho_S}{2}} + 1 \right) \sqrt{3\frac{\rho_F}{\rho_S} + 1} - 1 \right) c_F + \left(\sqrt{\frac{\tau\rho_S}{2}} - 1 \right) c_S + 1.$$

Let $p_F = \tau\rho_F$ and $p_S = \tau\rho_S$, the theoretical relative speed-up S is defined as the ratio between the two costs:

$$S = \frac{C_{\text{SK-ROCK}}}{C_{\text{mSK-ROCK}}} = \frac{(\sqrt{p_F + p_S} - \sqrt{2})(c_F + c_S) + \sqrt{2}}{\left((\sqrt{p_S} + \sqrt{2}) \sqrt{3\frac{p_F}{p_S} + 1} - \sqrt{2} \right) c_F + (\sqrt{p_S} - \sqrt{2}) c_S + \sqrt{2}}.$$

For some values of p_F, p_S we display S in function of c_F, c_S in Figure 4.5, with $c_F + c_S \in [0, 1]$. On the line $c_F + c_S = 1 - c_g$, with $c_g \in [0, 1]$ fixed, the speed-up increases as $c_S \rightarrow 1$; indeed, SK-ROCK needs more evaluations of f_S . In contrast, the mSK-ROCK method is slower than SK-ROCK ($S < 1$) if c_F is not sufficiently small, as it needs more evaluations of f_F . However, we recall that we intend to use the mSK-ROCK method when f_F is cheap to evaluate, otherwise we simply use the SK-ROCK scheme. Finally, if $c_g \rightarrow 1$ and thus $c_F + c_S \rightarrow 0$ then the cost of the two schemes is comparable; the effort of evaluating g becomes dominant and both methods have the same number of g evaluations, one each.

4.3.3 Mean-square stability analysis

We show here that when the mSK-ROCK method (3.36), (3.37) and (4.29) to (4.32) is applied to the stochastic multirate test equation (4.3) the scheme is stable.

In order to analyze the stability of the mRKC method in Section 3.4.3 we computed a closed expression for \bar{f}_η in (3.54) from an expression for u_η given by Lemma 3.16. In the next lemma, we will compute an expression for \bar{g}_η .

Lemma 4.12. *Let $\lambda, \mu \in \mathbb{R}$, $f_F(x) = \lambda x$, $g(x) = \mu x$, $\eta > 0$, $r \in \mathbb{N}$ and $v_0 \in \mathbb{R}$. Then, the modified diffusion \bar{g}_η (4.30), is given by*

$$\bar{g}_\eta(v_0) = \Psi_r(\eta\lambda)\mu v_0, \quad (4.34)$$

where

$$\Psi_r(z) = \frac{U_{r-1}(v_0 + v_1 z)}{U_{r-1}(v_0)} \left(1 + \frac{v_1}{2} z \right), \quad (4.35)$$

v_0, v_1 are as in (3.38) and $U_j(x)$ as in (1.26).

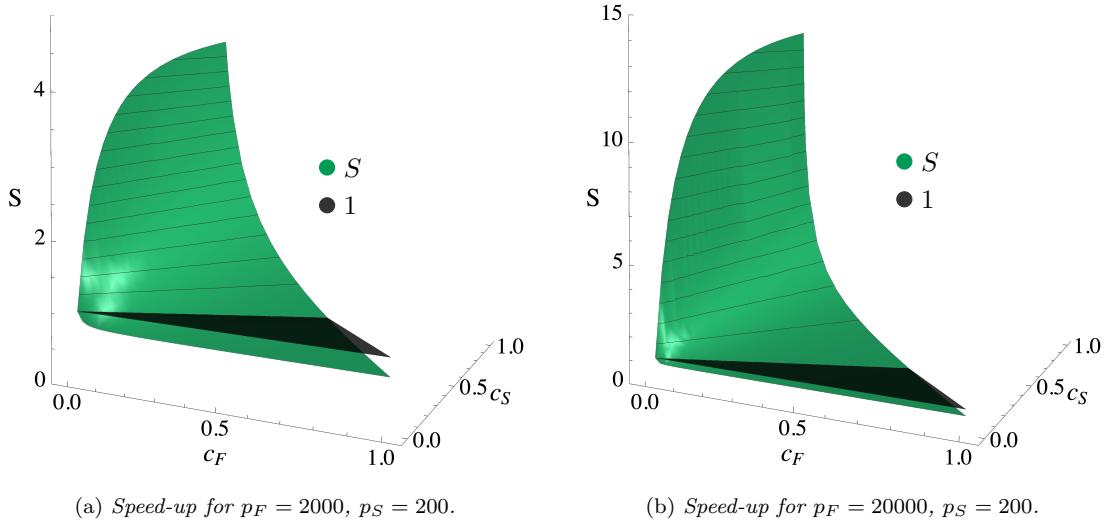


Figure 4.5. The relative speed-up S of mSK-ROCK over SK-ROCK, for some fixed values of $p_F = \tau\rho_F$ and $p_S = \tau\rho_S$.

Proof. Replacing $f_F(x) = \lambda x$ and $g(x) = \mu x$ in (4.31) yields

$$\begin{aligned} v_1 &= v_0 + \alpha_1 z v_0 + r_1, \\ v_j &= \beta_j v_{j-1} + \gamma_j v_{j-2} + \alpha_j z v_{j-1}, \quad j = 2, \dots, r, \end{aligned} \tag{4.36}$$

with $z = \eta\lambda$ and $r_1 = (\alpha_1\beta_1 z + \gamma_1)\theta_1\eta\mu v_0$. Scheme (4.36) is a perturbed RKC scheme, hence from (1.24), (1.25) and $v_\eta = v_r$ we deduce

$$\begin{aligned} v_\eta &= a_r T_r(v_0 + v_1 z) v_0 + \frac{a_r}{a_1} U_{r-1}(v_0 + v_1 z) r_1 \\ &= a_r T_r(v_0 + v_1 z) v_0 + \frac{a_r}{a_1} U_{r-1}(v_0 + v_1 z) (\alpha_1\beta_1 z + \gamma_1)\theta_1\eta\mu v_0. \end{aligned} \tag{4.37}$$

Using (4.33), $a_j = 1/T_j(v_0)$, $v_1 = T_m(v_0)/T'_m(v_0)$, $\alpha_1 = v_1/v_0$, $T_1(v_0) = v_0$ and the identity $T'_n(x) = nU_{n-1}(x)$, we compute

$$\begin{aligned} \frac{a_r}{a_1} (\alpha_1\beta_1 z + \gamma_1)\theta_1 &= \frac{T_1(v_0)}{T_r(v_0)} \left(m \frac{v_1^2}{2v_0} z + m \frac{v_1}{v_0} \right) \frac{1}{2v_1} \frac{T_r(v_0)}{T'_r(v_0)} = \frac{m}{2} \frac{1}{T'_r(v_0)} \left(1 + \frac{v_1}{2} z \right) \\ &= \frac{r}{T'_r(v_0)} \left(1 + \frac{v_1}{2} z \right) = \frac{1}{U_{r-1}(v_0)} \left(1 + \frac{v_1}{2} z \right). \end{aligned}$$

Therefore, from (4.37), we obtain

$$v_\eta = a_r T_r(v_0 + v_1 z) v_0 + \Psi_r(z) \eta\mu v_0, \tag{4.38}$$

where $\Psi_r(z)$ is given in (4.35). We have as well $\bar{v}_\eta = a_r T_r(v_0 + v_1 z) v_0$ and we obtain (4.34) by (4.30) and (4.38). \blacksquare

Letting $\Delta W_n = \tau^{1/2}\xi_n$ with $\xi_n \sim \mathcal{N}(0, 1)$ and plugging (3.54) and (4.34) into (4.29) yields (1.43) with p, q replaced by

$$p_m = \tau\Phi_m(\eta\lambda)(\lambda + \zeta), \quad q_r = \Psi_r(\eta\lambda)\mu\tau^{1/2}, \tag{4.39}$$

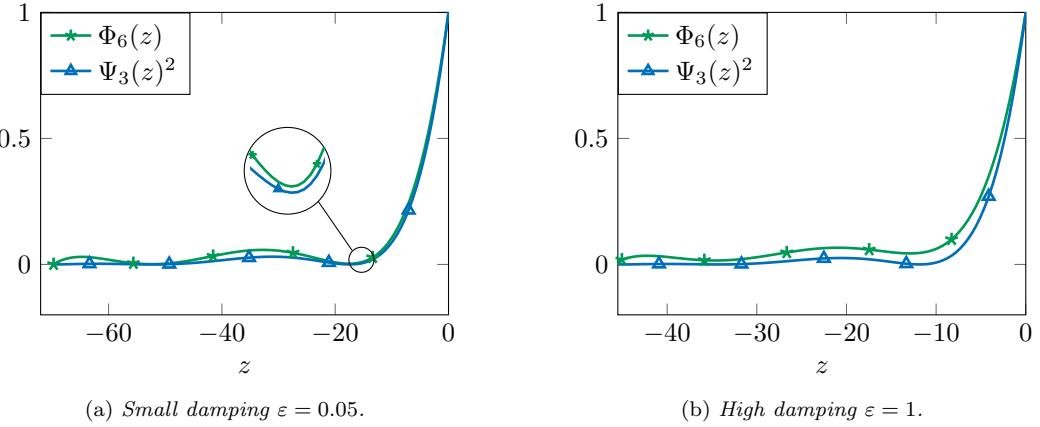


Figure 4.6. Illustration of $\Phi_m(z)$ and $\Psi_r(z)^2$ for $2r = m$ and $r = 3$, with damping $\varepsilon = 0.05$ (left) or $\varepsilon = 1$ (right).

respectively. Therefore, from (1.44), follows

$$X_{n+1} = (A_s(p_m) + B_s(p_m)q_r\xi)X_n,$$

which motivates the definition here below.

Definition 4.13. Let $\tau > 0$, $\eta \geq 0$, $\lambda, \zeta \leq 0$, $\mu \in \mathbb{R}$ and $\xi \sim \mathcal{N}(0, 1)$ a Gaussian random variable. The stability polynomial of the (s, m) -stage mSK-ROCK method is defined as

$$R_{s,m}(\tau, \eta, \lambda, \zeta, \mu, \xi) = A_s(p_m) + B_s(p_m)q_r\xi,$$

with A_s , B_s as in (1.45), p_m , q_r as in (4.39), $2m = r$ and Φ_m , Ψ_r as in (3.49) and (4.35).

Note that $R_{s,m}(\tau, \eta, \lambda, \zeta, \mu, \xi) = R_s(p_m, q_r, \xi)$, where $R_s(p, q, \xi)$ is the stability polynomial of the SK-ROCK scheme, see (1.44).

The next lemma is the numerical counterpart of Lemma 4.1 and therefore it is the main tool for proving stability of the scheme in Theorem 4.15 below. We postpone its proof to Section 4.5.2.

Lemma 4.14. Let $m, r \in \mathbb{N}$ with $2r = m$ and the damping parameter $\varepsilon = 0$. Then $\Psi_r(z)^2 \leq \Phi_m(z)$ for all $z \in [-\ell_m^\varepsilon, 0]$.

Numerical evidences show that Lemma 4.14 is valid for any damping parameter $\varepsilon > 0$. Indeed, we display $\Phi_m(z)$ and $\Psi_r(z)^2$ for $r = 3$ in Figure 4.6, for a small damping $\varepsilon = 0.05$ and a high damping $\varepsilon = 1$. In both cases relation $\Psi_r(z)^2 \leq \Phi_m(z)$ holds and is tight.

Theorem 4.15. Let the damping $\varepsilon = 0$ and $(\lambda + \zeta, \mu) \in \mathcal{S}^{MS}$ with $\lambda \leq 0$ and $\zeta < 0$. Then, for all τ, s, m and η with $m = 2r$ satisfying (3.56), i.e.

$$\tau|\zeta| \leq \ell_s^\varepsilon, \quad \eta|\lambda| \leq \ell_m^{\varepsilon_m}, \quad \text{with} \quad \eta \geq \frac{6\tau}{\ell_s^\varepsilon} \frac{m^2}{m^2 - 1}, \quad (4.40)$$

it holds $\mathbb{E}(|R_{s,m}(\tau, \eta, \lambda, \zeta, \mu, \xi)|^2) \leq 1$, i.e. the mSK-ROCK method is stable.

Proof. Since $R_{s,m}(\tau, \eta, \lambda, \zeta, \mu, \xi) = R_s(p_m, q_r, \xi)$, where $R_s(p, q, \xi)$ is the stability polynomial of the SK-ROCK scheme, we prove stability of the mSK-ROCK method showing that p_m, q_r satisfy the stability conditions for the SK-ROCK scheme, given in Theorem 1.1. Hence, we need to show

$$p_m \in [-\ell_s^\varepsilon, 0] \quad \text{and} \quad (p_m, q_r) \in \mathcal{S}^{MS}.$$

In Theorem 3.20 it was already proved that conditions (4.40) imply $p_m \in [-\ell_s^\varepsilon, 0]$. Hence we need only to prove $(p_m, q_r) \in \mathcal{S}^{MS}$, which holds if, and only if,

$$p_m + \frac{1}{2}|q_r|^2 \leq 0 \iff \Phi_m(\eta\lambda)(\lambda + \zeta) + \frac{1}{2}\Psi_r(\eta\lambda)^2|\mu|^2 \leq 0.$$

From Lemma 4.14 it follows $\Psi_r(\eta\lambda)^2 \leq \Phi_m(\eta\lambda)$ and thus

$$\Phi_m(\eta\lambda)(\lambda + \zeta) + \frac{1}{2}\Psi_r(\eta\lambda)^2|\mu|^2 \leq \Phi_m(\eta\lambda) \left(\lambda + \zeta + \frac{1}{2}|\mu|^2 \right) \leq 0, \quad (4.41)$$

as $\Phi_m(\eta\lambda) \geq 0$ (see Theorem 3.20) and $(\lambda + \zeta, \mu) \in \mathcal{S}^{MS}$. \blacksquare

Even if Theorem 4.15 is stated for $\varepsilon = 0$ numerical evidences show that it is valid for any damping $\varepsilon > 0$; indeed, Theorems 1.1 and 3.20 and Lemma 4.14 hold for $\varepsilon > 0$.

We see in (4.41) that stability of the mSK-ROCK scheme lies on the inequality $\Psi_r(z)^2 \leq \Phi_m(z)$. If for the integration of (4.8) we naively used a standard RKC method, instead of the modified RKC method (4.31), then we would obtain $q_r = \Phi_r(z/2)\mu\tau^{1/2}$ for some r satisfying the stability conditions of (4.8) (instead of $q_r = \Psi_r(z)\mu\tau^{1/2}$ with $2r = m$). Hence, stability of the scheme would hold only if $\Phi_r(z/2)^2 \leq \Phi_m(z)$ was true (as for $\varphi(z)$ in Lemma 4.1). However, such relation is not satisfied and a modified RKC method must be used in order to replace $\Phi_r(z/2)$ by a different polynomial $\Psi_r(z)$ satisfying $\Psi_r(z)^2 \leq \Phi_m(z)$. Observe that the $1/2$ factor in (4.8) disappears from (4.31) and (4.32) but is reflected on the fact that we take $r = m/2$ stages.

4.3.4 Convergence analysis

We prove here that the mSK-ROCK method has strong order $1/2$ and weak order 1. As in Section 4.2.4 we suppose that f_F, f_S, g are Lipschitz continuous and satisfy a linear growth condition, i.e. they satisfy (4.23) and (4.24). We recall as well that we denote by $C_p^4(\mathbb{R}^n, \mathbb{R}^n)$ the space of functions from \mathbb{R}^n to \mathbb{R}^n four times continuously differentiable having derivatives with at most polynomial growth. We start the convergence analysis showing a few technical lemmas, most of the proofs are purely technical and postponed to Section 4.5.2.

Lemma 4.16. *There exists $C > 0$ such that*

$$|\bar{f}_\eta(x) - \bar{f}_\eta(y)| + |\bar{g}_\eta(x) - \bar{g}_\eta(y)| \leq C|x - y|, \quad (4.42)$$

$$|\bar{f}_\eta(x) - f(x)| + |\bar{g}_\eta(x) - g(x)| \leq C(1 + |x|)\eta \quad (4.43)$$

for all $x, y \in \mathbb{R}^n$.

Lemma 4.17. *If $f_F, f_S, g \in C_p^4(\mathbb{R})$ then $\bar{f}_\eta, \bar{g}_\eta \in C_p^4(\mathbb{R})$.*

Lemma 4.18. *The stages K_j and \bar{Q}_η of (4.29) satisfy the estimate*

$$|\bar{Q}_\eta| + |K_j - X_n| \leq C(1 + |X_n|)(\tau + |\Delta W_n|), \quad (4.44)$$

$$|\mathbb{E}(\bar{Q}_\eta|X_n)| + |\mathbb{E}(K_j - X_n|X_n)| \leq C(1 + |X_n|)\tau \quad (4.45)$$

for $j = 1, \dots, s$.

Lemma 4.19. *The solution X_{n+1} of (4.29) satisfies*

$$X_{n+1} = X_n + \tau f(X_n) + g(X_n)\Delta W_n + R,$$

with $|\mathbb{E}(R|X_n)| \leq C(1 + |X_n|)\tau^{3/2}$ and $\mathbb{E}(|R|^2|X_n|^{1/2}) \leq C(1 + |X_n|)\tau^{3/2}$. If, furthermore, $f_F, f_S, g \in C_p^4(\mathbb{R}^n, \mathbb{R}^n)$, then $|\mathbb{E}(R|X_n)| \leq C(1 + |X_n|^q)\tau^2$.

Proof. Considering (4.29) we let $r_1 = \mu_1 \tau \bar{f}_\eta(K_0 + \nu_1 \bar{Q}_\eta) + \kappa_1 \bar{Q}_\eta$ and $r_j = \mu_j \tau \bar{f}_\eta(K_{j-1})$ for $j = 2, \dots, s$. From (1.25) with $z = 0$ we obtain

$$\begin{aligned} X_{n+1} &= X_n + \sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) r_k \\ &= X_n + \frac{b_s}{b_1} U_{s-1}(\omega_0) \kappa_1 \bar{Q}_\eta + \tau \sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k \bar{f}_\eta(\tilde{K}_{k-1}), \end{aligned}$$

where we let $\tilde{K}_0 = K_0 + \nu_1 \bar{Q}_\eta$ and $\tilde{K}_k = K_k$ for $k = 1, \dots, s-1$. Since $\frac{b_s}{b_1} U_{s-1}(\omega_0) \kappa_1 = 1$, $\sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k = 1$ and $\bar{Q}_\eta = \bar{g}_\eta(X_n) \Delta W_n$, we can write

$$\begin{aligned} X_{n+1} &= X_n + \tau \bar{f}_\eta(X_n) + \bar{g}_\eta(X_n) \Delta W_n + \tau \sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k (\bar{f}_\eta(\tilde{K}_{k-1}) - \bar{f}_\eta(X_n)) \\ &= X_n + \tau f(X_n) + g(X_n) \Delta W_n + R, \end{aligned}$$

with

$$\begin{aligned} R &= R_1 + R_2 + R_3, & R_1 &= \tau (\bar{f}_\eta(X_n) - f(X_n)), \\ R_2 &= (\bar{g}_\eta(X_n) - g(X_n)) \Delta W_n, & R_3 &= \tau \sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k (\bar{f}_\eta(\tilde{K}_{k-1}) - \bar{f}_\eta(X_n)). \end{aligned}$$

From (4.43),

$$|R_1|^2 \leq C(1 + |X_n|)^2 \eta^2 \tau^2, \quad |R_2|^2 \leq C(1 + |X_n|)^2 \eta^2 \Delta W_n^2. \quad (4.46)$$

Since $\frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k \geq 0$ and $\sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k = 1$, using (4.42) we obtain

$$\begin{aligned} |R_3|^2 &\leq \tau^2 \max_{k=1, \dots, s} |\bar{f}_\eta(\tilde{K}_{k-1}) - \bar{f}_\eta(X_n)|^2 \leq C \tau^2 \max_{k=1, \dots, s} |\tilde{K}_{k-1} - X_n|^2 \\ &\leq C(1 + |X_n|)^2 \tau^2 (\tau + |\Delta W_n|)^2, \end{aligned} \quad (4.47)$$

where we used (4.44) and thus $|\tilde{K}_{k-1} - X_n| \leq C(1 + |X_n|)(\tau + |\Delta W_n|)$. From $\mathbb{E}(R_2|X_n) = 0$, (4.46) and (4.47), we get

$$\begin{aligned} |\mathbb{E}(R|X_n)| &\leq C(1 + |X_n|)(\eta + \tau^{1/2})\tau, \\ \mathbb{E}(|R|^2|X_n) &\leq C(1 + |X_n|)^2 (\eta^2 \tau + \eta^2 + \tau^2)\tau. \end{aligned}$$

We conclude using (4.28), from where we deduce $\eta \leq 8\tau$.

Now, if $f_F, f_S \in C_p^4(\mathbb{R})$ then from Lemma 4.17 we have $\bar{f}_\eta \in C_p^4(\mathbb{R})$ and we can improve the estimate on $|\mathbb{E}(R|X_n)|$. From (4.45) it holds $|\mathbb{E}(\tilde{K}_{k-1} - X_n|X_n)| \leq C(1 + |X_n|)\tau$ and thus

$$\begin{aligned} |\mathbb{E}(R_3|X_n)| &\leq \tau \sum_{k=1}^s \frac{b_s}{b_k} U_{s-k}(\omega_0) \mu_k |\mathbb{E}(\bar{f}_\eta(\tilde{K}_{k-1}) - \bar{f}_\eta(X_n)|X_n)| \\ &\leq \tau \max_{k=1, \dots, s} |\mathbb{E}(\bar{f}_\eta(\tilde{K}_{k-1}) - \bar{f}_\eta(X_n)|X_n)| \\ &\leq C\tau(1 + |X_n|^q) \max_{k=1, \dots, s} |\mathbb{E}(\tilde{K}_{k-1} - X_n|X_n)| \leq C(1 + |X_n|^q)\tau^2, \end{aligned} \quad (4.48)$$

where we used (4.44) to bound the derivative of \bar{f}_η in $[X_n, \tilde{K}_{k-1}]$. Hence, using (4.48) and $|\mathbb{E}(R_1|X_n)| \leq C(1 + |X_n|)\tau^2$ yields $|\mathbb{E}(R|X_n)| \leq C(1 + |X_n|^q)\tau^2$. ■

Theorem 4.20. *The mSK-ROCK method has strong order 1/2 and weak order 1, i.e.*

$$\mathbb{E}(|X(t_n) - X_n|^2)^{1/2} \leq C(1 + \mathbb{E}(|X_0|^2))^{1/2}\tau^{1/2} \quad (4.49)$$

and if $f_F, f_S, g \in C_p^4(\mathbb{R})$ then

$$|\mathbb{E}(\psi(X(t_n)) - \mathbb{E}(\psi(X_n)))| \leq C(1 + \mathbb{E}(|X_0|^q))\tau \quad (4.50)$$

for all $\psi \in C_p^4(\mathbb{R})$.

Proof. As f_F, f_S, g are Lipschitz continuous, by the Itô formula applied to (4.1) with initial value $X(t_n) = X_n$, we obtain

$$X(t_{n+1}) = X_n + \tau f(X_n) + g(X_n)\Delta W_n + \bar{R}$$

with $|\mathbb{E}(\bar{R}|X_n)| \leq C(1 + |X_n|)\tau^{3/2}$ and $\mathbb{E}(|\bar{R}|^2|X_n)^{1/2} \leq C(1 + |X_n|)\tau$. Therefore, it follows from Lemma 4.19 that the local errors satisfy

$$\begin{aligned} |\mathbb{E}(X_{n+1} - X(t_{n+1})|X_n)| &\leq C(1 + |X_n|)\tau^{3/2}, \\ \mathbb{E}(|X_{n+1} - X(t_{n+1})|^2|X_n)^{1/2} &\leq C(1 + |X_n|)\tau. \end{aligned}$$

The classical result [93, Theorem 1.1], which asserts the global order of convergence from the local error, implies estimate (4.49).

Now, let f_F, f_S, g and $\psi \in C_p^4(\mathbb{R})$, from Lemma 4.19 follows

$$|\mathbb{E}(\psi(X(t_{n+1})) - \psi(X_{n+1})|X_n)| \leq C(1 + |X_n|^q)\tau^2.$$

From (4.44) and (4.45) we deduce as well that the mSK-ROCK method satisfies [93, Lemma 2.2] and thus all the conditions of [93, Theorem 2.1] are met, from which follows the weak convergence result (4.50). \blacksquare

4.4 Numerical experiments

Here we verify the accuracy of the mSK-ROCK method of Section 4.3.1 and compare its computational cost against the cost of the original SK-ROCK scheme of Section 1.4.1. At first, we verify the strong and weak convergence properties of the mSK-ROCK scheme on a nonstiff problem, where we fix the number of stages beforehand. Then we do the same but on a stiff problem, letting the scheme automatically choose the number of stages in function of the spectral radii and the step size. The next two experiments investigate the computational efficiency of the scheme, first on a chemical Langevin equation and then on a stochastic heat equation with multiplicative noise. The last experiment has been performed with the help of the C++ library `libMesh` [79].

4.4.1 Nonstiff problem convergence experiment

We verify the convergence properties of the mSK-ROCK scheme on the following SDE, taken from [2],

$$dX(t) = \left(\frac{1}{4}X(t) + \frac{1}{2}\sqrt{X(t)^2 + 1} \right) dt + \sqrt{\frac{X(t)^2 + 1}{2}} dW(t), \quad X(0) = 0,$$

where the exact solution is $X(t) = \sinh(\frac{t}{2} + \frac{W(t)}{\sqrt{2}})$. We let $f_F(X) = \frac{1}{2}\sqrt{X^2 + 1}$ and $f_S(X) = \frac{1}{4}X$. Considering the step sizes $\tau = 2^{-k}$, for $k = 1, \dots, 10$, we display the strong $\mathbb{E}(|X(T) - X_N|^2)^{1/2}$ and weak $|\mathbb{E}(\text{asinh}(X(T)) - \mathbb{E}(\text{asinh}(X_N))|$ errors at time $T = 1 = N\tau$ in Figure 4.7, using 10^6 samples and $(s, m) = (5, 4)$ or $(s, m) = (10, 10)$. We observe that the method converges with the predicted orders of accuracy and the error is essentially independent of the stage number.

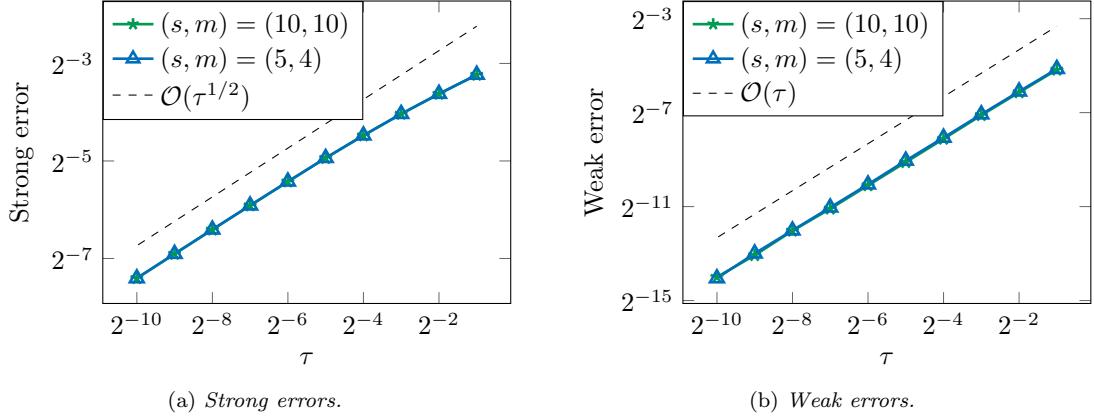


Figure 4.7. Nonstiff convergence experiment. Strong and weak errors of mSK-ROCK vs. the step size τ , for different stage choices.

4.4.2 Stiff problem convergence experiment

We consider a chemical Langevin model of dimerization reactions in a genetic network [31]. The model consists of 7 species and 10 reactions, described by the equations

$$dX(t) = \sum_{j=1}^m \nu_j a_j(X(t)) dt + \sum_{j=1}^m \nu_j \sqrt{a_j(X(t))} dW_j(t) \quad t \in (0, T], \quad X(0) = X_0, \quad (4.51)$$

where $m = 10$, $X(t) \in \mathbb{R}^7$, $T = 10$ and ν_j , $a_j(x)$ are derived as explained in Section 1.4.2 from the chemical reaction system introduced in [31]. We consider the same initial conditions as in [31] but multiplied by 10^3 , as the chemical Langevin model is accurate if there is an important number of reactants [70].

We order the reaction terms $\nu_j a_j(x)$ from the fastest to the slowest (the sequence ρ_j of the spectral radii of the Jacobians of $\nu_j a_j(x)$, evaluated on a typical path $X(t)$, is decreasing) and let

$$f_F(x) = \sum_{j=1}^3 \nu_j a_j(x), \quad f_S(x) = \sum_{j=4}^{10} \nu_j a_j(x),$$

hence f_F represent the three fastest reactions. We run the mSK-ROCK method over 10^5 Brownian paths with time steps $\tau = T/2^j$ for $j = 4, \dots, 10$ and measure the strong and weak errors committed against reference solutions computed on the same paths but using the SK-ROCK method with a time step $\tau = T/2^{12}$. As weak error we consider the error committed on the second moment of X_6 . Differently from Section 4.4.1 we let the mSK-ROCK method automatically choose the number of stages s, m . We observe in Figure 4.8 that the mSK-ROCK converges with the right orders and have similar errors as the SK-ROCK scheme.

4.4.3 E. Coli bacteria heat shock response

We consider a chemical Langevin equation modeling E. coli bacteria's protein denaturation under heat shocks. The original deterministic model is introduced in [80], while in [33, 74] it is considered as a chemical reaction system.

The model consists of 28 species and 61 reactions, it is described by (4.51) with $m = 61$ and $X(t) \in \mathbb{R}^{28}$. The initial condition is the same as in [74] but multiplied by 100 and we let $T = 10$.

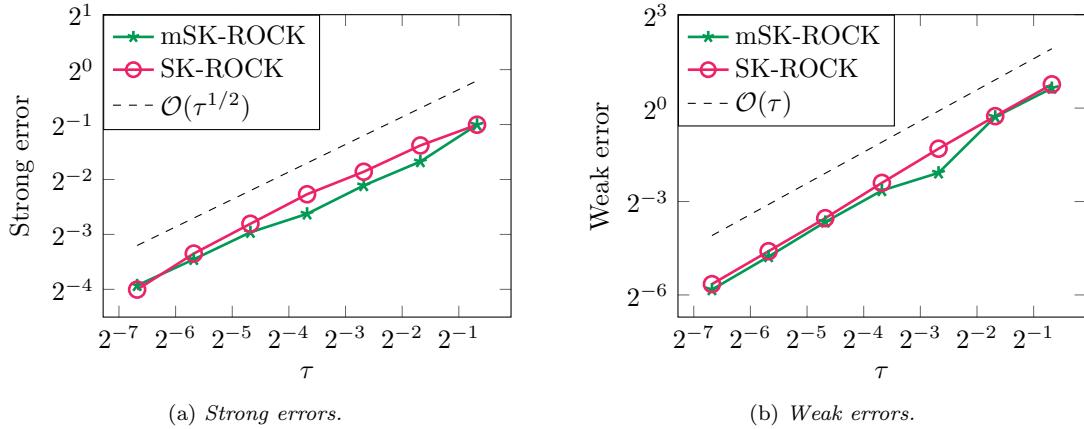


Figure 4.8. Stiff convergence experiment. Strong and weak errors of mSK-ROCK and SK-ROCK vs. step size τ .

The parameters $\nu_j, a_j(x)$ are derived as explained in Section 1.4.2 from the chemical reactions described in [74, Section 7.2] and the terms $\nu_j a_j(x)$ are ordered from the fastest to the slowest as explained in Section 4.4.2. For $r = 0, \dots, 10$ we define

$$f_F^r(x) = \sum_{j=1}^r \nu_j a_j(x), \quad f_S^r(x) = \sum_{j=r+1}^{61} \nu_j a_j(x),$$

hence f_F^r is defined by the r fastest reactions and f_S^r by the remaining ones. Observe that for $r = 0$ it holds $f_F^0 = 0$ and thus all the reactions are considered to be slow.

Let $\tau = T/2^{12}$ be fixed, for each value of $r = 0, \dots, 10$ we run the mSK-ROCK scheme and measure: the mean values of ρ_F, ρ_S, s, m along the integration interval and the code efficiency in terms of total multiplications needed to evaluate f_F^r and f_S^r . For $r = 0$ we have $f_F^0 = 0$ and thus the original SK-ROCK scheme is used with $f = f_S^0$. We display in Figures 4.9(a) and 4.9(b) the values of ρ_F, ρ_S and s, m , respectively. We see how ρ_S decreases as r increases and hence more fast reactions are put into f_F^r , as a consequence s decreases as well. In order to compensate the decreasing stabilization made by the “outer” scheme the “inner” method must increase the number of stages m , see Figure 4.9(b). In Figure 4.10(a) we show the cost of the scheme, defined as the total number of multiplications needed by mSK-ROCK in order to evaluate \bar{f}_η and \bar{g}_η . For $r = 0$ we have the cost of SK-ROCK and for $r = 1, \dots, 10$ the cost of mSK-ROCK. In Figure 4.10(b) we show the relative speed-up of mSK-ROCK with respect to SK-ROCK, defined as the cost of SK-ROCK ($r = 0$ in Figure 4.10(a)) divided by the cost of mSK-ROCK for $r = 1, \dots, 10$.

4.4.4 Diffusion across a narrow channel with multiplicative space-time noise

Here, we consider a stochastic heat equation with multiplicative noise defined in a domain which imposes local mesh refinement. We compare the efficiency of the mSK-ROCK and SK-ROCK method as the geometry constraints get more and more severe. The same problem is found in Section 3.5.3, but in the deterministic context.

We consider

$$\begin{aligned} du &= (\Delta u + b) dt + G(u) dW && \text{in } \Omega_\delta \times [0, T], \\ \nabla u \cdot \mathbf{n} &= 0 && \text{in } \partial\Omega_\delta \times [0, T], \\ u &= 0 && \text{in } \Omega_\delta \times \{0\}, \end{aligned} \tag{4.52}$$

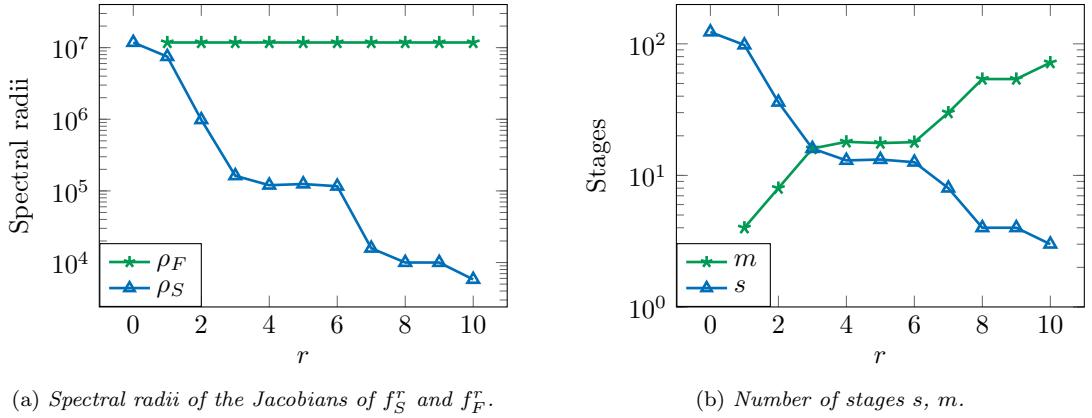


Figure 4.9. *E. Coli* experiment. Spectral radii and number of stages in function of the number r of fast reactions put into f_F^r .

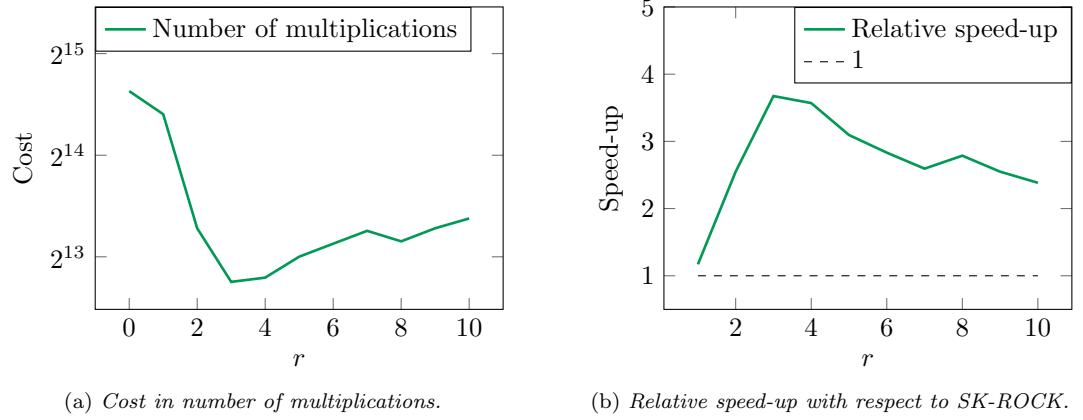


Figure 4.10. *E. Coli* experiment. Cost and relative speed-up of mSK-ROCK.

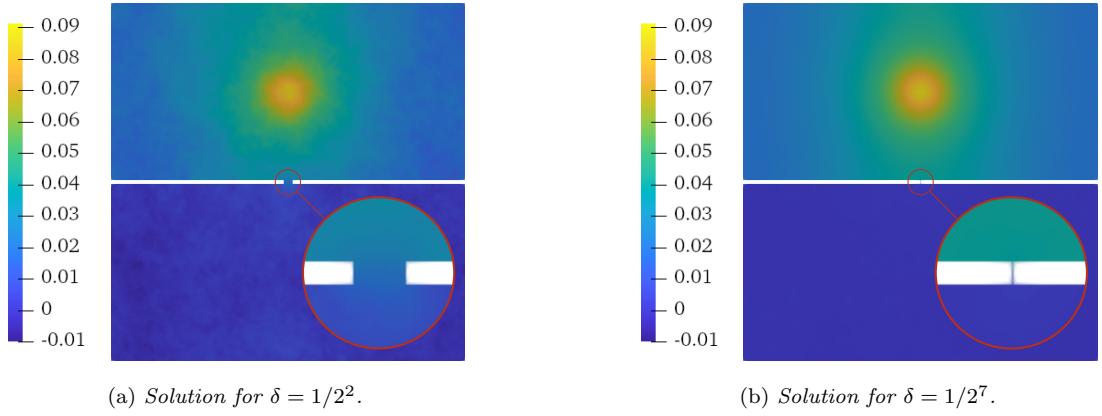


Figure 4.11. *Narrow channel*. Numerical solutions at $t = 10$ for a channel width $\delta = 1/2^2$ or $\delta = 1/2^7$.

where $T = 0.1$ and Ω_δ is a domain consisting in two 10×5 rectangles linked together by a narrow channel $\delta \times 0.05$ of width $\delta > 0$, see Figure 4.11. The source term $b(\mathbf{x}, t) = \sin(10\pi t)^2 e^{-5\|\mathbf{x}-\mathbf{c}\|^2}$ is a Gaussian centered in \mathbf{c} , the center of the upper rectangle in Ω_δ . We define $G : L^2(\Omega_\delta) \rightarrow \mathcal{L}(L^2(\Omega_\delta), L^2(\Omega_\delta))$ by $G(u)(v)(\mathbf{x}) = u(\mathbf{x})v(\mathbf{x})$ and $W(t)$ is a Q -Wiener process defined by a covariance operator $Q : L^2(\Omega_\delta) \rightarrow L^2(\Omega_\delta)$, i.e. $W(t)$ satisfies

$$\mathbb{E}(\langle W(t), h \rangle) = 0 \quad \text{and} \quad \mathbb{E}(\langle W(t), h \rangle^2) = t \langle Qh, h \rangle \quad (4.53)$$

for all $h \in L^2(\Omega_\delta)$, where $\langle \cdot, \cdot \rangle$ is the inner product in $L^2(\Omega_\delta)$. For $h \in L^2(\Omega_\delta)$ we define Q by

$$(Qh)(\mathbf{x}) = \int_{\Omega_\delta} q(\mathbf{x} - \mathbf{y})h(\mathbf{y}) d\mathbf{y}, \quad \text{where} \quad q(\mathbf{x}) = \frac{\alpha}{\pi} e^{-\alpha\|\mathbf{x}\|^2}$$

is an approximation of the Dirac delta function and $\alpha = 100$.

In Ω_δ , we define a Delaunay triangulation \mathcal{M} composed by simplicial elements having maximal size $H \approx 0.015$. Let $V = \text{span}\{\varphi_i : i = 1, \dots, N\}$ be a discontinuous Galerkin finite element (DG-FE) [39] space on \mathcal{M} and $\Delta_H : V \rightarrow V$ the DG-FE discretization of the Laplacian. Then, the semidiscrete problem corresponding to (4.52) is to find the process $u_H(t) = \sum_{i=1}^N u_i(t)\varphi_i$ satisfying

$$du_H = (\Delta_H u_H + P_H b) dt + P_H G(u_H) d\widehat{W}, \quad (4.54)$$

where $P_H : L^2(\Omega_\delta) \rightarrow V$ is the orthogonal projection operator and $\widehat{W}(t) \in V$ is the numerical counterpart of $W(t)$ in (4.53), hence it satisfies

$$\mathbb{E}(\langle \widehat{W}(t), h \rangle) = 0 \quad \text{and} \quad \mathbb{E}(\langle \widehat{W}(t), h \rangle^2) = t \langle Qh, h \rangle$$

for all $h \in V$. We set

$$\widehat{W}(t) = \sum_{i=1}^N \gamma_i^{1/2} e_i \beta_i(t),$$

where $\{e_i\}_{i=1}^N$ is an orthonormal basis of V , $\gamma_i \geq 0$ for $i = 1, \dots, N$ and $\{\beta_i(t)\}_{i=1}^N$ is a sequence of independently and identically distributed Brownian motions. Note that $\mathbb{E}(\langle \widehat{W}(t), h \rangle) = 0$ and since $\mathbb{E}(\langle \widehat{W}(t), e_i \rangle^2) = t\gamma_i$ we set

$$\gamma_i = \langle Qe_i, e_i \rangle = \int_{\Omega_\delta \times \Omega_\delta} q(\mathbf{x} - \mathbf{y}) e_i(\mathbf{x}) e_i(\mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

Observe that $\widehat{W}(t)$ is a Q -Wiener process in V with covariance operator Q_H defined by $Q_H e_i = \gamma_i e_i$.

Taking the inner product on both sides of (4.54) with respect to φ_j we obtain the equivalent equation

$$dX(t) = (AX(t) + M^{-1}\widehat{b}(t)) dt + M^{-1}\widehat{G}(X(t)) dB(t), \quad (4.55)$$

with $X(t) = (u_i(t))_{i=1}^N$, M the mass matrix, A the stiffness matrix, $B(t) = (\beta_i(t))_{i=1}^N$ an N -dimensional Wiener process and $\widehat{b}(t) \in \mathbb{R}^N$, $\widehat{G}(X) \in \mathbb{R}^{N \times N}$ are defined by

$$\widehat{b}_j(t) = \langle b(t), \varphi_j \rangle, \quad \widehat{G}(X)_{ji} = \gamma_i^{1/2} \langle G(u_H)(e_i), \varphi_j \rangle.$$

Letting

$$f(t, X) = AX + M^{-1}\widehat{b}(t), \quad g(X) = M^{-1}\widehat{G}(X)$$

we obtain (4.1a), in nonautonomous form. Note that the orthonormal basis $\{e_i\}_{i=1}^N$ can be computed locally on each element and M is easy to invert since it is block-diagonal. Therefore, the application of SK-ROCK to (4.55) leads to a truly explicit method.

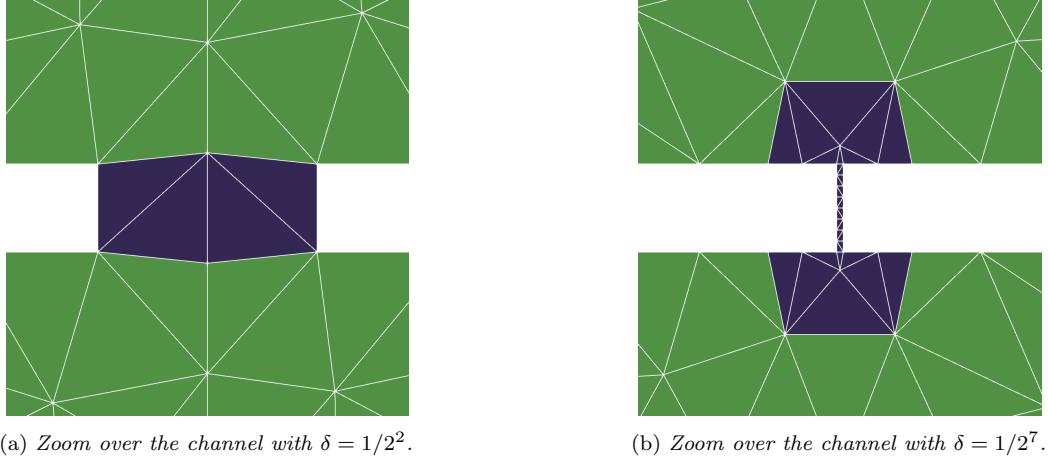


Figure 4.12. Narrow channel. Zoom of the FE mesh for channel width $\delta = 1/2^2$ or $1/2^7$, with the subdomain $\Omega_{F,\delta}$ (in blue).

We illustrate the triangulation \mathcal{M} in the neighborhood of the narrow channel in Figure 4.12, for two values of δ . We observe that for large δ the typical element size is small enough to resolve the channel (Figure 4.12(a)), while for small δ the elements in the channel are considerably smaller (Figure 4.12(b)). As the spectral radius of the discrete Laplacian behaves as $1/h^2$, where h is the size of the smallest elements in the mesh, then ρ increases as δ decreases, where ρ is the spectral radius of the Jacobian of f . Therefore, the cost of SK-ROCK applied to (4.55) increases as δ decreases.

Now, we want to decompose f in the two terms f_F and f_S such that as δ decreases then ρ_F increases but ρ_S remains constant, where ρ_F and ρ_S are the spectral radii of the Jacobians of f_F and f_S , respectively. We define a subdomain $\Omega_{F,\delta}$ consisting in the channel plus its neighboring elements having size smaller than the typical mesh size H , see Figure 4.12. Therefore, the size of the elements outside $\Omega_{F,\delta}$ is almost independent of δ . In order to identify f_F and f_S as the discrete Laplacian inside and outside of $\Omega_{F,\delta}$, respectively, we define a diagonal matrix $D \in \mathbb{R}^{N \times N}$ by $D_{jj} = 1$ if $\text{supp}(\varphi_j) \subset \Omega_{F,\delta}$ and $D_{jj} = 0$ else. We let

$$f_F(X) = DAX, \quad f_S(t, X) = (I - D)AX + M^{-1}\hat{b}(t),$$

with I the identity matrix. Thus, as δ decreases the size of the elements inside of $\Omega_{F,\delta}$ decrease and ρ_F increases, while ρ_S is independent of δ .

We will solve (4.55) for varying channel width δ and investigate the efficiency of the mSK-ROCK and SK-ROCK method. Hence, for each $\delta = 1/2^k$ with $k = 0, \dots, 15$ we solve once

$$dX(t) = f_S(t, X(t)) dt + f_F(X(t)) dt + g(X(t)) dB(t) \quad t \in (0, T], \quad X(0) = 0$$

with the mSK-ROCK and SK-ROCK methods, on the same sample $B(t)$ with $T = 0.1$ and the same step size $\tau = 0.01$. The relative speed-up S given by the mSK-ROCK scheme over the SK-ROCK method, in terms of CPU time, in function of δ is displayed in Figure 4.13(a). For large δ both methods have the same performance ($S \approx 1$), as δ decreases the mSK-ROCK becomes more efficient than SK-ROCK and it is at least 25 times faster for some values of δ .

The relative speed-up has been computed dividing the computational costs (CPU time) of the SK-ROCK and mSK-ROCK method, that are plotted in Figure 4.13(b). This choice is justified by the fact that the relative error between the two solutions, measured in the $L^2(\Omega_\delta)$ norm at time T , is less than 1%, see Figure 4.13(c).

The spectral radii ρ, ρ_F, ρ_S of the Jacobians of f, f_F, f_S are shown in Figure 4.13(d), for large δ the typical element size is sufficiently small to resolve the channel (Figure 4.12(a)) and thus $\rho \approx \rho_F \approx \rho_S$, implying that the costs of mSK-ROCK and SK-ROCK are similar. As δ decreases then ρ, ρ_F increase. Since ρ_S is almost constant the number of f_S evaluations in the mSK-ROCK method remains constant and only the number of f_F evaluations increase, therefore the cost of mSK-ROCK increases slower than the one of SK-ROCK. Finally, in Figure 4.13(e) we show the number of stages taken by the methods, which reflects the behavior of the spectral radii.

In Figure 4.13(a) we see a decrease in speed-up for δ extremely small, this is due to the fact that the cost of evaluating f_F with respect to f_S and g becomes important; a high number of tiny elements is indeed needed to resolve the channel. However, the mSK-ROCK scheme stays about 20 times faster than the SK-ROCK method.

4.5 Proofs of technical results

We prove here all the technical results needed to show the stability and convergence of the mEM and mSK-ROCK methods.

4.5.1 Technical results for the mEM method

In this section we prove the results needed by the analysis of the mEM method. We start with a preliminary result needed to show Lemma 4.7 below.

Lemma 4.21. *Let $M \in \mathbb{N}^*$, then $(1+z)^2 \Phi_{2M}^{EE}(z) \leq \Phi_{2(M+1)}^{EE}(z) \leq \Phi_{2M}^{EE}(z) \leq 1+z/2$ for all $z \in [-2, 0]$.*

Proof. We start proving $\Phi_{2(M+1)}^{EE}(z) \leq \Phi_{2M}^{EE}(z)$. For $z = 0$ it obviously holds, as $\Phi_N^{EE}(0) = 1$ for all $N \in \mathbb{N}$. For $z < 0$ we observe that

$$\Phi_N^{EE}(z) = \frac{1}{N} \frac{(1+z)^N - 1}{z} = \frac{1}{z} \int_0^z (1+s)^{N-1} ds, \quad (4.56)$$

hence replacing (4.56) into

$$\Phi_{2(M+1)}^{EE}(z) \leq \Phi_{2M}^{EE}(z) \quad (4.57)$$

and multiplying by $z < 0$ we obtain the equivalent inequality

$$\int_0^z (1+s)^{2M+1} ds \geq \int_0^z (1+s)^{2M-1} ds,$$

which is rewritten as

$$0 \leq \int_0^z ((1+s)^{2M+1} - (1+s)^{2M-1}) ds = \int_z^0 (1+s)^{2M-1} (1 - (1+s)^2) ds = I(z).$$

We observe that the integrand is an odd function with respect to $s = -1$, being strictly positive for $s \in (-1, 0)$ and strictly negative for $s \in (-2, -1)$. Therefore, $I(z)$ satisfies $I(0) = 0$, is strictly positive for $z \in [-1, 0)$ and reaches its maximum in $z = -1$. It is also positive for $z \in (-2, -1]$ and satisfies $I(-2) = 0$. Relation (4.57) is therefore proved.

Since $\Phi_2^{EE}(z) = 1+z/2$ it follows $\Phi_{2(M+1)}^{EE}(z) \leq \Phi_{2M}^{EE}(z) \leq 1+z/2$ by recursion. It remains to show

$$(1+z)^2 \Phi_{2M}^{EE}(z) \leq \Phi_{2(M+1)}^{EE}(z).$$

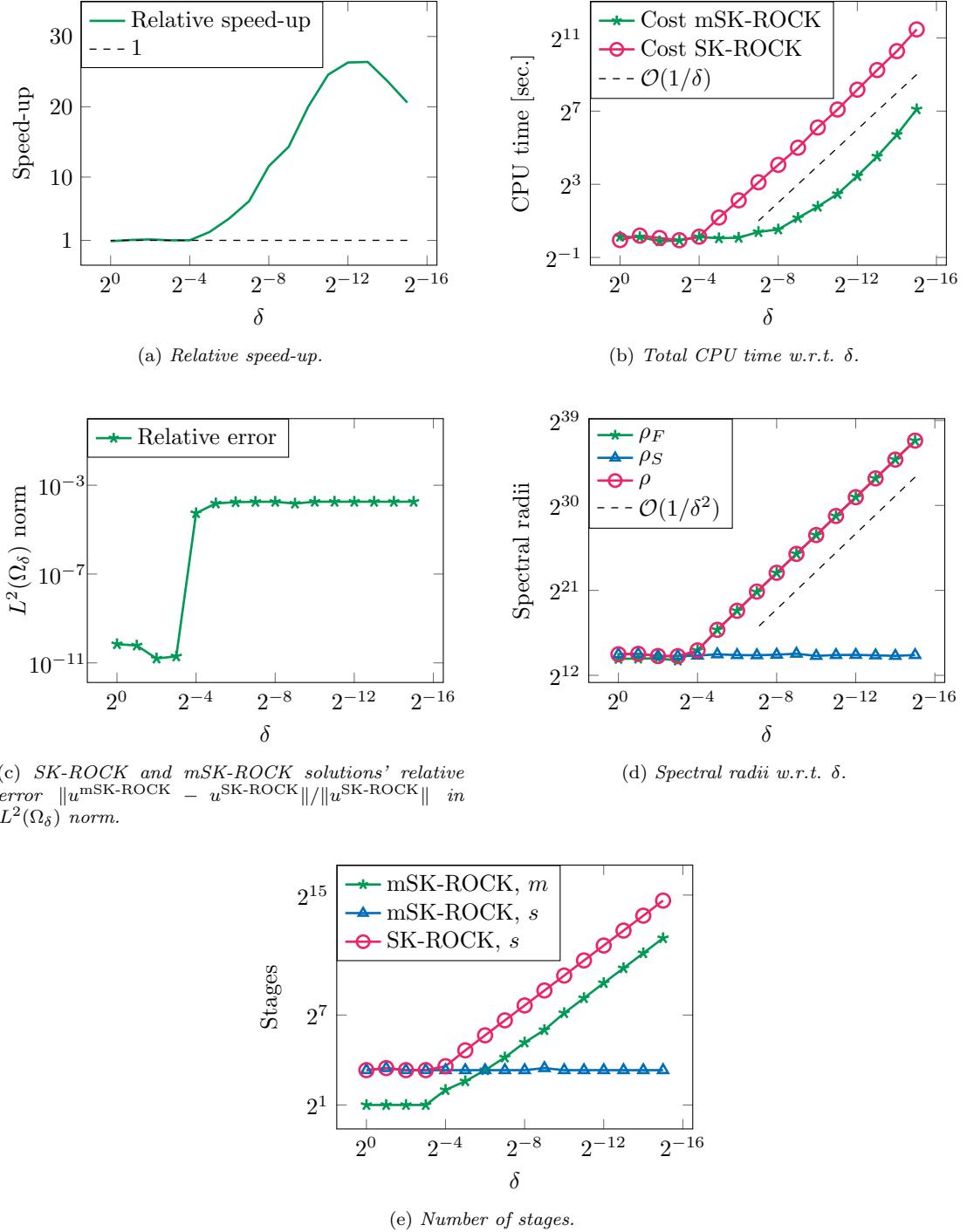


Figure 4.13. Narrow channel. Efficiency, evolution of spectral radii and stages.

We have, from (4.56) and the geometric sum identity,

$$\begin{aligned} (1+z)^2 \Phi_{2M}^{EE}(z) &= \frac{(1+z)^2}{2M} \sum_{k=0}^{2M-1} (1+z)^k = \frac{1}{2M} \sum_{k=2}^{2M+1} (1+z)^k \\ &= \frac{1}{2M} \left(\sum_{k=0}^{2M+1} (1+z)^k - 2 - z \right) = \frac{1}{M} \left((M+1)\Phi_{2(M+1)}^{EE}(z) - 1 - \frac{1}{2}z \right) \\ &\leq \Phi_{2(M+1)}^{EE}(z), \end{aligned}$$

as we already proved $\Phi_{2(M+1)}^{EE}(z) - 1 - z/2 \leq 0$. ■

Proof of Lemma 4.7. We will prove $\Psi_M^{EE}(z)^2 \leq \Phi_{2M}^{EE}(z)$ by recurrence over M . For $M = 1$ we verify

$$\Psi_1^{EE}(z)^2 = \left(1 + \frac{1}{2}z\right)^2 \leq 1 + \frac{1}{2}z = \Phi_2^{EE}(z),$$

as $z \in [-2, 0]$ and thus $1 + z/2 \in [0, 1]$. Supposing that $\Psi_M^{EE}(z)^2 \leq \Phi_{2M}^{EE}(z)$ holds, for $M + 1$ we have

$$\begin{aligned} \Psi_{M+1}^{EE}(z)^2 &= (1+z)^{2M} \left(1 + \frac{1}{2}z\right)^2 = (1+z)^2 (1+z)^{2(M-1)} \left(1 + \frac{1}{2}z\right)^2 \\ &= (1+z)^2 \Psi_M^{EE}(z)^2 \leq (1+z)^2 \Phi_{2M}^{EE}(z) \leq \Phi_{2(M+1)}^{EE}(z), \end{aligned}$$

where we used Lemma 4.21. ■

The error estimates for the mEM method in Theorem 4.11 follow from Lemma 4.10, we will first prove some technical results and the prove Lemma 4.10 at the end of this section.

Lemma 4.22. *The steps u_j of (3.21) satisfy*

$$|u_j - u_0| \leq C(1 + |u_0|)j\Delta\tau, \quad \text{for } j = 0, \dots, N, \quad (4.58)$$

and the averaged force \bar{f}_η (3.20) satisfies the linear growth condition

$$|\bar{f}_\eta(u_0)| \leq C(1 + |u_0|). \quad (4.59)$$

Proof. We start computing a bound on the steps u_j . It holds, from (3.21),

$$\begin{aligned} |u_{j+1}| &\leq |u_j| + \Delta\tau(|f_F(u_j)| + |f_S(u_0)|) \\ &\leq |u_j| + \Delta\tau C(1 + |u_j|) + \Delta\tau C(1 + |u_0|) \\ &\leq (1 + C\Delta\tau)|u_j| + C(1 + |u_0|)\Delta\tau \\ &\leq (1 + C\Delta\tau)^{j+1}|u_0| + C(1 + |u_0|)\Delta\tau \sum_{k=0}^j (1 + C\Delta\tau)^k. \end{aligned}$$

Using $\Delta\tau \sum_{k=0}^j (1 + C\Delta\tau)^k = C((1 + C\Delta\tau)^{j+1} - 1) \leq C(1 + C\Delta\tau)^{j+1}$, yields

$$\begin{aligned} |u_{j+1}| &\leq (1 + C\Delta\tau)^{j+1}|u_0| + C(1 + |u_0|)(1 + C\Delta\tau)^{j+1} \\ &\leq C(1 + C\Delta\tau)^{j+1}(1 + |u_0|) \leq Ce^{C(j+1)\Delta\tau}(1 + |u_0|) \\ &\leq Ce^{C\eta}(1 + |u_0|) = C(1 + |u_0|). \end{aligned}$$

as $(j+1)\Delta\tau \leq N\Delta\tau = \eta$. Now, let us estimate $|u_j - u_0|$. Starting from (3.21), we have

$$\begin{aligned} |u_{j+1} - u_0| &\leq |u_j - u_0| + \Delta\tau(|f_F(u_j)| + |f_S(u_0)|) \leq |u_j - u_0| + C\Delta\tau(1 + |u_j| + |u_0|) \\ &\leq |u_j - u_0| + C\Delta\tau(1 + |u_0|), \end{aligned}$$

from which (4.58) follows. Estimate (4.59) follows from (3.20), (4.58) with $j = N$ and $N\Delta\tau = \eta$. \blacksquare

Lemma 4.23. *The steps v_j, \bar{v}_j of (4.16) and (4.17) satisfy*

$$|v_j - \bar{v}_j| \leq C(1 + |v_0|)\eta, \quad \text{for } j = 1, \dots, M, \quad (4.60)$$

and the approximated diffusion \bar{g}_η (4.15) satisfies the linear growth condition

$$|\bar{g}_\eta(v_0)| \leq C(1 + |v_0|). \quad (4.61)$$

Proof. For the first step it holds

$$\begin{aligned} |v_1 - \bar{v}_1| &\leq \eta|g(v_0)| + \Delta\tau \left| f_F \left(v_0 + \frac{1}{2}\eta g(v_0) \right) - f_F(v_0) \right| \\ &\leq \eta|g(v_0)| + C\Delta\tau\eta|g(v_0)| \leq (1 + C\Delta\tau)(1 + |v_0|)\eta. \end{aligned}$$

For the next steps we have, for $j \geq 1$,

$$\begin{aligned} |v_{j+1} - \bar{v}_{j+1}| &\leq |v_j - \bar{v}_j| + \Delta\tau|f_F(v_j) - f_F(\bar{v}_j)| \leq (1 + C\Delta\tau)|v_j - \bar{v}_j| \\ &\leq (1 + C\Delta\tau)^j|v_1 - \bar{v}_1| \leq (1 + C\Delta\tau)^{j+1}(1 + |v_0|)\eta. \end{aligned}$$

We conclude using again $(1 + C\Delta\tau)^j \leq e^{Cj\Delta\tau} \leq e^{C\eta}$. For (4.61) we use (4.15) and (4.60). \blacksquare

Lemma 4.24. *For $u_0 \in \mathbb{R}^n$, the exact right-hand side f and the averaged force \bar{f}_η of (3.20) satisfy*

$$|\bar{f}_\eta(u_0) - f(u_0)| \leq C(1 + |u_0|)\eta.$$

For $v_0 \in \mathbb{R}^n$, the exact diffusion g and the approximated one \bar{g}_η of (4.15) satisfy

$$|\bar{g}_\eta(v_0) - g(v_0)| \leq C(1 + |v_0|)\eta.$$

Proof. We define \bar{u}_j by $\bar{u}_0 = u_0$ and $\bar{u}_{j+1} = \bar{u}_j + \Delta\tau f(u_0)$, hence

$$\bar{u}_N = \bar{u}_0 + N\Delta\tau f(u_0) = u_0 + \eta f(u_0)$$

and thus

$$\bar{f}_\eta(u_0) - f(u_0) = \frac{1}{\eta}(u_N - u_0) - \frac{1}{\eta}(\bar{u}_N - u_0) = \frac{1}{\eta}(u_N - \bar{u}_N).$$

Hence, if $|u_N - \bar{u}_N| \leq C(1 + |u_0|)\eta^2$ the proof is completed. From (3.21) and the definition of \bar{u}_j , we have

$$|u_{j+1} - \bar{u}_{j+1}| \leq |u_j - \bar{u}_j| + \Delta\tau|f_F(u_j) - f_F(u_0)|,$$

using (4.23) and (4.58) yields

$$|u_{j+1} - \bar{u}_{j+1}| \leq |u_j - \bar{u}_j| + C(1 + |u_0|)j\Delta\tau^2 \leq |u_j - \bar{u}_j| + C(1 + |u_0|)\Delta\tau\eta,$$

as $j\Delta\tau \leq \eta$. Therefore,

$$|u_{j+1} - \bar{u}_{j+1}| \leq C(1 + |u_0|)(j+1)\Delta\tau\eta \leq C(1 + |u_0|)\eta^2$$

for $j = 1, \dots, N - 1$. Thus $|u_N - \bar{u}_N| \leq C(1 + |u_0|)\eta^2$ holds.

For the second result, we show recursively that

$$|v_j - \bar{v}_j - \eta g(v_0)| \leq C(1 + |v_0|)j\Delta\tau\eta \quad \text{for } j = 1, \dots, M$$

and thus

$$|\bar{g}_\eta(v_0) - g(v_0)| = \frac{1}{\eta} |\bar{v}_M - v_M - \eta g(v_0)| \leq C(1 + |v_0|)M\Delta\tau \leq C(1 + |v_0|)\eta,$$

as $M\Delta\tau \leq \eta$. We have

$$\begin{aligned} v_{j+1} - \bar{v}_{j+1} &= v_j - \bar{v}_j + \Delta\tau(f_F(v_j) - f_F(\bar{v}_j)) \\ &= v_1 - \bar{v}_1 + \Delta\tau \sum_{k=1}^j f_F(v_k) - f_F(\bar{v}_k) \\ &= \eta g(v_0) + \Delta\tau \left(f_F \left(v_0 + \frac{1}{2}\eta g(v_0) \right) - f_F(v_0) \right) + \Delta\tau \sum_{k=1}^j f_F(v_k) - f_F(\bar{v}_k) \end{aligned}$$

and thus, using (4.60),

$$\begin{aligned} |v_{j+1} - \bar{v}_{j+1} - \eta g(v_0)| &\leq C\Delta\tau\eta|g(v_0)| + C\Delta\tau \sum_{k=1}^m |v_k - \bar{v}_k| \\ &\leq C(1 + |v_0|)\Delta\tau\eta + C(1 + |v_0|)\Delta\tau m\eta \\ &\leq C(1 + |v_0|)\Delta\tau(m + 1)\eta. \end{aligned}$$
■

Proof of Lemma 4.10. Comparing (4.14) and (4.25) we readily find

$$R = \tau(\bar{f}_\eta(X_n) - f(X_n)) + (\bar{g}_\eta(X_n) - g(X_n))\Delta W_n = R_1 + R_2. \quad (4.62)$$

Using Lemma 4.24 we estimate

$$\begin{aligned} |\mathbb{E}(R_1|X_n)| &\leq C(1 + |X_n|)\eta\tau, & \mathbb{E}(|R_1|^2|X_n)^{1/2} &\leq C(1 + |X_n|)\eta\tau, \\ |\mathbb{E}(R_2|X_n)| &= 0, & \mathbb{E}(|R_2|^2|X_n)^{1/2} &\leq C(1 + |X_n|)\eta\tau^{1/2}. \end{aligned}$$

Hence, $|\mathbb{E}(R|X_n)| \leq C(1 + |X_n|)\eta\tau$ and $\mathbb{E}(|R|^2|X_n)^{1/2} \leq C(1 + |X_n|)\eta\tau^{1/2}$. Estimate (4.26) is obtained applying a Taylor expansion to $\psi(X_{n+1})$ and using (4.62). ■

4.5.2 Technical results for the mSKROCK method

We prove here the lemmas needed to show stability and convergence of the mSK-ROCK scheme.

Proof of Lemma 4.14. Since $\Psi_r(0)^2 = \Phi_m(0) = 1$ we consider $z \in [-\ell_m^\varepsilon, 0]$. For $\varepsilon = 0$ we have $v_0 = 1$, $v_1 = 1/m^2$. Letting $x = v_0 + v_1 z = 1 + z/m^2 \in [-1, 1]$ and using $U_{r-1}(1) = r$, we have

$$\Psi_r(z) = \frac{U_{r-1}(x)}{2r}(x + 1). \quad (4.63)$$

Using the identity $2T_n(x)T_m(x) = T_{n+m}(x) + T_{|n-m|}(x)$ we obtain $2T_r(x)^2 = T_m(x) + 1$ and

$$\Phi_m(z) = \frac{T_m(x) - 1}{z} = \frac{T_m(x) - 1}{m^2(x - 1)} = \frac{T_r(x)^2 - 1}{2r^2(x - 1)}. \quad (4.64)$$

From (4.63) and (4.64) and $x - 1 < 0$, $\Psi_r(z)^2 \leq \Phi_m(z)$ is equivalent to

$$0 \leq U_{r-1}(x)^2(x^2 - 1)(x + 1) - 2(T_r(x)^2 - 1).$$

Using the identity $T_r(x)^2 - 1 = U_{r-1}(x)^2(x^2 - 1)$ and $x \in [-1, 1]$ the result follows. \blacksquare

Lemma 4.25. *The steps u_j of (3.37) satisfy*

$$|u_j - u_0| \leq C(1 + |u_0|)\eta \quad (4.65)$$

and the averaged force \bar{f}_η (3.36) satisfies the linear growth condition

$$|\bar{f}_\eta(u_0)| \leq C(1 + |u_0|). \quad (4.66)$$

Proof. We start computing a bound on $|u_j|$. We define $d_0 = |u_0|$ and

$$d_j = (1 + C\alpha_j\eta)d_{j-1} + C\alpha_j\eta(1 + |u_0|) \quad \text{for } j = 1, \dots, s. \quad (4.67)$$

Then, $d_j \leq d_{j+1}$ and $|u_j| \leq d_j$, indeed using (3.37) and $\beta_j + \gamma_j = 1$ yields

$$|u_1| \leq |u_0| + C\alpha_1\eta(1 + |u_0|) \leq (1 + C\alpha_1\eta)|u_0| + C\alpha_1\eta(1 + |u_0|) = d_1$$

and

$$\begin{aligned} |u_j| &\leq \beta_j d_{j-1} + \gamma_j d_{j-2} + C\alpha_j\eta(1 + d_{j-1}) + C\alpha_j\eta(1 + |u_0|) \\ &\leq d_{j-1} + C\alpha_j\eta(1 + d_{j-1}) + C\alpha_j\eta(1 + |u_0|) \leq d_j. \end{aligned}$$

From (4.67) we deduce

$$d_j = \prod_{k=1}^j (1 + C\alpha_k\eta)|u_0| + \sum_{k=1}^j \left(\prod_{i=k+1}^j (1 + C\alpha_i\eta) \right) C\alpha_k\eta(1 + |u_0|).$$

Using $1 + x \leq e^x$ and $\sum_{k=1}^s \alpha_k < C$ with C independent from s , it yields

$$\begin{aligned} d_j &\leq e^{C\eta}|u_0| + e^{C\eta} \sum_{k=1}^j C\alpha_k\eta(1 + |u_0|) \leq e^{C\eta}|u_0| + Ce^{C\eta}\eta(1 + |u_0|) \\ &\leq e^{C\eta}(1 + C\eta)|u_0| + Ce^{C\eta}\eta \leq C(1 + |u_0|). \end{aligned}$$

Let us now estimate $|u_j - u_0|$. From (3.37) we have

$$\begin{aligned} u_1 - u_0 &= \alpha_1\eta(f_F(u_0) + f_S(u_0)) \\ u_j - u_0 &= \beta_j(u_{j-1} - u_0) + \gamma_j(u_{j-2} - u_0) + \alpha_j\eta(f_F(u_{j-1}) + f_S(u_0)) \quad j = 2, \dots, s. \end{aligned}$$

Using (1.25) with $z = 0$ and $r_j = \alpha_j\eta(f_F(u_j) + f_S(u_0))$ for $j = 1, \dots, s-1$ we obtain

$$u_j - u_0 = \sum_{k=1}^j \frac{a_j}{a_k} U_{j-k}(v_0) \alpha_k \eta(f_F(u_{k-1}) + f_S(u_0))$$

Following the lines of the proof of Lemma 3.17, we obtain

$$\sum_{k=1}^j \frac{a_j}{a_k} U_{j-k}(v_0 + v_1 z) \alpha_k = \frac{a_j T_j(v_0 + v_1 z) - 1}{z}$$

and thus, taking the limit $z \rightarrow 0$ and using $a_j T_j(v_0) = 1$,

$$\sum_{k=1}^j \frac{a_j}{a_k} U_{j-k}(v_0) \alpha_k = a_j v_1 T'_j(v_0) = \frac{T_m(v_0) T'_j(v_0)}{T'_m(v_0) T_j(v_0)} \leq 1,$$

where the last inequality follows from [125, Eqs. (2.7),(2.18)]. Since $\frac{a_j}{a_k} U_{j-k}(v_0) \alpha_k \geq 0$, it yields

$$\begin{aligned} |u_j - u_0| &\leq \eta \max_{k=1,\dots,j} |f_F(u_{k-1}) + f_S(u_0)| \leq C\eta \max_{k=1,\dots,j} (1 + |u_{k-1}| + |u_0|) \\ &\leq C(1 + |u_0|)\eta. \end{aligned}$$

Estimate (4.66) follows from (3.36) and (4.65). \blacksquare

Lemma 4.26. *The stages v_j, \bar{v}_j of (4.31) and (4.32) satisfy*

$$|v_j - \bar{v}_j| \leq C(1 + |v_0|)\eta \quad (4.68)$$

and the approximated diffusion \bar{g}_η (4.30) satisfies the linear growth condition

$$|\bar{g}_\eta(v_0)| \leq C(1 + |v_0|). \quad (4.69)$$

Proof. For $j = 0$ holds $|v_0 - \bar{v}_0| = 0$ and for $j = 1$ we compute, using (4.31) and (4.32),

$$\begin{aligned} |v_1 - \bar{v}_1| &\leq \gamma_1 \theta_1 \eta |g(v_0)| + \alpha_1 \eta |f_F(v_0 + \beta_1 \theta_1 \eta g(v_0)) - f_F(v_0)| \\ &\leq C(1 + |v_0|)\eta + C\alpha_1 \beta_1 \theta_1 \eta^2 |g(v_0)| \leq C(1 + |v_0|)\eta. \end{aligned}$$

For $j \geq 2$, we have

$$\begin{aligned} |v_j - \bar{v}_j| &\leq \beta_j |v_{j-1} - \bar{v}_{j-1}| + \gamma_j |v_{j-2} - \bar{v}_{j-2}| + \alpha_j \eta |f_F(v_{j-1}) - f_F(\bar{v}_{j-1})| \\ &\leq C(1 + |v_0|)\eta + C\eta |v_{j-1} - \bar{v}_{j-1}| \leq C(1 + |v_0|)\eta. \end{aligned}$$

Estimate (4.69) follows from (4.30) and (4.68). \blacksquare

Proof of Lemma 4.16. We first prove (4.42) for \bar{f}_η, f . Let u_j, \bar{u}_j be the stages (3.36) where u_0 is replaced by x, y , respectively. From (3.36) we have

$$\begin{aligned} |(u_0 - x) - (\bar{u}_0 - y)| &= 0, \\ |(u_1 - x) - (\bar{u}_1 - y)| &\leq \alpha_1 \eta |f_F(x) + f_S(x) - f_F(y) - f_S(y)| \leq C|x - y|\eta. \end{aligned}$$

Let $j \geq 2$ and suppose $|(u_k - x) - (\bar{u}_k - y)| \leq C|x - y|\eta$ for $k < j$, (3.36) yields

$$\begin{aligned} (u_j - x) - (\bar{u}_j - y) &= \beta_j ((u_{j-1} - x) - (\bar{u}_{j-1} - y)) + \gamma_j ((u_{j-2} - x) - (\bar{u}_{j-2} - y)) \\ &\quad + \alpha_j \eta (f_F(u_{j-1}) + f_S(x) - f_F(\bar{u}_{j-1}) - f_S(y)) \end{aligned}$$

and therefore

$$|(u_j - x) - (\bar{u}_j - y)| \leq C|x - y|\eta + C\alpha_j \eta (|u_{j-1} - \bar{u}_{j-1}| + |x - y|) \leq C|x - y|\eta$$

by triangular inequality. Estimate (4.42) for \bar{f}_η follows from

$$\bar{f}_\eta(x) - \bar{f}_\eta(y) = \frac{1}{\eta} ((u_s - x) - (\bar{u}_s - y)).$$

Now we prove (4.42) for \bar{g}_η, g . Let v_j, \bar{v}_j bet the stages (4.31) and (4.32) where v_0 is replaced by x and z_j, \bar{z}_j be the stages (4.31) and (4.32) where v_0 is replaced by y . We have

$$\begin{aligned} v_0 - z_0 &= x - y, \\ v_1 - z_1 &= x - y + \alpha_1 \eta (f_F(x + \beta_1 \theta_1 \eta g(x)) - f_F(y + \beta_1 \theta_1 \eta g(y))) + \gamma_1 \theta_1 \eta (g(x) - g(y)) \end{aligned}$$

and therefore

$$\begin{aligned} |(v_0 - z_0) - (x - y)| &= 0, \\ |(v_1 - z_1) - (x - y)| &\leq C\eta|(x + \beta_1 \theta_1 \eta g(x)) - (y + \beta_1 \theta_1 \eta g(y))| + C|x - y|\eta \leq C|x - y|\eta. \end{aligned}$$

Let $j \geq 2$ and suppose $|(v_k - z_k) - (x - y)| \leq C|x - y|\eta$ for $k < j$, it holds

$$\begin{aligned} (v_j - z_j) - (x - y) &= \beta_j((v_{j-1} - z_{j-1}) - (x - y)) + \gamma_j((v_{j-2} - z_{j-2}) - (x - y)) \\ &\quad + \alpha_j \eta (f_F(v_{j-1}) - f_F(z_{j-1})) \end{aligned}$$

and

$$|(v_j - z_j) - (x - y)| \leq C|x - y|\eta + C\eta|v_{j-1} - z_{j-1}| \leq C|x - y|\eta,$$

by triangular inequality. In a similar fashion we obtain $|(\bar{v}_j - \bar{z}_j) - (x - y)| \leq C|x - y|\eta$ for $j = 0, \dots, r$. Hence,

$$\begin{aligned} |\bar{g}_\eta(x) - \bar{g}_\eta(y)| &= \frac{1}{\eta}|(v_r - \bar{v}_r) - (z_r - \bar{z}_r)| \\ &= \frac{1}{\eta}|(v_r - z_r - (x - y)) - (\bar{v}_r - \bar{z}_r - (x - y))| \leq C|x - y|. \end{aligned}$$

Let us prove $|\bar{f}_\eta(u_0) - f(u_0)| \leq C(1 + |u_0|)\eta$, i.e. (4.43) for \bar{f}_η, f . We define F_j for $0 = 1, \dots, m$ by

$$\begin{aligned} F_0 &= 0, \\ F_1 &= \alpha_1 \eta f(u_0), \\ F_j &= \beta_j F_{j-1} + \gamma_j F_{j-2} + \alpha_j \eta f(u_0) \quad j = 2, \dots, m, \end{aligned}$$

using (1.24) and (1.25) with $z = 0$ and $r_j = \alpha_j \eta f(u_0)$ yields

$$F_m = \sum_{k=1}^m \frac{a_m}{a_k} U_{m-k}(v_0) \alpha_k \eta f(u_0) = \Phi_m(0) \eta f(u_0) = \eta f(u_0)$$

and thus

$$\bar{f}_\eta(u_0) - f(u_0) = \frac{1}{\eta}(u_m - u_0 - F_m).$$

We will show that $|u_m - u_0 - F_m| \leq C(1 + |u_0|)\eta^2$ and (4.43) for \bar{f}_η, f follows. We have

$$\begin{aligned} u_0 - u_0 - F_0 &= 0, \\ u_1 - u_0 - F_1 &= 0, \\ u_j - u_0 - F_j &= \beta_j(u_{j-1} - u_0 - F_{j-1}) + \gamma_j(u_{j-2} - u_0 - F_{j-2}) + \alpha_j \eta (f_F(u_{j-1}) - f_F(u_0)), \end{aligned}$$

hence, using (1.24), (1.25) and (4.65), we obtain

$$\begin{aligned} |u_m - u_0 - F_m| &\leq \left| \sum_{k=1}^m \frac{a_m}{a_k} U(v_0) \alpha_k \eta (f_F(u_{k-1}) - f_F(u_0)) \right| \leq C \sum_{k=1}^m \frac{a_m}{a_k} U(v_0) \alpha_k \eta |u_{k-1} - u_0| \\ &\leq C\eta \max_{k=1, \dots, m} |u_{k-1} - u_0| \leq C(1 + |u_0|)\eta^2. \end{aligned}$$

Now we show $|\bar{g}_\eta(v_0) - g(v_0)| \leq C(1 + |v_0|)\eta$ with $v_0 \in \mathbb{R}^n$, hence (4.43) for \bar{g}_η, g . We define

$$\begin{aligned} G_0 &= 0, \\ G_1 &= \gamma_1 \theta_1 \eta g(v_0), \\ G_j &= \beta_j G_{j-1} + \gamma_j G_{j-2} \quad j = 2, \dots, r. \end{aligned} \tag{4.70}$$

From (1.25) with $z = 0$, $r_1 = \gamma_1 \theta_1 \eta g(v_0)$ and the definition of γ_1, θ_1 in (4.33) follows

$$G_r = \frac{a_r}{a_1} U_{r-1}(v_0) \gamma_1 \theta_1 \eta g(v_0) = \eta g(v_0)$$

and therefore, from (4.30),

$$\bar{g}_\eta(v_0) - g(v_0) = \frac{1}{\eta} (v_r - \bar{v}_r - G_r),$$

with v_j, \bar{v}_j as in (4.31) and (4.32). Let us prove $|v_r - \bar{v}_r - G_r| \leq C(1 + |v_0|)\eta^2$, which implies $|\bar{g}_\eta(v_0) - g(v_0)| \leq C(1 + |v_0|)\eta$. Subtracting (4.32) and (4.70) from (4.31) yields

$$\begin{aligned} v_0 - \bar{v}_0 - G_0 &= 0, \\ v_1 - \bar{v}_1 - G_1 &= \alpha_1 \eta (f_F(v_0 + \beta_1 \theta_1 \eta g(v_0)) - f_F(\bar{v}_0)), \\ v_j - \bar{v}_j - G_j &= \beta_j (v_{j-1} - \bar{v}_{j-1} - G_{j-1}) + \gamma_j (v_{j-2} - \bar{v}_{j-2} - G_{j-2}) \quad j = 2, \dots, r. \\ &\quad + \alpha_j \eta (f_F(v_{j-1}) - f_F(\bar{v}_{j-1})) \end{aligned}$$

Using

$$\begin{aligned} |f_F(v_0 + \beta_1 \theta_1 \eta g(v_0)) - f_F(\bar{v}_0)| &\leq C(1 + |v_0|)\eta, \\ |f_F(v_{j-1}) - f_F(\bar{v}_{j-1})| &\leq C|v_{j-1} - \bar{v}_{j-1}| \leq C(1 + |v_0|)\eta \end{aligned}$$

and then (1.25) yields $|v_r - \bar{v}_r - G_r| \leq C(1 + |v_0|)\eta^2$ and thus (4.43) for g, \bar{g}_η . \blacksquare

Proof of Lemma 4.17. We prove recursively, using (3.37), (4.31) and (4.32), that u_j, v_j, \bar{v}_j seen as functions of u_0, v_0 are in $C_p^4(\mathbb{R})$ and hence $\bar{f}_\eta, \bar{g}_\eta \in C_p^4(\mathbb{R})$. \blacksquare

Proof of Lemma 4.18. We start proving (4.44). From (4.69) and $\bar{Q}_\eta = \bar{g}_\eta(X_n)\Delta W_n$ we readily get $|\bar{Q}_\eta| \leq C(1 + |X_n|)|\Delta W_n|$. We also have $|K_0 - X_n| = 0$ and

$$\begin{aligned} |K_1 - X_n| &\leq \mu_1 \tau |\bar{f}_\eta(X_n + \nu_1 \bar{Q}_\eta)| + \kappa_1 |\bar{Q}_\eta| \leq C\tau(1 + |X_n| + |\bar{Q}_\eta|) + |\bar{Q}_\eta| \\ &\leq C(1 + |X_n|)(\tau + |\Delta W_n|). \end{aligned}$$

Let us suppose $|K_k - X_n| \leq C(1 + |X_n|)(\tau + |\Delta W_n|)$ for $k < j$, then

$$|K_k| \leq C(1 + |X_n|)(1 + \tau + |\Delta W_n|)$$

and thus, using (4.66),

$$\begin{aligned} |K_j - X_n| &\leq \nu_j |K_{j-1} - X_n| + \kappa_j |K_{j-2} - X_n| + \mu_j \tau |\bar{f}_\eta(K_{j-1})| \\ &\leq C(1 + |X_n|)(\tau + |\Delta W_n|) + C\tau(1 + |K_{j-1}|) \\ &\leq C(1 + |X_n|)(\tau + |\Delta W_n|) + C\tau(1 + (1 + |X_n|)(1 + \tau + |\Delta W_n|)) \\ &\leq C(1 + |X_n|)(\tau + |\Delta W_n|). \end{aligned}$$

Let us now prove (4.45). Since $\bar{Q}_\eta = \bar{g}_\eta(X_n)\Delta W_n$ then $|\mathbb{E}(\bar{Q}_\eta|X_n)| = 0$. Using (4.44) and (4.66) yields

$$\begin{aligned} |\mathbb{E}(K_1 - X_0|X_n)| &= \mu_1\tau|\mathbb{E}(\bar{f}_\eta(X_n + \nu_1\bar{Q}_\eta)|X_n)| \leq C\tau(1 + |X_n| + \mathbb{E}(|\bar{Q}_\eta||X_n)) \\ &\leq C\tau(1 + |X_n| + C(1 + |X_n|)\tau^{1/2}) \leq C(1 + |X_n|)\tau. \end{aligned}$$

Let $j \geq 2$ and suppose $|\mathbb{E}(K_k - X_n|X_n)| \leq C(1 + |X_n|)\tau$ for $k < j$. Then, using (4.66),

$$\begin{aligned} |\mathbb{E}(K_j - X_0|X_n)| &\leq \nu_j|\mathbb{E}(K_{j-1} - X_0|X_n)| + \kappa_j|\mathbb{E}(K_{j-2} - X_0|X_n)| + \mu_j\tau\mathbb{E}(|\bar{f}_\eta(K_{j-1})||X_n)) \\ &\leq C(1 + |X_n|)\tau + C\tau(1 + \mathbb{E}(|K_{j-1}||X_n)) \leq C(1 + |X_n|)\tau, \end{aligned}$$

by triangular inequality $|K_{j-1}| \leq |X_n| + |K_{j-1} - X_n|$. ■

5 Conclusion of Part I

In this part of the thesis, we introduced multirate methods for stiff deterministic and stochastic differential equations. We first proposed an interpolation based scheme which, however, is prone to instabilities. Therefore, we derived a new framework for deterministic and stochastic differential equations which allows for the development of a class of interpolation-free and stable explicit multirate methods.

In Chapter 1 we recalled stabilized explicit methods, which are the core schemes used in this part of the thesis, and introduced a new stabilized explicit method for stochastic differential equations driven by discrete Poisson noise. First, we studied the optimal stability polynomial for first-order Runge–Kutta schemes and then introduced first-order stabilized schemes, as the RKC method. Next we briefly discussed higher order methods and then we recalled the SK-ROCK scheme for stochastic differential equations. Finally, we introduced the new SK- τ -ROCK scheme, which is an SK-ROCK-like method for SDEs driven by discrete noise. We showed accuracy and stability of the scheme and its post-processor in Theorems 1.3 and 1.4, in particular the SK- τ -ROCK scheme enjoys the same stability conditions as the RKC and SK-ROCK methods and furthermore its post-processed version captures the exact statistics of the test equation. We presented as well a numerical example on a nonlinear chemical system which confirms the accuracy and efficiency of the method.

In Chapter 2 we proposed an interpolation based additive RKC method for multirate ordinary differential equations. The method is based on a decomposition of the original problem in two subproblems and integrates both problems with an RKC scheme, where the number of stages is adapted to the stiffness of each subproblem. The different stages number leads to an asynchronous integration procedure and linear interpolation in time between stages is employed whenever coupling values are needed. The scheme is explicit and straightforward to implement. However, we have shown on a model problem that linear interpolations might render the scheme unstable. Furthermore, the second-order additive RKC method suffers from an order reduction phenomenon. Numerical examples corroborate the theoretical findings.

Chapter 3 is where we introduced the new framework based on modified equations for stiff ordinary differential equations with disparate time scales and also presented interpolation-free explicit multirate methods, as the multirate RKC (mRKC) method. Our methodology is based on an averaged right-hand side whose stiffness is decreased thanks to an auxiliary stiff problem involving only a small number of terms. In Theorem 3.4, we show that this averaged right-hand side preserves the contractivity properties of the original equation and in Theorem 3.5 that the error committed when solving the modified equation, instead of the original problem, is of first-order and independent of the problem’s stiffness.

Departing from the modified equation a whole class of explicit multirate methods can be derived, by discretizing the modified equation and the auxiliary problem by a time integration scheme, or even two different schemes. For the resulting methods, the number of function evaluations is independent of any severe stiffness induced by a few degrees of freedom and depends solely on the slow dynamics. In this thesis we discussed three such multirate methods, the most flexible of them being the mRKC method defined as the discretization of the modified equation and the auxiliary problem by an RKC scheme; the algorithm is given by (3.34) to (3.37). The mRKC method is first-order accurate, as proved in Theorem 3.21, and its stability is analyzed for a model problem in Theorem 3.20.

In contrast to the multiscale schemes in [42, 57], the mRKC method does not assume any scale separation, in the sense that the eigenvalues of the Jacobian of the fast component f_F need not cluster at the left end of the stability domain. Hence, for parabolic problems the mRKC method permits to overcome the crippling effect of local mesh refinement and thus recover the well-known efficiency of stabilized Runge–Kutta methods without sacrificing explicitness. In fact, as shown in Section 3.5.4, the stability conditions can even be weakened in that particular context, as those in (3.44) are sufficient, thus leading to even higher efficiency.

The mRKC method retains the explicitness of classical stabilized Runge–Kutta schemes: thus it is inherently parallel and does not require any special data structure. Numerical experiments show that the new scheme overcomes the bottleneck caused by the stiffness of a few degrees of freedom without sacrificing the accuracy or explicitness of stabilized methods.

In Chapter 4 we generalized the modified equation framework introduced in Chapter 3, allowing for the development of explicit multirate methods for stiff stochastic differential equations with different time-scales but without any clear-cut scale separation. To do so, we defined a damped diffusion term whose stiffness is decreased thanks to two deterministic auxiliary problems. The new diffusion term is such that the mean-square stability properties of the original problem are inherited by the modified equation.

Departing from the stochastic modified equation we derived a multirate method based on stochastic stabilized explicit Runge–Kutta methods, namely the SK-ROCK scheme, where evaluation of the modified drift and damped diffusion term demands the solution to stiff but cheap deterministic problems in a short time interval. The efficiency of the scheme is hardly affected by the severe stiffness introduced by a few degrees of freedom and the number of expensive function evaluations depends on the slow scales only. The scheme is given in (3.36), (3.37) and (4.29) to (4.32), its stability is proven on a model problem in Theorem 4.15 and in Theorem 4.20 we show that the method has strong order 1/2 and weak order 1.

The method is straightforward to implement and retains the explicitness of stabilized Runge–Kutta schemes. Numerical experiments verify the stability and accuracy properties predicted by the theory and demonstrate that the computational cost is significantly reduced without sacrificing any accuracy; the numerical solutions of the multirate and classical scheme are indeed essentially the same.

An important topic for further research is to develop a multirate stabilized scheme for stochastic differential equations driven by discrete noise. A building block is the SK- τ -ROCK scheme developed during this thesis, see Section 1.4.2. Next, one would need to investigate a modified equation for SDEs driven by discrete noise. As the modified equation would have different damping properties with respect to the original SDE a post-processing procedure might be necessary in order to capture the exact invariant measure.

Local adaptive discontinuous Galerkin schemes for elliptic equations

Part II

In this part of the thesis, we develop a local discontinuous Galerkin scheme for elliptic partial differential equations with high contrasts. As space discretization method we consider the Symmetric Weighted Interior Penalty (SWIP) scheme [39, 49], it is a generalization of the Symmetric Interior Penalty (SIP) method [23] which employs diffusivity-dependent averages and is therefore better suited for problems with discontinuous diffusivity, for instance. The local method relies on a coarse solution computed by the SWIP scheme and improves the accuracy by solving a sequence of local elliptic problems in refined subdomains, where artificial Dirichlet boundary conditions are imposed. No iterations between the coarse and refined subdomains are performed. A priori and a posteriori error analysis of the scheme are presented.

Chapters 6 and 7 are dedicated to the introduction of the local scheme and the a priori error analysis, for that we consider the elliptic model problem

$$\begin{aligned} -\nabla \cdot (A(u)\nabla u) &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega, \end{aligned}$$

where $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$ is an open bounded polytopal connected set, $u \in H_0^1(\Omega)$ and $f \in H^{-1}(\Omega)$. The matrix A is symmetric, positive definite and can possibly depend on u , since we consider both a linear and a quasilinear case.

Although our local scheme is based on the SWIP scheme, for the a priori error analysis we use the gradient discretization method (GDM) [41]. The GDM is a framework for proving the convergence of gradient schemes (GS); it allows for minimal assumptions and analysis of, for instance, quasilinear problems, fully nonlinear problems and degenerate problems. Hence, any space discretization method that can be written and a GS is directly known to converge for such problems. Many well known schemes can be written as a GS: the finite element method, the discontinuous finite element method, the hybrid finite volume method, the mimetic finite difference method and others [41, Part III]. In order to prove convergence of our local scheme we will use the GDM; therefore, we must recast the SWIP method into a GS. In Chapter 6 we first introduce the notation for the GDM and then define a new GS which is equivalent to the SWIP method. To do so, we follow [52] (see also [41, Chapter 11]), where the SIP method is written as a GS.

In Chapter 7 we define the local method and present the a priori error analysis under minimal regularity assumptions, i.e. $u \in H_0^1(\Omega)$ and $f \in H^{-1}(\Omega)$, using the GDM. The method is shown to converge for linear and quasilinear equations. The chapter is concluded by a sequence of numerical examples corroborating the theory and assessing the computational efficiency of the scheme.

In Chapter 7 we supposed that the high error regions are known and therefore the refined subdomains are defined upon an a priori knowledge of the solution's behavior. However, this is not always possible in practical applications and we thus need a posteriori error estimators in order to identify the subdomains to be refined. The a posteriori error analysis of the local SWIP scheme is presented in Chapter 8 for the linear advection-diffusion-reaction equation

$$\begin{aligned} -\nabla \cdot (A\nabla u) + \beta \cdot \nabla u + \mu u &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega, \end{aligned}$$

where β is the velocity field and μ the reaction coefficient. Here, $f \in L^2(\Omega)$ and A does not depend on u . Furthermore, in this chapter we do not use the GDM but the standard setting for discontinuous Galerkin methods [39]. The a posteriori error analysis that we present is based on the results of [48]; there, the authors introduce robust a posteriori error estimators for the SWIP scheme based on cutoff functions and conforming flux and potential reconstructions. Following the same strategy, we derive estimators for our local scheme; however, we must relax the regularity

requirements for the reconstructed fluxes and allow for jumps at the subdomains boundaries. Nevertheless, the new estimators inherit two main properties of the estimators introduced in [48]: they are robust in singularly perturbed regimes and free of undetermined constants. Furthermore, they are employed to identify the local subdomains and provide robust error bounds on the solution given by the local SWIP method.

Chapters 6 and 7 are based on [12] and Chapter 8 is essentially taken from [13].

6 The weighted discontinuous Galerkin gradient discretization scheme

In this chapter, we present the gradient discretization method (GDM) and the gradient scheme (GS) for the elliptic problem

$$\begin{aligned} -\nabla \cdot (A(u)\nabla u) &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega. \end{aligned} \quad (6.1)$$

Then we introduce the symmetric Weighted Discontinuous Galerkin Gradient Discretization (WDGGD).

We consider a linear and a quasilinear case; for the linear case we make the following assumption.

Assumption 6.1.

- $\Omega \subset \mathbb{R}^d$ is an open bounded polytopal domain,
- $A : \Omega \rightarrow \mathbb{R}^{d \times d}$ is such that $A(\boldsymbol{x})$ is a symmetric matrix measurable with respect to \boldsymbol{x} and there exists $\underline{\lambda}, \bar{\lambda} > 0$ such that it has eigenvalues in $[\underline{\lambda}, \bar{\lambda}]$,
- the forcing term is $f \in H^{-1}(\Omega)$.

For the sake of simplicity, we omit in the following the dependence on \boldsymbol{x} of the tensor A . Under Assumption 6.1 the unique weak solution of (6.1) is $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} A \nabla u \cdot \nabla v \, d\boldsymbol{x} = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega), \quad (6.2)$$

where $\langle \cdot, \cdot \rangle$ denotes the pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

We will consider also a quasilinear case, where we make the following

Assumption 6.2.

- $\Omega \subset \mathbb{R}^d$ is an open bounded polytopal domain,
- $A(\boldsymbol{x}, s) = (a_{ij}(\boldsymbol{x}, s))_{i,j=1}^d$ is such that $a_{ij} : \bar{\Omega} \times \mathbb{R} \rightarrow \mathbb{R}$ is continuous in \boldsymbol{x} and Lipschitz continuous in s . Furthermore $A(\boldsymbol{x}, s)$ is a symmetric matrix with eigenvalues in $[\underline{\lambda}, \bar{\lambda}]$,
- the forcing term is $f \in H^{-1}(\Omega)$.

Under Assumption 6.2 we have existence and uniqueness of the solution of the quasilinear problem

$$\int_{\Omega} A(u) \nabla u \cdot \nabla v \, d\boldsymbol{x} = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega). \quad (6.3)$$

The continuity of A with respect to \boldsymbol{x} is needed to ensure uniqueness of the solution and simplify the presentation but it is not needed by the local scheme.

This chapter starts in Section 6.1 with the presentation of the GDM for (6.1), to do so we mainly follow [41]. Then, inspired from [52], where the Symmetric Interior Penalty scheme is recast into a GS, in Section 6.2 we define the WDGGD scheme; which is the Symmetric Weighted Interior Penalty method but written as a GS. In the same section, we prove that the WDGGD satisfies the core properties of a gradient discretization and therefore the associated GS converges to the exact solution of (6.1), under the assumptions that $u \in H_0^1(\Omega)$ and $f \in H^{-1}(\Omega)$. Equivalence of the WDGGD scheme and the SWIG scheme is proved at the end of the chapter.

6.1 The gradient discretization method for homogeneous Dirichlet boundary conditions

The gradient discretization method is a convenient framework for studying the convergence properties of discretization schemes for diffusion problems of many types: linear and nonlinear, stationary or time dependent. The key idea is to define a discrete scheme, called gradient scheme, by replacing in the weak formulation of the problem, as (6.2) and (6.3), the continuous operators by discrete ones defined by a gradient discretization. A gradient discretization \mathcal{D} is defined by three objects: a real vector space $X_{\mathcal{D}}$ representing the degrees of freedom and two operators $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ which, from elements of $X_{\mathcal{D}}$, construct functions in $L^2(\Omega)$ and gradients in $L^2(\Omega)^d$, respectively. Hence, given a gradient discretization \mathcal{D} and a weak formulation, a gradient scheme is defined by replacing the continuous entities $u, \nabla u, v, \nabla v$ in (6.2) and (6.3) by the discrete ones $\Pi_{\mathcal{D}}\vartheta, \nabla_{\mathcal{D}}\vartheta, \Pi_{\mathcal{D}}\phi, \nabla_{\mathcal{D}}\phi$, with $\vartheta, \phi \in X_{\mathcal{D}}$. In the next section it is also explained how to deal with the pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Many classic discretization schemes can be written as a gradient scheme, upon the definition of an appropriate gradient discretization. For instance, the conforming \mathbb{P}_1 finite element Galerkin method: given a partition of Ω into simplices, let $V_h \subset H_0^1(\Omega)$ be the set of piecewise linear and continuous functions on this partition, $\{e_i\}_{i \in I}$ be a basis of V_h and define $X_{\mathcal{D}} = \{\phi = (\zeta_i)_{i \in I} : \zeta_i \in \mathbb{R} \text{ for all } i \in I\}$, $\Pi_{\mathcal{D}}\phi = \sum_{i \in I} \zeta_i e_i$ and $\nabla_{\mathcal{D}}\phi = \sum_{i \in I} \zeta_i \nabla e_i$.

If a given discretization scheme satisfies the core properties of a gradient discretization, thus enters in the gradient discretization method framework, then it converges for all problems for which gradient schemes are known to converge; for instance, a gradient scheme converges for the p -Laplace problem and therefore the conforming finite element method too. Furthermore, the gradient discretization method naturally allows for minimal regularity assumptions.

Following [41, Section 2.1], in this section we define the gradient discretization and prove convergence of the associated gradient scheme for (6.2) using the gradient discretization method. Convergence of the gradient scheme for (6.3) is stated, for the proof we refer to [41, Theorem 2.35].

6.1.1 Definition and approximation properties

We start by defining the gradient discretization (GD), then we recall its key properties and finally we define the gradient scheme (GS) used to solve (6.1).

Definition 6.3. A GD \mathcal{D} for homogeneous Dirichlet boundary conditions is defined by $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$, where

- 1) the set $X_{\mathcal{D}}$ is a finite dimensional real vector space,

- 2) the reconstruction function $\Pi_{\mathcal{D}} : X_{\mathcal{D}} \rightarrow L^2(\Omega)$ is a linear mapping that reconstructs, from an element in $X_{\mathcal{D}}$, a function over Ω ,
- 3) the gradient reconstruction $\nabla_{\mathcal{D}} : X_{\mathcal{D}} \rightarrow L^2(\Omega)^d$ is a linear mapping which reconstructs, from an element of $X_{\mathcal{D}}$, a gradient over Ω ,
- 4) the gradient reconstruction is such that $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ is a norm on $X_{\mathcal{D}}$.

Now, let us define some properties that a GD must satisfy in order to prove the convergence of the associated GS. In the following $(\mathcal{D}_n)_{n \in \mathbb{N}}$ is a sequence of GD, it is useful to think that each \mathcal{D}_n is associated to a mesh of size h_n with $\lim_{n \rightarrow \infty} h_n = 0$.

Definition 6.4. If \mathcal{D} is a GD, define $C_{\mathcal{D}}$ as the norm of $\Pi_{\mathcal{D}}$:

$$C_{\mathcal{D}} := \max_{\phi \in X_{\mathcal{D}} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}}\phi\|_{L^2(\Omega)}}{\|\nabla_{\mathcal{D}}\phi\|_{L^2(\Omega)^d}}.$$

A sequence $(\mathcal{D}_n)_{n \in \mathbb{N}}$ of GD is coercive if there exists $C_p \in \mathbb{R}_+$ such that $C_{\mathcal{D}_n} \leq C_p$ for all $n \in \mathbb{N}$.

We observe that coercivity implies a kind of Poincaré inequality.

Definition 6.5. If \mathcal{D} is a GD, define $S_{\mathcal{D}} : H_0^1(\Omega) \rightarrow [0, \infty[$ by

$$S_{\mathcal{D}}(v) := \min_{\phi \in X_{\mathcal{D}}} (\|\Pi_{\mathcal{D}}\phi - v\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\phi - \nabla v\|_{L^2(\Omega)^d}).$$

A sequence $(\mathcal{D}_n)_{n \in \mathbb{N}}$ of GD is space-consistent if $\lim_{n \rightarrow \infty} S_{\mathcal{D}_n}(v) = 0$ for all $v \in H_0^1(\Omega)$.

Note that Definition 6.5 makes sense since the norms $\|\cdot\|_{L^2(\Omega)}, \|\cdot\|_{L^2(\Omega)^d}$ are strictly convex; therefore, a minimum exists and is unique.

Definition 6.6. If \mathcal{D} is a GD, define $W_{\mathcal{D}} : H_{\text{div}}(\Omega) \rightarrow [0, \infty[$ by

$$W_{\mathcal{D}}(\mathbf{v}) = \sup_{\phi \in X_{\mathcal{D}} \setminus \{0\}} \frac{\left| \int_{\Omega} (\nabla_{\mathcal{D}}\phi \cdot \mathbf{v} + \Pi_{\mathcal{D}}\phi \nabla \cdot \mathbf{v}) \, d\mathbf{x} \right|}{\|\nabla_{\mathcal{D}}\phi\|_{L^2(\Omega)^d}}. \quad (6.4)$$

A sequence $(\mathcal{D}_n)_{n \in \mathbb{N}}$ of GD is limit-conforming if $\lim_{n \rightarrow \infty} W_{\mathcal{D}_n}(\mathbf{v}) = 0$ for all $\mathbf{v} \in H_{\text{div}}(\Omega)$.

The limit-conformity implies that the gradient discretization method satisfies asymptotically the divergence theorem.

Definition 6.7. A sequence $(\mathcal{D}_n)_{n \in \mathbb{N}}$ of GD is compact if, for any sequence $\phi_n \in X_{\mathcal{D}_n}$ such that $(\|\nabla_{\mathcal{D}_n}\phi_n\|_{L^2(\Omega)^d})_{n \in \mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathcal{D}_n}\phi_n)_{n \in \mathbb{N}}$ is relatively compact in $L^2(\Omega)$.

In order to use the GD to solve (6.2) it is useful to write $f \in H^{-1}(\Omega)$ as

$$f = f_0 + \sum_{i=1}^d \frac{\partial f_i}{\partial x_i} = f_0 + \nabla \cdot \mathbf{F}, \quad (6.5)$$

where $\mathbf{x} = (x_1, \dots, x_d) \in \Omega$, $f_0, f_1, \dots, f_d \in L^2(\Omega)$ and $\mathbf{F} = (f_1, \dots, f_d)^T \in L^2(\Omega)^d$. With this notation, (6.2) becomes

$$\int_{\Omega} A \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} (f_0 v - \mathbf{F} \cdot \nabla v) \, d\mathbf{x} \quad \forall v \in H_0^1(\Omega). \quad (6.6)$$

We next define the GS used to approximate u solution of (6.6).

Definition 6.8. For a given GD \mathcal{D} , the GS for problem (6.6) is defined by: find $\vartheta \in X_{\mathcal{D}}$ such that

$$\int_{\Omega} A \nabla_{\mathcal{D}} \vartheta \cdot \nabla_{\mathcal{D}} \phi \, d\mathbf{x} = \int_{\Omega} (f_0 \Pi_{\mathcal{D}} \phi - \mathbf{F} \cdot \nabla_{\mathcal{D}} \phi) \, d\mathbf{x} \quad \forall \phi \in X_{\mathcal{D}}. \quad (6.7)$$

In a similar fashion, using (6.5), problem (6.3) can be written as

$$\int_{\Omega} A(u) \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} (f_0 v - \mathbf{F} \cdot \nabla v) \, d\mathbf{x} \quad \forall v \in H_0^1(\Omega) \quad (6.8)$$

and the associated GS is given by the next definition.

Definition 6.9. For a given GD \mathcal{D} , the GS for problem (6.8) is defined by: find $\vartheta \in X_{\mathcal{D}}$ such that

$$\int_{\Omega} A(\Pi_{\mathcal{D}} \vartheta) \nabla_{\mathcal{D}} \vartheta \cdot \nabla_{\mathcal{D}} \phi \, d\mathbf{x} = \int_{\Omega} (f_0 \Pi_{\mathcal{D}} \phi - \mathbf{F} \cdot \nabla_{\mathcal{D}} \phi) \, d\mathbf{x} \quad \forall \phi \in X_{\mathcal{D}}. \quad (6.9)$$

6.1.2 A priori error analysis for linear and quasilinear elliptic equations

We summarize the convergence properties of the gradient scheme for linear and quasilinear equations in this section. Theorem 6.10 below is the main tool to prove convergence of the gradient scheme (6.7) for linear equations, the proof is mainly taken from [41, Theorem 2.28] and reported here for completeness.

Theorem 6.10. *Let \mathcal{D} be a GD, then there exists one and only one $\vartheta \in X_{\mathcal{D}}$ solution to (6.7) and it satisfies*

$$\|\nabla u - \nabla_{\mathcal{D}} \vartheta\|_{L^2(\Omega)^d} \leq \frac{1}{\lambda} W_{\mathcal{D}}(A \nabla u + \mathbf{F}) + (1 + \kappa(A)) S_{\mathcal{D}}(u), \quad (6.10)$$

where $\kappa(A) = \bar{\lambda}/\underline{\lambda}$ is the condition number of A .

Proof. We start proving that if $\vartheta \in X_{\mathcal{D}}$ satisfies (6.7) then (6.10) holds. Notice that under Assumption 6.1 we have $A \nabla u + \mathbf{F} \in H_{\text{div}}(\Omega)$, indeed $-\nabla \cdot (A \nabla u + \mathbf{F}) = f_0 \in L^2(\Omega)$ by (6.6). Taking $\mathbf{v} = A \nabla u + \mathbf{F}$, thus $\nabla \cdot \mathbf{v} = -f_0$, (6.4) yields

$$\left| \int_{\Omega} (\nabla_{\mathcal{D}} \psi \cdot (A \nabla u + \mathbf{F}) - \Pi_{\mathcal{D}} \psi f_0) \, d\mathbf{x} \right| \leq W_{\mathcal{D}}(A \nabla u + \mathbf{F}) \|\nabla_{\mathcal{D}} \psi\|_{L^2(\Omega)^d}$$

for all $\psi \in X_{\mathcal{D}}$. From (6.7) follows

$$\left| \int_{\Omega} \nabla_{\mathcal{D}} \psi \cdot (A \nabla u - A \nabla_{\mathcal{D}} \vartheta) \, d\mathbf{x} \right| \leq W_{\mathcal{D}}(A \nabla u + \mathbf{F}) \|\nabla_{\mathcal{D}} \psi\|_{L^2(\Omega)^d}$$

Let $\varphi \in X_{\mathcal{D}}$ such that $\|\nabla_{\mathcal{D}} \varphi - \nabla u\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}}(u)$ (see Definition 6.5), using the symmetricity of A and the triangular inequality gives

$$\begin{aligned} \left| \int_{\Omega} A \nabla_{\mathcal{D}} \psi \cdot (\nabla_{\mathcal{D}} \varphi - \nabla_{\mathcal{D}} \vartheta) \, d\mathbf{x} \right| &\leq W_{\mathcal{D}}(A \nabla u + \mathbf{F}) \|\nabla_{\mathcal{D}} \psi\|_{L^2(\Omega)^d} \\ &\quad + \left| \int_{\Omega} A \nabla_{\mathcal{D}} \psi \cdot (\nabla_{\mathcal{D}} \varphi - \nabla u) \, d\mathbf{x} \right| \\ &\leq (W_{\mathcal{D}}(A \nabla u + \mathbf{F}) + \bar{\lambda} \|\nabla_{\mathcal{D}} \varphi - \nabla u\|_{L^2(\Omega)^d}) \|\nabla_{\mathcal{D}} \psi\|_{L^2(\Omega)^d} \\ &\leq (W_{\mathcal{D}}(A \nabla u + \mathbf{F}) + \bar{\lambda} S_{\mathcal{D}}(u)) \|\nabla_{\mathcal{D}} \psi\|_{L^2(\Omega)^d} \end{aligned}$$

Letting $\psi = \varphi - \vartheta$ yields

$$\|\nabla_{\mathcal{D}}\varphi - \nabla_{\mathcal{D}}\vartheta\|_{L^2(\Omega)^d} \leq \frac{1}{\lambda} W_{\mathcal{D}}(A\nabla u + \mathbf{F}) + \frac{\bar{\lambda}}{\lambda} S_{\mathcal{D}}(u).$$

By the triangular inequality $\|\nabla u - \nabla_{\mathcal{D}}\vartheta\|_{L^2(\Omega)^d} \leq \|\nabla_{\mathcal{D}}\varphi - \nabla u\|_{L^2(\Omega)^d} + \|\nabla_{\mathcal{D}}\varphi - \nabla_{\mathcal{D}}\vartheta\|_{L^2(\Omega)^d}$ we obtain (6.10). Now, let us prove that a unique solution to (6.7) exists. We equip the finite dimensional real vector space $X_{\mathcal{D}}$ with the inner product

$$(\phi, \psi)_{\mathcal{D}} = \int_{\Omega} \nabla_{\mathcal{D}}\phi \cdot \nabla_{\mathcal{D}}\psi \, d\mathbf{x},$$

which induces the norm $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$; therefore, $X_{\mathcal{D}}$ is an Hilbert space. The functional

$$\mathcal{F}(\phi) = \int_{\Omega} (f_0 \Pi_{\mathcal{D}}\phi - \mathbf{F} \cdot \nabla_{\mathcal{D}}\phi) \, d\mathbf{x}$$

is linear and bounded, thanks to the coercivity of the GD scheme. The bilinear form

$$a(\vartheta, \phi) = \int_{\Omega} A\nabla_{\mathcal{D}}\vartheta \cdot \nabla_{\mathcal{D}}\phi \, d\mathbf{x}$$

is bounded and coercive thanks to Assumption 6.1. From the Lax-Milgram theorem a solution to (6.7) exists and is unique. \blacksquare

Corollary 6.11. *If $(\mathcal{D}_n)_{n \in \mathbb{N}}$ is a space-consistent and limit-conforming sequence of GD and $\vartheta_n \in \mathcal{D}_n$ is a sequence of solutions to (6.7), then*

$$\lim_{n \rightarrow \infty} \|\nabla u - \nabla_{\mathcal{D}_n}\vartheta_n\|_{L^2(\Omega)^d} = 0.$$

Proof. Follows from (6.10) and Definitions 6.5 and 6.6. \blacksquare

Convergence rates are obtained under stronger regularity hypothesis on the data and the solution, upon the introduction of a mesh and depend on the approximation properties of the GD. We refer to Corollary 6.18 at the end of Section 6.2 for such results in the case of the WDGGD. The compactness hypothesis of Definition 6.7 is needed to establish convergence of the GS when applied to nonlinear problems.

The next theorem states the convergence of the GS (6.9) for quasilinear equations, for the proof we refer to [41, Theorem 2.35].

Theorem 6.12. *Let $(\mathcal{D}_n)_{n \in \mathbb{N}}$ be a space-consistent, limit-conforming and compact sequence of GD. Then, for any $n \in \mathbb{N}$ there exists a unique solution $\vartheta_n \in \mathcal{D}_n$ to (6.9). Furthermore, the sequence $\Pi_{\mathcal{D}_n}\vartheta_n$ converges strongly in $L^2(\Omega)$ to the solution u of (6.8) and $\nabla_{\mathcal{D}_n}\vartheta_n$ converges strongly in $L^2(\Omega)^d$ to ∇u , as $n \rightarrow \infty$.*

6.2 The weighted discontinuous Galerkin gradient discretization

Inspired from the Discontinuous Galerkin Gradient Discretization (DGGD) introduced in [52] we define the Weighted Discontinuous Galerkin Gradient Discretization (WDGGD).

6.2.1 The gradient discretization

Here we define a GD for which the associated GS for (6.6) is equivalent to the SWIP scheme.

A polytopal mesh $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P})$ is defined as follows. \mathcal{M} is a finite family of non empty polytopal open disjoint elements $K \subset \Omega$ such that $\overline{\Omega} = \cup_{K \in \mathcal{M}} \overline{K}$. We suppose that K is star shaped with respect to an $\mathbf{x}_K \in K$ and denote $\mathcal{P} = (\mathbf{x}_K)_{K \in \mathcal{M}}$. Let $\mathcal{F} = \mathcal{F}_b \cup \mathcal{F}_i$ be the set of faces of the mesh, where $\mathcal{F}_b, \mathcal{F}_i$ are the boundary and internal faces, respectively. The set of faces of K is $\mathcal{F}_K = \{\sigma \in \mathcal{F} : \sigma \subset \partial K\}$. For each $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$ we denote by $d_{K,\sigma}$ the orthogonal distance between \mathbf{x}_K and σ , hence

$$d_{K,\sigma} = (\mathbf{y} - \mathbf{x}_K) \cdot \mathbf{n}_{K,\sigma} \quad \forall \mathbf{y} \in \sigma,$$

where $\mathbf{n}_{K,\sigma}$ is the unit vector normal to σ outward to K . We denote by $D_{K,\sigma}$ the cone with vertex \mathbf{x}_K and basis σ , that is

$$D_{K,\sigma} = \{\mathbf{x}_K + s(\mathbf{y} - \mathbf{x}_K) : s \in]0, 1[, \mathbf{y} \in \sigma\}.$$

Finally, we define the mesh size and a constant measuring the regularity of the mesh. For $\sigma \in \mathcal{F}$ let $\mathcal{M}_\sigma = \{K \in \mathcal{M} : \sigma \in \mathcal{F}_K\}$ and let h_K be the diameter of $K \in \mathcal{M}$, then

$$\begin{aligned} h_{\mathcal{M}} &= \max \{h_K : K \in \mathcal{M}\}, \\ \eta_{\mathfrak{T}} &= \max \left(\left\{ \frac{h_T}{h_K} + \frac{h_K}{h_T} : \sigma \in \mathcal{F}_i, \mathcal{M}_\sigma = \{K, T\} \right\} \cup \left\{ \frac{h_K}{d_{K,\sigma}} : K \in \mathcal{M}, \sigma \in \mathcal{F}_K \right\} \right. \\ &\quad \left. \cup \{\#\mathcal{F}_K : K \in \mathcal{M}\} \right), \end{aligned}$$

the term $\{\#\mathcal{F}_K : K \in \mathcal{M}\}$ is needed in [52, Lemma 3.14] to bound the jumps on the faces of the elements.

Let $V = \{v \in L^2(\Omega) : v|_K \in \mathbb{P}_\ell(K), \forall K \in \mathcal{M}\}$, where $\mathbb{P}_\ell(K)$ is the space of polynomials in K of total degree ℓ . Let $(e_i)_{i \in I}$ be a basis of V such that $\text{supp}(e_i)$ is restricted to one element of \mathcal{M} . We set

$$X_{\mathcal{D}} = \{\phi = (\zeta_i)_{i \in I} : \zeta_i \in \mathbb{R} \text{ for all } i \in I\} \quad (6.11)$$

and define the operator $\Pi_{\mathcal{D}} : X_{\mathcal{D}} \rightarrow L^2(\Omega)$ by

$$\Pi_{\mathcal{D}}\phi = \sum_{i \in I} \zeta_i e_i. \quad (6.12)$$

For $K \in \mathcal{M}$ we note by $\Pi_{\overline{K}}\phi := \Pi_{\mathcal{D}}\phi|_{\overline{K}}$ the restriction of $\Pi_{\mathcal{D}}\phi$ to K extended to \overline{K} and define $\nabla_{\overline{K}}\phi = \nabla \Pi_{\overline{K}}\phi$. Let $\alpha \in]0, 1[$ be a user parameter and $\eta : [0, 1] \rightarrow \mathbb{R}$ such that $\eta(s) = 0$ on $[0, \alpha[$ and $\eta|_{[\alpha, 1]} \in \mathbb{P}_{\ell-1}([\alpha, 1])$ satisfying

$$\int_0^1 \eta(s)s^{d-1}ds = 1 \quad \text{and} \quad \int_0^1 (1-s)^i \eta(s)s^{d-1}ds = 0 \quad \text{for } i = 1, \dots, \ell-1. \quad (6.13)$$

In the case where $\ell = 1$ we have $\eta(s)|_{[\alpha, 1]} = d/(1-\alpha^d)$. This choice of η is fundamental to show the equivalence with the SWIP method, see Section 6.2.3. The discrete gradient $\nabla_{\mathcal{D}} : X_{\mathcal{D}} \rightarrow L^2(\Omega)^d$ is defined as follows. For $\phi \in X_{\mathcal{D}}$, $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, we set, for a.e. $\mathbf{x} \in D_{K,\sigma}$

$$\nabla_{\mathcal{D}}\phi(\mathbf{x}) = \nabla_{\overline{K}}\phi(\mathbf{x}) + \eta(s) \frac{[\![\phi]\!]_{K,\sigma}(\mathbf{y})}{d_{K,\sigma}} \mathbf{n}_{K,\sigma}, \quad (6.14)$$

where $\mathbf{x} = \mathbf{x}_K + s(\mathbf{y} - \mathbf{x}_K)$ with $s \in]0, 1[, \mathbf{y} \in \sigma$ and

- if $\sigma \in \mathcal{F}_i$ with $\sigma = \partial K \cap \partial T$ then $[\![\phi]\!]_{K,\sigma}(\mathbf{y}) = \omega_{K,\sigma}(\Pi_{\overline{T}}\phi(\mathbf{y}) - \Pi_{\overline{K}}\phi(\mathbf{y}))$,
- if $\sigma \in \mathcal{F}_b$ with $\sigma = \partial K \cap \partial \Omega$ then $[\![\phi]\!]_{K,\sigma}(\mathbf{y}) = 0 - \Pi_{\overline{K}}\phi(\mathbf{y})$.

For $\sigma \in \mathcal{F}_b$ with $\sigma = \partial K \cap \partial \Omega$ and $K \in \mathcal{M}$ it is useful to set $\omega_{K,\sigma} = 1$. If instead $\sigma \in \mathcal{F}_i$ with $\sigma = \partial K \cap \partial T$ and $K, T \in \mathcal{M}$ the weights $\omega_{K,\sigma}, \omega_{T,\sigma}$ are two non negative numbers such that

$$\omega_{K,\sigma} + \omega_{T,\sigma} = 1. \quad (6.15)$$

In the original DGGD introduced in [52] the weights are $(\omega_{K,\sigma}, \omega_{T,\sigma}) = (1/2, 1/2)$ and it is proven that $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$, with $X_{\mathcal{D}}$, $\Pi_{\mathcal{D}}$, $\nabla_{\mathcal{D}}$ as in (6.11), (6.12) and (6.14), is a GD. Moreover, any sequence $(\mathcal{D}_n)_{n \in \mathbb{N}}$ of DGGD defined from polytopal meshes $(\mathfrak{T}_n)_{n \in \mathbb{N}}$ with $(\eta_{\mathfrak{T}_n})_{n \in \mathbb{N}}$ bounded and $h_{\mathcal{M}_n} \rightarrow 0$ is a coercive, space-consistent, limit-conforming and compact sequence of GD. Thanks to the particular choice of η in (6.13) it is possible to show that in the linear case with piecewise constant diffusion the DGGD scheme is equivalent to the well known SIP method.

In our case, we want to be equivalent to the SWIP method, hence we define the weights as follows. Let $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, we set

$$\delta_{K,\sigma} = \mathbf{n}_{K,\sigma}^\top A|_K \mathbf{n}_{K,\sigma}.$$

For $\sigma \in \mathcal{F}_i$ such that $\sigma = \partial K \cap \partial T$ with $K, T \in \mathcal{M}$ we define

$$\omega_{K,\sigma} = \frac{\delta_{T,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}, \quad \omega_{T,\sigma} = \frac{\delta_{K,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}. \quad (6.16)$$

Upon changing the constants in [52, Lemma 3.8] we deduce from [52, Lemma 3.10] that $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ with the choice of weights given by (6.16) is a norm on $X_{\mathcal{D}}$ and hence $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ with $(\omega_{K,\sigma}, \omega_{T,\sigma})$ as in (6.16) is a GD. It can be used to solve diffusion problems with homogeneous boundary conditions as in Definitions 6.8 and 6.9. From now on we refer to this GD as the Weighted DGGD (WDGGD). Apart from the weights definition, the only difference with respect to the DGGD is a factor

$$C_\omega := \frac{1}{2} \max_{K \in \mathcal{M}, \sigma \in \mathcal{F}_K} \omega_{K,\sigma}^{-1} \quad (6.17)$$

multiplying the constant $C_{\mathcal{D}}$ of Definition 6.4.

6.2.2 Analysis of approximation properties

In the foregoing analysis we need the jump seminorm on $X_{\mathcal{D}}$, defined by

$$|\phi|_J^2 := \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} [\![\phi]\!]_{K,\sigma}^2(\mathbf{y}) d\mathbf{y}.$$

We define a stronger version of $S_{\mathcal{D}}$ which controls the jumps.

Definition 6.13. If \mathcal{D} is a WDGGD, define $S_{\mathcal{D},J} : H_0^1(\Omega) \rightarrow [0, \infty[$ by

$$S_{\mathcal{D},J}(v) := \min_{\phi \in X_{\mathcal{D}}} (\|\Pi_{\mathcal{D}}\phi - v\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}\phi - \nabla v\|_{L^2(\Omega)^d} + |\phi|_J).$$

We quote two improved estimates on $S_{\mathcal{D}}$, $S_{\mathcal{D},J}$ and $W_{\mathcal{D}}$.

Lemma 6.14. There exists $C_S > 0$ depending only on $|\Omega|$, α , ℓ , d and $\eta_{\mathfrak{T}}$ such that for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$

$$S_{\mathcal{D}}(v) \leq C_S h_{\mathcal{M}} \|v\|_{H^2(\Omega)} \quad \text{and} \quad S_{\mathcal{D},J}(v) \leq C_S h_{\mathcal{M}} \|v\|_{H^2(\Omega)}.$$

The result for $S_{\mathcal{D},J}$ is obtained following the lines of the proof for $S_{\mathcal{D}}$, which is given in [52, Lemma 3.14].

Lemma 6.15. *There exists $C_W > 0$ depending only on $|\Omega|$, α , ℓ , d and $\eta_{\mathfrak{T}}$ such that for all $\mathbf{v} \in H^1(\Omega)^d$*

$$W_{\mathcal{D}}(\mathbf{v}) \leq C_W h_{\mathcal{M}} \|\mathbf{v}\|_{H^1(\Omega)^d}.$$

Lemma 6.15 has been proven for the DGGD in [52, Lemma 3.15]. The proof uses the fact that $(1/2, 1/2)$ is a partition of unity. Thanks to (6.15) the same result holds for the WDGGD. Next, Theorem 6.16 establishes the asymptotic properties of the WDGGD.

Theorem 6.16. *Let $(\mathcal{D}_n)_{n \in \mathbb{N}}$ be a sequence of polytopal meshes $(\mathfrak{T}_n)_{n \in \mathbb{N}}$ with $(\eta_{\mathfrak{T}_n})_{n \in \mathbb{N}}$ bounded and $h_{\mathcal{M}_n} \rightarrow 0$ for $n \rightarrow \infty$. Then it is a coercive, space-consistent, limit-conforming and compact sequence of GD.*

Proof. Coercivity and compactness are proven as in [52, Lemma 3.12, Lemma 3.13]. Space-consistency follows from Lemma 6.14 and [41, Lemma 2.16]. Limit-conformity follows from the compactness of the scheme, Lemma 6.15 and [41, Lemma 2.17]. ■

In the WDGGD the C_p constant in Definition 6.4 depends continuously on C_ω from (6.17). We note that, even if C_ω is mesh dependent it can be bounded by terms depending only on A . In the following lemma we show, by usual density arguments, that even if v is only in $H_0^1(\Omega)$ we have $\lim_{n \rightarrow \infty} S_{\mathcal{D}_n, J}(v) = 0$.

Lemma 6.17. *Consider the same assumptions of Theorem 6.16 and $v \in H_0^1(\Omega)$. Then we have $\lim_{n \rightarrow \infty} S_{\mathcal{D}_n, J}(v) = 0$.*

Proof. Let $v \in H_0^1(\Omega)$ and $\varepsilon > 0$. Then there exists $v_\varepsilon \in H^2(\Omega) \cap H_0^1(\Omega)$ such that $\|v - v_\varepsilon\|_{L^2(\Omega)} + \|\nabla v - \nabla v_\varepsilon\|_{L^2(\Omega)^d} \leq \varepsilon$. Let

$$\phi_n = \operatorname{argmin}_{\phi \in X_{\mathcal{D}_n}} (\|\Pi_{\mathcal{D}_n} \phi - v_\varepsilon\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}_n} \phi - \nabla v_\varepsilon\|_{L^2(\Omega)^d} + |\phi|_J).$$

Hence

$$S_{\mathcal{D}_n, J}(v) \leq \|\Pi_{\mathcal{D}_n} \phi_n - v\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}_n} \phi_n - \nabla v\|_{L^2(\Omega)^d} + |\phi_n|_J \leq \varepsilon + C_S h_{\mathcal{M}_n} \|v_\varepsilon\|_{H^2(\Omega)},$$

using Lemma 6.14 $\lim_{n \rightarrow \infty} S_{\mathcal{D}_n, J}(v) \leq \varepsilon$. Since ε is arbitrary the result follows. ■

Corollary 6.18 (Of Theorem 6.10). *Let \mathcal{D} be a WDGGD, under the same assumptions of Theorem 6.10, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $\mathbf{F} \in H^1(\Omega)^d$, the solution $\vartheta \in X_{\mathcal{D}}$ to (6.7) satisfies*

$$\|\nabla u - \nabla_{\mathcal{D}} \vartheta\|_{L^2(\Omega)^d} \leq h_{\mathcal{M}} \left(\frac{1}{\lambda} C_W \|A \nabla u + \mathbf{F}\|_{H^1(\Omega)^d} + (1 + \kappa(A)) C_S \|u\|_{H^2(\Omega)} \right).$$

Proof. Follows from Theorem 6.10 together with Lemmas 6.14 and 6.15. ■

6.2.3 Equivalence to the symmetric weighted interior penalty method

Here we show that for the WDGGD the GS scheme of Definition 6.8 is equivalent to the SWIP scheme [39, equation 4.63] (see also [49]). In order to do that, we follow [52], where the equivalence of the GS corresponding to the DGGD is shown to be equivalent to the SIP method. We suppose that $A(\mathbf{x}, u) = A(\mathbf{x})$ and $A_K := A|_K$ the restriction of A to an element $K \in \mathcal{M}$ is constant, that $f \in L^2(\Omega)$ and hence $f = f_0$.

Starting from (6.7) and developing the discrete gradients (6.14) we get

$$\begin{aligned}
 & \int_{\Omega} A \nabla_{\mathcal{D}} \vartheta \cdot \nabla_{\mathcal{D}} \phi \, d\mathbf{x} \\
 &= \sum_{K \in \mathcal{M}} \int_K A_K \nabla_{\bar{K}} \vartheta \cdot \nabla_{\bar{K}} \phi \, d\mathbf{x} \\
 &\quad + \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K,\sigma}} \frac{\eta(s)}{d_{K,\sigma}} A_K (\llbracket \vartheta \rrbracket_{K,\sigma}(\mathbf{y}) \nabla_{\bar{K}} \phi(\mathbf{x}) + \llbracket \phi \rrbracket_{K,\sigma}(\mathbf{y}) \nabla_{\bar{K}} \vartheta(\mathbf{x})) \cdot \mathbf{n}_{K,\sigma} \, d\mathbf{x} \\
 &\quad + \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K,\sigma}} \frac{A_K \mathbf{n}_{K,\sigma} \cdot \mathbf{n}_{K,\sigma}}{d_{K,\sigma}^2} \eta(s)^2 \llbracket \vartheta \rrbracket_{K,\sigma}(\mathbf{y}) \llbracket \phi \rrbracket_{K,\sigma}(\mathbf{y}) \, d\mathbf{x} \\
 &= I_1 + I_2 + I_3.
 \end{aligned}$$

Since $\mathbf{x} = \mathbf{x}_K + s(\mathbf{y} - \mathbf{x}_K)$ for $s \in]0, 1[$, $\mathbf{y} \in \sigma$ and $\nabla_{\bar{K}} \vartheta \in \mathbb{P}_{\ell-1}(K)^d$ then

$$\nabla_{\bar{K}} \vartheta(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} = \nabla_{\bar{K}} \vartheta(\mathbf{y}) \cdot \mathbf{n}_{K,\sigma} + \sum_{j=1}^{\ell-1} p_j(\mathbf{y})(1-s)^j,$$

with $p_j(\mathbf{y})$ polynomials of degree $\ell-1$ in the components of \mathbf{y} . It follows from (6.13) that

$$\int_0^1 \nabla_{\bar{K}} \vartheta(\mathbf{x}) \cdot \mathbf{n}_{K,\sigma} s^{d-1} \eta(s) \, ds = \nabla_{\bar{K}} \vartheta(\mathbf{y}) \cdot \mathbf{n}_{K,\sigma},$$

hence, using the change of variables $d\mathbf{x} = s^{d-1} d_{K,\sigma} \, ds \, d\mathbf{y}$ we get

$$I_2 = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} A_K (\llbracket \vartheta \rrbracket_{K,\sigma}(\mathbf{y}) \nabla_{\bar{K}} \phi(\mathbf{y}) + \llbracket \phi \rrbracket_{K,\sigma}(\mathbf{y}) \nabla_{\bar{K}} \vartheta(\mathbf{y})) \cdot \mathbf{n}_{K,\sigma} \, d\mathbf{y}.$$

For $\sigma \in \mathcal{F}_i$ with $\sigma = \partial K \cap \partial T$ let $\mathbf{n}_{\sigma} = \mathbf{n}_{K,\sigma}$ and

$$[\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} = \Pi_{\bar{K}} \vartheta - \Pi_{\bar{T}} \vartheta, \quad \{ A \nabla \Pi_{\mathcal{D}} \vartheta \}_{\omega,\sigma} = \omega_{K,\sigma} A|_K \nabla \Pi_{\bar{K}} \vartheta + \omega_{T,\sigma} A|_T \nabla \Pi_{\bar{T}} \vartheta.$$

If $\sigma \in \mathcal{F}_b$ with $\sigma = \partial K \cap \partial \Omega$ let $\mathbf{n}_{\sigma} = \mathbf{n}_{K,\sigma}$ and

$$[\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} = \Pi_{\bar{K}} \vartheta, \quad \{ A \nabla \Pi_{\mathcal{D}} \vartheta \}_{\omega,\sigma} = A|_K \nabla \Pi_{\bar{K}} \vartheta.$$

It holds $\llbracket \vartheta \rrbracket_{K,\sigma} \cdot \mathbf{n}_{K,\sigma} = -\omega_{K,\sigma} [\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} \cdot \mathbf{n}_{\sigma}$ and similarly for ϕ , hence

$$\begin{aligned}
 I_2 &= - \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} \omega_{K,\sigma} ([\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} A_K \nabla \Pi_{\bar{K}} \phi + [\![\Pi_{\mathcal{D}} \phi]\!]_{\sigma} A_K \nabla \Pi_{\bar{K}} \vartheta) \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \\
 &= - \sum_{\sigma \in \mathcal{F}} \int_{\sigma} ([\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} \{ A \nabla \Pi_{\mathcal{D}} \phi \}_{\omega,\sigma} + [\![\Pi_{\mathcal{D}} \phi]\!]_{\sigma} \{ A \nabla \Pi_{\mathcal{D}} \vartheta \}_{\omega,\sigma}) \cdot \mathbf{n}_{\sigma} \, d\mathbf{y}.
 \end{aligned}$$

For I_3 , using $C_{\eta}^2 = \int_{\alpha}^1 \eta(s)^2 s^{d-1} \, ds$ and by the change of variables $d\mathbf{x} = s^{d-1} d_{K,\sigma} \, ds \, d\mathbf{y}$, we obtain

$$I_3 = C_{\eta}^2 \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \frac{\delta_{K,\sigma}}{d_{K,\sigma}} \omega_{K,\sigma}^2 \int_{\sigma} [\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} [\![\Pi_{\mathcal{D}} \phi]\!]_{\sigma} \, d\mathbf{y} = \sum_{\sigma \in \mathcal{F}} \eta_{\sigma} \frac{\gamma_{\sigma}}{h_{\sigma}} \int_{\sigma} [\![\Pi_{\mathcal{D}} \vartheta]\!]_{\sigma} [\![\Pi_{\mathcal{D}} \phi]\!]_{\sigma} \, d\mathbf{y},$$

where h_{σ} is the diameter of σ and $\gamma_{\sigma}, \eta_{\sigma}$ for $\sigma \in \mathcal{F}_i$ are defined by

$$\begin{aligned}
 \gamma_{\sigma} &= \frac{2 \delta_{K,\sigma} \delta_{T,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}, \\
 \eta_{\sigma} &= C_{\eta}^2 \left(\frac{\delta_{K,\sigma}}{d_{K,\sigma}} \omega_{K,\sigma}^2 + \frac{\delta_{T,\sigma}}{d_{T,\sigma}} \omega_{T,\sigma}^2 \right) \frac{h_{\sigma}}{\gamma_{\sigma}} = C_{\eta}^2 h_{\sigma} \left(\frac{\omega_{K,\sigma}}{d_{K,\sigma}} + \frac{\omega_{T,\sigma}}{d_{T,\sigma}} \right)
 \end{aligned} \tag{6.18}$$

and for $\sigma \in \mathcal{F}_b$ by

$$\gamma_\sigma = \delta_{K,\sigma}, \quad \eta_\sigma = C_\eta^2 \frac{h_\sigma}{d_{K,\sigma}}.$$

Summing $I_1 + I_2 + I_3$ yields

$$\begin{aligned} \int_{\Omega} A \nabla_{\mathcal{D}} \vartheta \cdot \nabla_{\mathcal{D}} \phi \, d\mathbf{x} &= \sum_{K \in \mathcal{M}} \int_K A_K \nabla_{\bar{K}} \vartheta \cdot \nabla_{\bar{K}} \phi \, d\mathbf{x} \\ &\quad - \sum_{\sigma \in \mathcal{F}} \int_{\sigma} ([\Pi_{\mathcal{D}} \vartheta] \{A \nabla \Pi_{\mathcal{D}} \phi\}_{\omega} + [\Pi_{\mathcal{D}} \phi] \{A \nabla \Pi_{\mathcal{D}} \vartheta\}_{\omega}) \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \quad (6.19) \\ &\quad + \sum_{\sigma \in \mathcal{F}} \eta_{\sigma} \frac{\gamma_{\sigma}}{h_{\sigma}} \int_{\sigma} [\Pi_{\mathcal{D}} \vartheta] [\Pi_{\mathcal{D}} \phi] \, d\mathbf{y}, \end{aligned}$$

where σ in $[\cdot]_{\sigma}$, $\{\cdot\}_{\omega,\sigma}$ is left out of notation for simplicity. Therefore, we get the equivalence of (6.19) and [39, equation 4.64], where the parameter η_{σ} chosen as in (6.18). Under the additional hypothesis that the mesh sequence satisfies

$$\min \left\{ \frac{h_{\sigma}}{d_{K,\sigma}} : K \in \mathcal{M}, \sigma \in \mathcal{F}_K \right\} \geq C_{\mathcal{F}} > 0,$$

we have that $\eta_{\sigma} \geq C_{\eta}^2 C_{\mathcal{F}}$. Since $C_{\eta}^2 \geq d/(1 - \alpha^d)$, letting $\alpha \rightarrow 1$ we can have the penalty parameter η_{σ} as large as desired.

7 A local weighted discontinuous Galerkin gradient discretization scheme for linear and quasilinear elliptic equations

Based on the Weighted Discontinuous Galerkin Gradient Discretization (WDGGD) introduced in Chapter 6, here we define and analyze a local WDGGD for linear and quasilinear elliptic equations as (6.1). The local scheme is based on a coarse grid and successively improves the solution solving a sequence of local elliptic problems in under resolved regions. Using the GDM we prove convergence of the scheme for linear and quasilinear equations under minimal regularity assumptions. In a particular case, the error due to artificial boundary conditions is also analyzed, shown to be of higher order and shown to depend only locally on the regularity of the solution.

We begin with some preliminary definitions and results in Section 7.1. The local WDGGD for linear elliptic equations (6.2) is defined and analyzed in Section 7.2, the scheme for quasilinear equations (6.3) is defined and analyzed in Section 7.3. Numerical experiments are presented in Section 7.4, where we confirm the theoretical findings and the local method's accuracy is compared against the non local approach.

7.1 Notation and preliminary results

Here we define the local gradient discretization and analyze its approximation properties. The associated local gradient scheme is defined in the next section. The local domains Ω_k identifying the under resolved regions defined here below are supposed to be known; we are aware that this is not always possible in practical applications and therefore in Chapter 8 we derive a posteriori error estimators used to mark the local domains Ω_k when they are not known beforehand.

Let $\{\Omega_k\}_{k=1}^M$ be a sequence of polytopal domains with $\Omega_1 = \Omega$ and $\Omega_k \subset \Omega$. We consider as well a sequence $(\mathfrak{T}_k)_{k=1}^M = ((\mathcal{M}_k, \mathcal{F}_k, \mathcal{P}_k))_{k=1}^M$ of polytopal meshes on Ω and denote $\mathcal{F}_k = \mathcal{F}_{k,b} \cup \mathcal{F}_{k,i}$ with $\mathcal{F}_{k,b}$ and $\mathcal{F}_{k,i}$ the set of boundary and internal faces of \mathcal{M}_k . Moreover, $(\mathfrak{T}_k)_{k=1}^M$ satisfies the following.

Assumption 7.1.

- a) For each $k = 1, \dots, M$, $\overline{\Omega}_k = \cup_{K \in \mathcal{M}_k, K \subset \Omega_k} \overline{K}$.
- b) For $k = 1, \dots, M - 1$
 - i) $\{K \in \mathcal{M}_{k+1} : K \subset \Omega \setminus \Omega_{k+1}\} = \{K \in \mathcal{M}_k : K \subset \Omega \setminus \Omega_{k+1}\}$,
 - ii) if $K, T \in \mathcal{M}_k$ with $K \subset \Omega_{k+1}$, $T \subset \Omega \setminus \Omega_{k+1}$ and $\partial K \cap \partial T \neq \emptyset$ then $K \in \mathcal{M}_{k+1}$,
 - iii) if $K \in \mathcal{M}_k$ and $K \subset \Omega_{k+1}$, either $K \in \mathcal{M}_{k+1}$ or K is a union of elements in \mathcal{M}_{k+1} .
- c) We suppose the existence of $C_r > 0$ such that
 - i) for $k = 1, \dots, M - 1$, if $K \in \mathcal{M}_k$ and $\hat{K} \in \mathcal{M}_{k+1}$ with $\hat{K} \subset K$ and $\sigma \in \mathcal{F}_K$, $\hat{\sigma} \in \mathcal{F}_{\hat{K}}$

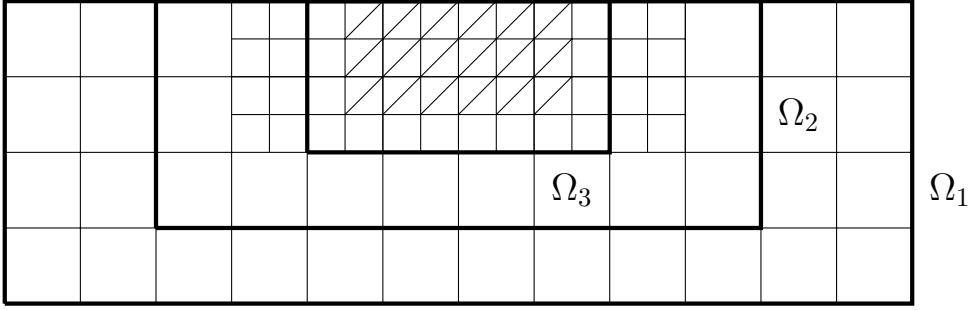


Figure 7.1. Example of possible meshes satisfying Assumption 7.1.

with $\hat{\sigma} \subset \sigma$ then $d_{K,\sigma} \leq C_r d_{\hat{K},\hat{\sigma}}$,

- ii) for $k = 1, \dots, M$, if $\sigma = \partial K \cap \partial T$ with $K, T \in \mathcal{M}_k$, $T \subset \Omega \setminus \Omega_k$ and $K \subset \Omega_k$ then $d_{K,\sigma} \leq C_r d_{T,\sigma}$.
- d) It exists $\rho > 0$ such that $\eta_{\mathfrak{T}_k} \leq \rho$ for $k = 1, \dots, M$.

The above assumptions on $(\mathfrak{T}_k)_{k=1}^M$ ensure that \mathfrak{T}_{k+1} is a refinement of \mathfrak{T}_k and that this refinement occurs inside the subdomain Ω_{k+1} . Let $\widehat{\mathfrak{T}}_k = (\widehat{\mathcal{M}}_k, \widehat{\mathcal{F}}_k, \widehat{\mathcal{P}}_k)$, with $\widehat{\mathcal{M}}_k = \{K \in \mathcal{M}_k : K \subset \Omega_k\}$, $\widehat{\mathcal{P}}_k = \{\mathbf{x}_k \in \mathcal{P}_k : \mathbf{x}_k \in \Omega_k\}$ and $\widehat{\mathcal{F}}_k = \widehat{\mathcal{F}}_{k,b} \cup \widehat{\mathcal{F}}_{k,i}$ the set of faces of $\widehat{\mathcal{M}}_k$, with $\widehat{\mathcal{F}}_{k,b}$ and $\widehat{\mathcal{F}}_{k,i}$ the boundary and internal faces of Ω_k , respectively. Condition a) in Assumption 7.1 assures that $\widehat{\mathfrak{T}}_k$ is a polytopal mesh on Ω_k . b) guarantees that in $\Omega \setminus \Omega_{k+1}$ and in the neighborhood of $\partial\Omega_{k+1}$ the meshes \mathcal{M}_k and \mathcal{M}_{k+1} are equal and that \mathcal{M}_{k+1} is a refinement of \mathcal{M}_k in Ω_{k+1} . c) and d) ensure mesh regularity, will permit equivalences between jump norms and make the constant C_S of Lemma 6.14 uniform in k . An example of meshes satisfying Assumption 7.1 is given in Figure 7.1.

Given $(\mathfrak{T}_k)_{k=1}^M$ we define a sequence $\mathcal{D}_k = (X_{\mathcal{D}_k}, \Pi_{\mathcal{D}_k}, \nabla_{\mathcal{D}_k})$ of WDGGD. Let

$$V_k = \{v_k \in L^2(\Omega) : v_k|_K \in \mathbb{P}_\ell(K), \forall K \in \mathcal{M}_k\}$$

and $(e_{k,i})_{i \in I_k}$ be a basis of V_k such that $\text{supp}(e_{k,i})$ is restricted to one element of \mathcal{M}_k . We set

$$X_{\mathcal{D}_k} = \{\phi_k = (\zeta_{k,i})_{i \in I_k} : \zeta_{k,i} \in \mathbb{R} \text{ for all } i \in I_k\}.$$

The operators $\Pi_{\mathcal{D}_k}$ and $\nabla_{\mathcal{D}_k}$ are defined as in (6.12), (6.14) and (6.16).

We can write $X_{\mathcal{D}_k} = Y_{\mathcal{D}_k} \oplus Z_{\mathcal{D}_k}$, where $\text{supp}(\Pi_{\mathcal{D}_k} \varphi_k) \subset \Omega_k$ for $\varphi_k \in Y_{\mathcal{D}_k}$ and $\text{supp}(\Pi_{\mathcal{D}_k} \xi_k) \subset \Omega \setminus \Omega_k$ for $\xi_k \in Z_{\mathcal{D}_k}$. For $k = 1$ we have $Y_{\mathcal{D}_1} = X_{\mathcal{D}_1}$ and $Z_{\mathcal{D}_1} = \{0\}$. For $k \geq 2$ and $\phi_{k-1} \in X_{\mathcal{D}_{k-1}}$ there exists $\xi_k \in Z_{\mathcal{D}_k}$ such that $\Pi_{\mathcal{D}_{k-1}} \phi_{k-1} \chi_{\Omega \setminus \Omega_k} = \Pi_{\mathcal{D}_k} \xi_k$. By abuse of notation we will denote $\xi_k = \phi_{k-1} \chi_{\Omega \setminus \Omega_k}$, hence $\chi_{\Omega \setminus \Omega_k}$ is seen as an operator from $X_{\mathcal{D}_{k-1}}$ to $Z_{\mathcal{D}_k}$.

In what follows $\Pi_{\widehat{\mathcal{D}}_k}$ is the restriction of $\Pi_{\mathcal{D}_k}$ to $Y_{\mathcal{D}_k}$. Let us define as well a gradient on $Y_{\mathcal{D}_k}$ which will be used to impose inhomogeneous Dirichlet boundary conditions. Let $\varphi_k \in Y_{\mathcal{D}_k}$ and $\xi_k \in Z_{\mathcal{D}_k}$, for $K \in \widehat{\mathcal{M}}_k$, $\sigma \in \mathcal{F}_K$ and $\mathbf{x} \in D_{K,\sigma}$ the gradient $\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k(\mathbf{x})$ is defined by

$$\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k(\mathbf{x}) = \nabla_{\overline{K}} \varphi_k(\mathbf{x}) + \eta(s) \frac{[\![\varphi_k]\!]_{K,\sigma, \xi_k}(\mathbf{y})}{d_{K,\sigma}} \mathbf{n}_{K,\sigma},$$

where $\mathbf{x} = \mathbf{x}_K + s(\mathbf{y} - \mathbf{x}_K)$ and

$$\begin{aligned}\llbracket \varphi_k \rrbracket_{K,\sigma,\xi_k}(\mathbf{y}) &= \llbracket \varphi_k \rrbracket_{K,\sigma}(\mathbf{y}) && \text{if } \sigma \in \widehat{\mathcal{F}}_{k,i} \text{ or } \sigma \in \widehat{\mathcal{F}}_{k,b} \cap \mathcal{F}_{k,b}, \\ \llbracket \varphi_k \rrbracket_{K,\sigma,\xi_k}(\mathbf{y}) &= \Pi_{\overline{T}} \xi_k - \Pi_{\overline{K}} \varphi_k && \text{if } \sigma \in \widehat{\mathcal{F}}_{k,b} \setminus \mathcal{F}_{k,b} \text{ with } \sigma = \partial K \cap \partial T \\ &&& \text{and } K \in \widehat{\mathcal{M}}_k, T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k.\end{aligned}$$

We will denote $\nabla_{\widehat{\mathcal{D}}_k,0}$ by $\nabla_{\widehat{\mathcal{D}}_k}$.

Theorem 7.2. *The triple $\widehat{\mathcal{D}}_k = (Y_{\mathcal{D}_k}, \Pi_{\widehat{\mathcal{D}}_k}, \nabla_{\widehat{\mathcal{D}}_k})$ is a WDGGD for each $k = 1, \dots, M$.*

Proof. We notice that $\widehat{\mathcal{D}}_k$ is the WDGGD corresponding to the local polytopal mesh $\widehat{\mathfrak{T}}_k$, hence it is a WDGGD by construction. ■

In what follows we will call $\widehat{\mathcal{D}}_k$ the local WDGGD. Remark that Lemmas 6.14 and 6.15 and Theorem 6.16 are valid if we replace \mathcal{D} , Ω , $h_{\mathcal{M}}$ and \mathfrak{T} with $\widehat{\mathcal{D}}_k$, Ω_k , $h_{\widehat{\mathcal{M}}_k}$ and $\widehat{\mathfrak{T}}_k$.

Note that for $\varphi_k \in Y_{\mathcal{D}_k}$ we have $\nabla_{\widehat{\mathcal{D}}_k} \varphi_k \neq \nabla_{\mathcal{D}_k} \varphi_k$, indeed $\nabla_{\widehat{\mathcal{D}}_k}$ is missing the weight $\omega_{K,\sigma}$ in the jumps at the faces $\sigma \in \widehat{\mathcal{F}}_{k,b} \setminus \mathcal{F}_{k,b}$. Adding the weight $\omega_{K,\sigma}$ at those faces would prevent the limit-consistency of $\widehat{\mathcal{D}}_k$.

In what follows $S_{\widehat{\mathcal{D}}_k}$ and $W_{\widehat{\mathcal{D}}_k}$ are the operators defined by Definitions 6.5 and 6.6 but with Ω , \mathcal{D} , and $X_{\mathcal{D}}$ replaced by Ω_k , $\widehat{\mathcal{D}}_k$ and $Y_{\mathcal{D}_k}$. We define as well the jump seminorms on $X_{\mathcal{D}_k}$ and $Y_{\mathcal{D}_k}$. For $\phi_k \in X_{\mathcal{D}_k}$ we define

$$|\phi_k|_{J(k)}^2 := \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} \llbracket \phi_k \rrbracket_{K,\sigma}(\mathbf{y})^2 d\mathbf{y}$$

and for $\xi_k \in Z_{\mathcal{D}_k}$, $\varphi_k \in Y_{\mathcal{D}_k}$ we set

$$|\varphi_k|_{\widehat{J}(k),\xi_k}^2 := \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} \llbracket \varphi_k \rrbracket_{K,\sigma,\xi_k}(\mathbf{y})^2 d\mathbf{y}.$$

Since in our local scheme (to be defined in Section 7.2) we solve local elliptic problems with artificial boundary conditions we need a local version of $S_{\mathcal{D}_k,J}$ which measures the error of the method on the boundary.

Definition 7.3. Let $\xi_k \in Z_{\mathcal{D}_k}$ and $\widehat{\mathcal{D}}_k$ be a local WDGGD, define
 $S_{\widehat{\mathcal{D}}_k,J,\xi_k} : H_0^1(\Omega) \rightarrow [0, \infty[$ by

$$S_{\widehat{\mathcal{D}}_k,J,\xi_k}(v) := \min_{\varphi \in Y_{\mathcal{D}_k}} (\|\nabla_{\widehat{\mathcal{D}}_k,\xi_k} \varphi - \nabla v\|_{L^2(\Omega_k)^d} + |\varphi|_{\widehat{J}(k),\xi_k}).$$

The $L^2(\Omega_k)$ norm is not taken into account in $S_{\widehat{\mathcal{D}}_k,J,\xi_k}$ since our convergence results are in energy and jump norms.

Lemma 7.4. *Let $v \in H_0^1(\Omega) \cap H^2(\Omega)$, then for $k = 1, \dots, M$*

$$\min_{\xi \in Z_{\mathcal{D}_k}} S_{\widehat{\mathcal{D}}_k,J,\xi}(v) \leq C_S h_{\widehat{\mathcal{M}}_k} \|v\|_{H^2(\Omega_k)}.$$

Proof. Follows the lines of [52, Lemma 3.14]. ■

In order to provide bounds on $S_{\widehat{\mathcal{D}}_k, J, \xi_k}$ we need an additional norm to measure the error at the interface between subdomains. Let $\phi_k \in X_{\mathcal{D}_k}$, we define

$$|\phi_k|_{\partial\Omega_k^-}^2 := \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K,\sigma}} \int_{\sigma} \Pi_{\bar{T}} \phi_k(\mathbf{y})^2 d\mathbf{y}.$$

The minus sign in $|\cdot|_{\partial\Omega_k^-}$ refers to the fact that in the integral the argument lives outside $\widehat{\mathcal{M}}_k$. Later, $|\cdot|_{\partial\Omega_k^+}$ will also be defined.

Lemma 7.5. *Let $\kappa_k, \xi_k \in Z_{\mathcal{D}_k}$ and $v \in H_0^1(\Omega)$. Then*

$$S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(v) \leq S_{\widehat{\mathcal{D}}_k, J, \xi_k}(v) + C_{\partial} |\kappa_k - \xi_k|_{\partial\Omega_k^-},$$

where $C_{\partial} = 1 + C_{\eta}$ and $C_{\eta}^2 = \int_{\alpha}^1 \eta(s)^2 s^{d-1} ds$. If moreover $v \in H^2(\Omega) \cap H_0^1(\Omega)$ we have

$$S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(v) \leq C_S h_{\widehat{\mathcal{M}}_k} \|v\|_{H^2(\Omega_k)} + C_{\partial} |\kappa_k - \xi_k|_{\partial\Omega_k^-} \quad \text{for } \xi_k = \operatorname{argmin}_{\xi \in Z_{\mathcal{D}_k}} S_{\widehat{\mathcal{D}}_k, J, \xi}(v).$$

Proof. Let $\kappa_k, \xi_k \in Z_{\mathcal{D}_k}$, $v \in H_0^1(\Omega)$ and $\varphi_k \in Y_{\mathcal{D}_k}$ defined by

$$\varphi_k = \operatorname{argmin}_{\varphi \in Y_{\mathcal{D}_k}} (\|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi - \nabla v\|_{L^2(\Omega_k)^d} + |\varphi|_{\widehat{\mathcal{J}}(k), \xi_k}),$$

we have

$$\begin{aligned} & \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k\|_{L^2(\Omega_k)^d}^2 \\ &= \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K \cap \widehat{\mathcal{F}}_{k,b}} \int_{D_{K,\sigma}} \frac{\eta(s)^2}{d_{K,\sigma}^2} (\llbracket \varphi_k \rrbracket_{K,\sigma, \kappa_k}(\mathbf{y}) - \llbracket \varphi_k \rrbracket_{K,\sigma, \xi_k}(\mathbf{y}))^2 d\mathbf{x}, \end{aligned}$$

where $\mathbf{x} = \mathbf{x}_K + s(\mathbf{y} - \mathbf{x}_K)$ for $s \in [0, 1]$ and $\mathbf{y} \in \sigma$. Using the change of variables $d\mathbf{x} = d_{K,\sigma} s^{d-1} ds d\mathbf{y}$ yields

$$\begin{aligned} & \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k\|_{L^2(\Omega_k)^d}^2 \\ &= \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K \cap \widehat{\mathcal{F}}_{k,b}} \int_{\sigma} \int_{\alpha}^1 \frac{\eta(s)^2}{d_{K,\sigma}^2} (\llbracket \varphi_k \rrbracket_{K,\sigma, \kappa_k}(\mathbf{y}) - \llbracket \varphi_k \rrbracket_{K,\sigma, \xi_k}(\mathbf{y}))^2 d_{K,\sigma} s^{d-1} ds d\mathbf{y} \\ &= C_{\eta}^2 \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K \cap \widehat{\mathcal{F}}_{k,b}} \frac{1}{d_{K,\sigma}} \int_{\sigma} (\llbracket \varphi_k \rrbracket_{K,\sigma, \kappa_k}(\mathbf{y}) - \llbracket \varphi_k \rrbracket_{K,\sigma, \xi_k}(\mathbf{y}))^2 d\mathbf{y}. \end{aligned}$$

If in the above sum $\sigma \in \mathcal{F}_{k,b}$, then $\llbracket \varphi_k \rrbracket_{K,\sigma, \kappa_k} - \llbracket \varphi_k \rrbracket_{K,\sigma, \xi_k} = 0$. Else, if $\sigma \in \mathcal{F}_{k,i}$ there is $T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k$ such that $\sigma = \partial K \cap \partial T$ and

$$\llbracket \varphi_k \rrbracket_{K,\sigma, \kappa_k} - \llbracket \varphi_k \rrbracket_{K,\sigma, \xi_k} = \Pi_{\bar{T}} \kappa_k - \Pi_{\bar{T}} \xi_k,$$

which implies

$$\begin{aligned} & \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k\|_{L^2(\Omega_k)^d}^2 \\ &= C_{\eta}^2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K,\sigma}} \int_{\sigma} \Pi_{\bar{T}} (\kappa_k - \xi_k)(\mathbf{y})^2 d\mathbf{y} = C_{\eta}^2 |\kappa_k - \xi_k|_{\partial\Omega_k^-}^2. \end{aligned} \quad (7.1)$$

For the jump term $|\varphi_k|_{\widehat{J}(k), \kappa_k}$, we have

$$\begin{aligned} |\varphi_k|_{\widehat{J}(k), \kappa_k}^2 &= |\varphi_k|_{\widehat{J}(k), \xi_k}^2 \\ &+ \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K \cap \widehat{\mathcal{F}}_{k,b}} \frac{1}{d_{K,\sigma}} \int_{\sigma} ([\![\varphi_k]\!]_{K,\sigma,\kappa_k}(\mathbf{y})^2 - [\![\varphi_k]\!]_{K,\sigma,\xi_k}(\mathbf{y})^2) d\mathbf{y}. \end{aligned} \quad (7.2)$$

If $\sigma \in \mathcal{F}_{k,b}$ then $[\![\varphi_k]\!]_{K,\sigma,\kappa_k}^2 - [\![\varphi_k]\!]_{K,\sigma,\xi_k}^2 = 0$, else, if $\sigma \in \mathcal{F}_{k,i}$ with $\sigma = \partial T \cap \partial K$, $T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k$ we have

$$\begin{aligned} [\![\varphi_k]\!]_{K,\sigma,\kappa_k}^2 - [\![\varphi_k]\!]_{K,\sigma,\xi_k}^2 &= ([\![\varphi_k]\!]_{K,\sigma,\kappa_k} - [\![\varphi_k]\!]_{K,\sigma,\xi_k})([\![\varphi_k]\!]_{K,\sigma,\kappa_k} + [\![\varphi_k]\!]_{K,\sigma,\xi_k}) \\ &= (\Pi_{\bar{T}} \kappa_k - \Pi_{\bar{T}} \xi_k)(\Pi_{\bar{T}} \kappa_k - \Pi_{\bar{T}} \xi_k + 2[\![\varphi_k]\!]_{K,\sigma,\xi_k}). \end{aligned} \quad (7.3)$$

Using (7.2) and (7.3) we obtain

$$\begin{aligned} |\varphi_k|_{\widehat{J}(k), \kappa_k}^2 &\leq |\varphi_k|_{\widehat{J}(k), \xi_k}^2 + \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K,\sigma}} \int_{\sigma} \Pi_{\bar{T}}(\kappa_k - \xi_k)(\mathbf{y})^2 d\mathbf{y} \\ &+ 2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K,\sigma}} \int_{\sigma} |[\![\varphi_k]\!]_{K,\sigma,\xi_k}(\mathbf{y}) \Pi_{\bar{T}}(\kappa_k - \xi_k)(\mathbf{y})| d\mathbf{y} \\ &\leq |\varphi_k|_{\widehat{J}(k), \xi_k}^2 + |\kappa_k - \xi_k|_{\partial \Omega_k^-}^2 + 2|\varphi_k|_{\widehat{J}(k), \xi_k} |\kappa_k - \xi_k|_{\partial \Omega_k^-} \\ &= (|\varphi_k|_{\widehat{J}(k), \xi_k} + |\kappa_k - \xi_k|_{\partial \Omega_k^-})^2. \end{aligned}$$

Using the above estimate and (7.1) we get

$$\begin{aligned} S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(v) &\leq \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi - \nabla v\|_{L^2(\Omega_k)^d} + |\varphi_k|_{\widehat{J}(k), \kappa_k} \\ &\leq \|\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k - \nabla v\|_{L^2(\Omega_k)^d} + |\varphi_k|_{\widehat{J}(k), \xi_k} + (1 + C_{\eta}) |\kappa_k - \xi_k|_{\partial \Omega_k^-} \\ &= S_{\widehat{\mathcal{D}}_k, J, \xi_k}(v) + (1 + C_{\eta}) |\kappa_k - \xi_k|_{\partial \Omega_k^-}. \end{aligned}$$

If moreover $v \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\xi_k = \operatorname{argmin}_{\xi \in Z_{\mathcal{D}_k}} S_{\widehat{\mathcal{D}}_k, J, \xi}(v)$, Lemma 7.4 yields $S_{\widehat{\mathcal{D}}_k, J, \xi_k}(v) \leq C_S h_{\widehat{\mathcal{M}}_k} \|v\|_{H^2(\Omega_k)}$. \blacksquare

7.2 The local WDGGD scheme for linear elliptic problems

We introduce here our local WDGGD scheme before embarking into its a priori error analysis.

Set $\vartheta_0 = 0$ and define iteratively $\vartheta_k \in X_{\mathcal{D}_k}$ for $k = 1, \dots, M$ as

$$\vartheta_k = \hat{\vartheta}_k + \kappa_k, \quad (7.4a)$$

where $\kappa_k \in Z_{\mathcal{D}_k}$ is defined as

$$\kappa_k = \vartheta_{k-1} \chi_{\Omega \setminus \Omega_k} \quad (7.4b)$$

and $\hat{\vartheta}_k \in Y_{\mathcal{D}_k}$ is the solution of the local problem

$$\int_{\Omega_k} A \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi d\mathbf{x} = \int_{\Omega_k} (f_0 \Pi_{\widehat{\mathcal{D}}_k} \varphi - \mathbf{F} \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi) d\mathbf{x} \quad (7.4c)$$

for all $\varphi \in Y_{\mathcal{D}_k}$.

Remember that $\nabla_{\widehat{\mathcal{D}}_k} = \nabla_{\widehat{\mathcal{D}}_k, 0}$, hence we use homogeneous boundary conditions for φ . Due to the definition of $\nabla_{\widehat{\mathcal{D}}_k, \kappa_k}$ the inhomogeneous Dirichlet boundary condition κ_k is weakly imposed on $\widehat{\vartheta}_k$. We have $\kappa_1 = 0$, hence $\vartheta_1 = \widehat{\vartheta}_1 \in X_{\mathcal{D}_1}$. Then, for $k \geq 2$ the scheme (7.4) computes a new local solution $\widehat{\vartheta}_k$ on a refined mesh $\widehat{\mathcal{M}}_k$, where the boundary condition is inherited from the previous solution ϑ_{k-1} .

In Section 7.2.1 we perform the a priori error analysis for the local solutions $\widehat{\vartheta}_k$ and provide bounds for the errors in the local domains Ω_k . Section 7.2.2 improves the results of Section 7.2.1 in a particular case, showing that the error due to artificial boundary conditions is of higher order. Finally, Section 7.2.3 provides error bounds for the global solution ϑ_k .

7.2.1 A priori error analysis for the local solution

In this section we proceed with the a priori error analysis of the local scheme presented in Section 7.2. Before proving convergence of the scheme we need the following interpolation result.

Lemma 7.6. *Let $\xi_{k-1} \in Z_{\mathcal{D}_{k-1}}$, $\varphi_{k-1} \in Y_{\mathcal{D}_{k-1}}$ and $\xi_k = (\xi_{k-1} + \varphi_{k-1})\chi_{\Omega \setminus \Omega_k} \in Z_{\mathcal{D}_k}$. Then there exists $\varphi_k \in Y_{\mathcal{D}_k}$ such that*

$$\|\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_{k-1}, \xi_{k-1}} \varphi_{k-1}\|_{L^2(\Omega_k)^d} \leq C_i |\varphi_{k-1}|_{\widehat{\mathcal{J}}(k-1), \xi_{k-1}}, \quad (7.5a)$$

$$|\varphi_k|_{\widehat{\mathcal{J}}(k), \xi_k} \leq C_i |\varphi_{k-1}|_{\widehat{\mathcal{J}}(k-1), \xi_{k-1}}, \quad (7.5b)$$

with $C_i = \sqrt{2}C_\eta(1 + C_{\omega, k}^2 C_r)^{1/2}$, $C_{\omega, k} = \max_{K \in \mathcal{M}_k, \sigma \in \mathcal{F}_K} \omega_{K, \sigma}^{-1}$, C_η from Lemma 7.5 and C_r from Assumption 7.1.

Proof. Since $\widehat{\mathcal{M}}_k$ is a refinement of $\widehat{\mathcal{M}}_{k-1}$ in Ω_k , there exists $\varphi_k \in Y_{\mathcal{D}_k}$ such that $\Pi_{\widehat{\mathcal{D}}_k} \varphi_k = \Pi_{\widehat{\mathcal{D}}_{k-1}} \varphi_{k-1}|_{\Omega_k}$. Hence

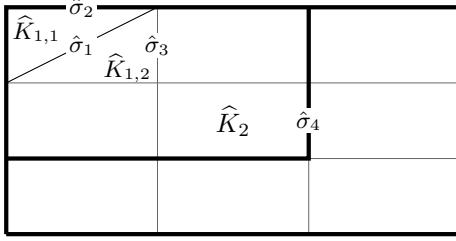
$$\begin{aligned} & \|\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_{k-1}, \xi_{k-1}} \varphi_{k-1}\|_{L^2(\Omega_k)^d}^2 \\ & \leq 2 \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K, \sigma}} \left| \eta(s) \frac{[\![\varphi_k]\!]_{K, \sigma, \xi_k}(\mathbf{y})}{d_{K, \sigma}} \right|^2 d\mathbf{x} \\ & \quad + 2 \sum_{\substack{K \in \mathcal{M}_{k-1} \\ K \subset \Omega_k}} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K, \sigma}} \left| \eta(s) \frac{[\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}(\mathbf{y})}{d_{K, \sigma}} \right|^2 d\mathbf{x}, \end{aligned}$$

since the broken gradients of $\Pi_{\widehat{\mathcal{D}}_{k-1}} \varphi_{k-1}|_{\Omega_k}$ and $\Pi_{\widehat{\mathcal{D}}_k} \varphi_k$ cancel each other out. With the change of variables $d\mathbf{x} = d_{K, \sigma} s^{d-1} ds d\mathbf{y}$ we have

$$\int_{D_{K, \sigma}} \left| \eta(s) \frac{[\![\varphi_k]\!]_{K, \sigma, \xi_k}(\mathbf{y})}{d_{K, \sigma}} \right|^2 d\mathbf{x} = \frac{1}{d_{K, \sigma}} \int_\alpha^1 \eta(s)^2 s^{d-1} ds \int_\sigma [\![\varphi_k]\!]_{K, \sigma, \xi_k}(\mathbf{y})^2 d\mathbf{y}$$

and similarly for φ_{k-1} . Using $C_\eta^2 = \int_\alpha^1 \eta(s)^2 s^{d-1} ds$ yields

$$\begin{aligned} \|\nabla_{\widehat{\mathcal{D}}_k, \xi_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_{k-1}, \xi_{k-1}} \varphi_{k-1}\|_{L^2(\Omega_k)^d}^2 & \leq 2C_\eta^2 \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K, \sigma}} \int_\sigma [\![\varphi_k]\!]_{K, \sigma, \xi_k}(\mathbf{y})^2 d\mathbf{y} \\ & \quad + 2C_\eta^2 \sum_{\substack{K \in \mathcal{M}_{k-1} \\ K \subset \Omega_k}} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K, \sigma}} \int_\sigma [\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}(\mathbf{y})^2 d\mathbf{y} \\ & \leq 2C_\eta^2 (|\varphi_k|_{\widehat{\mathcal{J}}(k), \xi_k}^2 + |\varphi_{k-1}|_{\widehat{\mathcal{J}}(k-1), \xi_{k-1}}^2). \end{aligned}$$



Ω_k and Ω_{k-1} bounded by thick lines —,
 $\hat{\sigma}_1 \subset K_1 = \hat{K}_{1,1} \cup \hat{K}_{1,2}$,
 $\hat{\sigma}_2 = \sigma_2$ with $(\hat{\sigma}_2, \sigma_2) \in \hat{\mathcal{F}}_{k,b} \times \hat{\mathcal{F}}_{k-1,b}$,
 $\hat{\sigma}_3 \in \hat{\mathcal{F}}_{k,i}$,
 $\hat{\sigma}_4 = \sigma_4$ with $(\hat{\sigma}_4, \sigma_4) \in \hat{\mathcal{F}}_{k,b} \times \hat{\mathcal{F}}_{k-1,i}$.

Figure 7.2. Example of a situation described in the proof of Lemma 7.6.

To obtain (7.5a) it remains to prove $|\varphi_k|_{\hat{J}(k), \xi_k}^2 \leq C_{\omega,k}^2 C_r |\varphi_{k-1}|_{\hat{J}(k-1), \xi_{k-1}}^2$. We write $|\varphi_k|_{\hat{J}(k), \xi_k}^2$ as

$$|\varphi_k|_{\hat{J}(k), \xi_k}^2 = \sum_{\substack{K \in \mathcal{M}_{k-1} \\ K \subset \Omega_k}} \sum_{\substack{\hat{K} \in \mathcal{M}_k \\ \hat{K} \subset K}} \sum_{\hat{\sigma} \in \mathcal{F}_{\hat{K}}} \frac{1}{d_{\hat{K}, \hat{\sigma}}} \int_{\hat{\sigma}} [\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k}(\mathbf{y})^2 d\mathbf{y}.$$

Let K , \hat{K} and $\hat{\sigma}$ be as in the above sum, either $\hat{\sigma} \subset K$ and so $[\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k} = 0$ or there exists $\sigma \in \mathcal{F}_K$ such that $\hat{\sigma} \subseteq \sigma$. In that latter case, if $(\hat{\sigma}, \sigma) \in \hat{\mathcal{F}}_{k,b} \times \hat{\mathcal{F}}_{k-1,b}$ or $\hat{\sigma} \in \hat{\mathcal{F}}_{k,i}$ then $[\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k} = [\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}$. If instead $(\hat{\sigma}, \sigma) \in \hat{\mathcal{F}}_{k,b} \times \hat{\mathcal{F}}_{k-1,i}$ then $[\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k} = \omega_{K,\sigma}^{-1} [\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}$. See Figure 7.2 for an illustration of the above cases. Since $\omega_{K,\sigma}^{-1} \geq 1$, we obtain in all cases

$$|[\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k}| \leq \omega_{K,\sigma}^{-1} |[\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}| \leq C_{\omega,k} |[\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}|.$$

Furthermore, by Assumption 7.1 we have $d_{K,\sigma} \leq C_r d_{\hat{K}, \hat{\sigma}}$. These considerations together give

$$\sum_{\substack{\hat{K} \in \mathcal{M}_k \\ \hat{K} \subset K}} \sum_{\hat{\sigma} \in \mathcal{F}_{\hat{K}}} \frac{1}{d_{\hat{K}, \hat{\sigma}}} \int_{\hat{\sigma}} [\![\varphi_k]\!]_{\hat{K}, \hat{\sigma}, \xi_k}(\mathbf{y})^2 d\mathbf{y} \leq C_{\omega,k}^2 C_r \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} [\![\varphi_{k-1}]\!]_{K, \sigma, \xi_{k-1}}(\mathbf{y})^2 d\mathbf{y},$$

hence $|\varphi_k|_{\hat{J}(k), \xi_k}^2 \leq C_{\omega,k}^2 C_r |\varphi_{k-1}|_{\hat{J}(k-1), \xi_{k-1}}^2$ and (7.5a) is proved. In [52, section 6.1] it is shown that $C_\eta \geq 1$, hence $C_{\omega,k} C_r^{1/2} < C_i = \sqrt{2} C_\eta (1 + C_{\omega,k}^2 C_r)^{1/2}$ and (7.5b) follows. ■

The next lemma has been proved for the DGGD in [52] and is valid for the local WDGGD as well.

Lemma 7.7. *Let $\hat{\mathcal{D}}_k$ be a local WDGGD, then there exists $C_{eq} > 0$ depending only on α, ℓ and d such that*

$$|\varphi_k|_{\hat{J}(k), 0} \leq C_{eq} \|\nabla_{\hat{\mathcal{D}}_k} \varphi_k\|_{L^2(\Omega_k)^d} \quad \forall \varphi_k \in Y_{\mathcal{D}_k}.$$

Proof. Follows the lines of [52, Lemma 3.8]. ■

The next lemma shows that the error of the local solution depends as usual on the regularity of the solution and data but also on the error committed on the artificial boundary condition. The proof is inspired from the one of Theorem 6.10 (see also [41, Theorem 2.28]) and uses Lemma 7.5. Note that the next result is valid for any $\kappa_k \in Z_{\mathcal{D}_k}$ and not only κ_k given by scheme (7.4).

Lemma 7.8. Let $u \in H_0^1(\Omega)$ be the exact solution to (6.6), $\kappa_k \in Z_{\mathcal{D}_k}$ and $\hat{\vartheta}_k \in Y_{\mathcal{D}_k}$ be solution of (7.4c). Then

$$\begin{aligned} & \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} \\ & \leq \frac{1+C_{eq}}{\lambda} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + C_A \min_{\xi_k \in Z_{\mathcal{D}_k}} (S_{\widehat{\mathcal{D}}_k, J, \xi_k}(u) + C_\partial |\kappa_k - \xi_k|_{\partial\Omega_k^-}) \end{aligned} \quad (7.6)$$

with $C_A := C_{eq}(1 + \kappa(A))$ and C_∂ from Lemma 7.5.

Proof. Since $\widehat{\mathcal{D}}_k$ is a WDGGD, by Definition 6.6 for any $\mathbf{v} \in H_{\text{div}}(\Omega_k)$ and $\psi_k \in Y_{\mathcal{D}_k}$ we have

$$\left| \int_{\Omega_k} (\nabla_{\widehat{\mathcal{D}}_k} \psi_k \cdot \mathbf{v} + \Pi_{\widehat{\mathcal{D}}_k} \psi_k \nabla \cdot \mathbf{v}) \, d\mathbf{x} \right| \leq \|\nabla_{\widehat{\mathcal{D}}_k} \psi_k\|_{L^2(\Omega_k)^d} W_{\widehat{\mathcal{D}}_k}(\mathbf{v}).$$

As $-\nabla \cdot (A\nabla u + \mathbf{F}) = f_0 \in L^2(\Omega_k)$ we can take $\mathbf{v} = A\nabla u + \mathbf{F}$ and obtain

$$\left| \int_{\Omega_k} (\nabla_{\widehat{\mathcal{D}}_k} \psi_k \cdot (A\nabla u + \mathbf{F}) - \Pi_{\widehat{\mathcal{D}}_k} \psi_k f_0) \, d\mathbf{x} \right| \leq \|\nabla_{\widehat{\mathcal{D}}_k} \psi_k\|_{L^2(\Omega_k)^d} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}).$$

Using (7.4c) we get

$$\left| \int_{\Omega_k} A(\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k) \cdot \nabla_{\widehat{\mathcal{D}}_k} \psi_k \, d\mathbf{x} \right| \leq \|\nabla_{\widehat{\mathcal{D}}_k} \psi_k\|_{L^2(\Omega_k)^d} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}).$$

Let $\varphi_k \in Y_{\mathcal{D}_k}$, we have

$$\begin{aligned} & \int_{\Omega_k} A(\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k) \cdot \nabla_{\widehat{\mathcal{D}}_k} \psi_k \, d\mathbf{x} \\ & \leq \|\nabla_{\widehat{\mathcal{D}}_k} \psi_k\|_{L^2(\Omega_k)^d} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + \int_{\Omega_k} A(\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla u) \cdot \nabla_{\widehat{\mathcal{D}}_k} \psi_k \, d\mathbf{x} \\ & \leq \|\nabla_{\widehat{\mathcal{D}}_k} \psi_k\|_{L^2(\Omega_k)^d} (W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + \bar{\lambda} \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla u\|_{L^2(\Omega_k)^d}). \end{aligned}$$

We choose $\psi_k = \varphi_k - \hat{\vartheta}_k$, since $\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k = \nabla_{\widehat{\mathcal{D}}_k, 0}(\varphi_k - \hat{\vartheta}_k) = \nabla_{\widehat{\mathcal{D}}_k}(\varphi_k - \hat{\vartheta}_k)$ we get

$$\underline{\lambda} \|\nabla_{\widehat{\mathcal{D}}_k}(\varphi_k - \hat{\vartheta}_k)\|_{L^2(\Omega_k)^d}^2 \leq \int_{\Omega_k} A(\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k) \cdot \nabla_{\widehat{\mathcal{D}}_k}(\varphi_k - \hat{\vartheta}_k) \, d\mathbf{x}$$

and hence

$$\|\nabla_{\widehat{\mathcal{D}}_k}(\varphi_k - \hat{\vartheta}_k)\|_{L^2(\Omega_k)^d} \leq \frac{1}{\underline{\lambda}} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + \kappa(A) \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla u\|_{L^2(\Omega_k)^d}. \quad (7.7)$$

This gives

$$\begin{aligned} \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} & \leq \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k\|_{L^2(\Omega_k)^d} + \|\nabla_{\widehat{\mathcal{D}}_k}(\varphi_k - \hat{\vartheta}_k)\|_{L^2(\Omega_k)^d} \\ & \leq \frac{1}{\underline{\lambda}} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + (1 + \kappa(A)) \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k\|_{L^2(\Omega_k)^d}. \end{aligned} \quad (7.8)$$

Using $|\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} \leq |\varphi_k|_{\widehat{\mathcal{J}}(k), \kappa_k} + |\hat{\vartheta}_k - \varphi_k|_{\widehat{\mathcal{J}}(k), 0}$, Lemma 7.7 and (7.7) yields

$$|\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} \leq |\varphi_k|_{\widehat{\mathcal{J}}(k), \kappa_k} + \frac{C_{eq}}{\underline{\lambda}} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + C_{eq}\kappa(A) \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi_k - \nabla u\|_{L^2(\Omega_k)^d}. \quad (7.9)$$

Summing (7.8) and (7.9) and taking the infimum over $\varphi_k \in Y_{\mathcal{D}_k}$ we get

$$\|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} \leq \frac{1+C_{eq}}{\underline{\lambda}} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + C_{eq}(1 + \kappa(A)) S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(u).$$

We conclude using Lemma 7.5 and taking the inf over ξ_k . ■

Lemma 7.9. Let $((\kappa_k, \hat{\vartheta}_k))_{k=1}^M$ be the sequence defined by the local scheme (7.4a) to (7.4c). Then for $k \geq 2$

$$\begin{aligned} \min_{\xi_k \in Z_{\mathcal{D}_k}} (S_{\widehat{\mathcal{D}}_k, J, \xi_k}(u) + C_\partial |\kappa_k - \xi_k|_{\partial \Omega_k^-}) &\leq 2C_i \left(\|\nabla_{\widehat{\mathcal{D}}_k, \kappa_{k-1}} \hat{\vartheta}_{k-1} - \nabla u\|_{L^2(\Omega_k)^d} \right. \\ &\quad \left. + |\hat{\vartheta}_{k-1}|_{\widehat{J}(k-1), \kappa_{k-1}} \right), \end{aligned}$$

where C_i is defined in Lemma 7.6.

Proof. Taking $\xi_k = \kappa_k$ we have

$$\min_{\xi_k \in Z_{\mathcal{D}_k}} (S_{\widehat{\mathcal{D}}_k, J, \xi_k}(u) + C_\partial |\kappa_k - \xi_k|_{\partial \Omega_k^-}) \leq S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(u).$$

Since $\kappa_k = (\kappa_{k-1} + \hat{\vartheta}_{k-1})\chi_{\Omega \setminus \Omega_k}$ by Lemma 7.6 there exists $\hat{\varphi}_k \in Y_{\mathcal{D}_k}$ satisfying

$$\begin{aligned} \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\varphi}_k - \nabla_{\widehat{\mathcal{D}}_{k-1}, \kappa_{k-1}} \hat{\vartheta}_{k-1}\|_{L^2(\Omega_k)^d} &\leq C_i |\hat{\vartheta}_{k-1}|_{\widehat{J}(k-1), \kappa_{k-1}}, \\ |\hat{\varphi}_k|_{\widehat{J}(k), \kappa_k} &\leq C_i |\hat{\vartheta}_{k-1}|_{\widehat{J}(k-1), \kappa_{k-1}} \end{aligned}$$

and so

$$\begin{aligned} S_{\widehat{\mathcal{D}}_k, J, \kappa_k}(u) &= \min_{\varphi \in Y_{\mathcal{D}_k}} (\|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \varphi - \nabla u\|_{L^2(\Omega_k)^d} + |\varphi|_{\widehat{J}(k), \kappa_k}) \\ &\leq \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\varphi}_k - \nabla u\|_{L^2(\Omega_k)^d} + |\hat{\varphi}_k|_{\widehat{J}(k), \kappa_k} \\ &\leq \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_{k-1}} \hat{\vartheta}_{k-1} - \nabla u\|_{L^2(\Omega_k)^d} + 2C_i |\hat{\vartheta}_{k-1}|_{\widehat{J}(k-1), \kappa_{k-1}}. \end{aligned} \quad \blacksquare$$

Let $\mathcal{H} \subset \mathbb{R}_+$ be a countable set with zero as only accumulation point. For each $h \in \mathcal{H}$ we consider a polytopal mesh sequence $(\mathfrak{T}_{h,k})_{k=1}^M = ((\mathcal{M}_{h,k}, \mathcal{F}_{h,k}, \mathcal{P}_{h,k}))_{k=1}^M$ satisfying Assumption 7.1 with $h = \max_{k=1, \dots, M} h_{\mathcal{M}_{h,k}}$, where $h_{\mathcal{M}_{h,k}} = \max\{h_K : K \in \mathcal{M}_{h,k}\}$. Let $\mathcal{D}_{h,k}$ and $\widehat{\mathcal{D}}_{h,k}$ be the global and local WDGGD given by those meshes $\mathfrak{T}_{h,k}$. In the following the index h in $\mathcal{D}_{h,k}$ and $\widehat{\mathcal{D}}_{h,k}$ is left out of notation for the sake of simplicity.

Theorem 7.10. Let \mathcal{D}_k and $\widehat{\mathcal{D}}_k$ be global and local WDGGD. Let $((\kappa_k, \hat{\vartheta}_k))_{k=1}^M$ be the sequence defined by the local scheme (7.4a) to (7.4c) and $u \in H_0^1(\Omega)$ the exact solution to (6.6). Then for $k = 1, \dots, M$

$$\lim_{h \rightarrow 0} \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{J}(k), \kappa_k} = 0. \quad (7.10a)$$

Moreover, if $u \in H_0^1(\Omega) \cap H^2(\Omega)$, the coefficients of A are Lipschitz continuous and $\mathbf{F} \in H^1(\Omega)^d$; then, there exists C_1, C_2, C_3 depending on $\alpha, \ell, d, \rho, C_r, |\Omega|, A, \mathbf{F}$ and u such that

$$\|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{J}(k), \kappa_k} \leq C_1 h, \quad (7.10b)$$

$$\|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{J}(k), \kappa_k} \leq C_2 h_{\widehat{\mathcal{M}}_k} + C_3 |\kappa_k - \xi_k|_{\partial \Omega_k^-}, \quad (7.10c)$$

where $\xi_k = \operatorname{argmin}_{\xi \in Z_{\mathcal{D}_k}} S_{\widehat{\mathcal{D}}_k, J, \xi}(u)$.

The above theorem gives three important results. The first one (7.10a) asserts that the numerical solution given by the local scheme (7.4a) to (7.4c) converges to the exact solution even under weak regularity of the solution and data. Assuming more regularity we recover in (7.10b) the usual convergence rate. In (7.10c) we establish that the error in the local domain depends on the local mesh size and the error due to the artificial boundary condition.

Proof of Theorem 7.10. Let

$$E_{\widehat{\mathcal{D}}_k} := \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{v}_k\|_{L^2(\Omega_k)^d} + |\hat{v}_k|_{\widehat{\mathcal{J}}(k), \kappa_k}.$$

Since $\kappa_1 = 0$ and $Z_{\mathcal{D}_1} = \{0\}$ by Lemma 7.8 we have

$$E_{\widehat{\mathcal{D}}_1} \leq \frac{1 + C_{\text{eq}}}{\lambda} W_{\widehat{\mathcal{D}}_1}(A\nabla u + \mathbf{F}) + C_A S_{\widehat{\mathcal{D}}_1, J, 0}(u)$$

and by Lemmas 7.8 and 7.9 we have, for $k \geq 2$,

$$E_{\widehat{\mathcal{D}}_k} \leq \frac{1 + C_{\text{eq}}}{\lambda} W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) + 2C_i C_A E_{\widehat{\mathcal{D}}_{k-1}}.$$

Let $C_{Ai} = 2C_i C_A$, since $S_{\widehat{\mathcal{D}}_1, J, 0}(u) \leq S_{\mathcal{D}_1, J}(u)$ it holds

$$\begin{aligned} E_{\widehat{\mathcal{D}}_k} &\leq C_{Ai}^{k-1} E_{\widehat{\mathcal{D}}_1} + \frac{1 + C_{\text{eq}}}{\lambda} \sum_{j=2}^k C_{Ai}^{k-j} W_{\widehat{\mathcal{D}}_j}(A\nabla u + \mathbf{F}) \\ &\leq C_A C_{Ai}^{k-1} S_{\mathcal{D}_1, J}(u) + \frac{1 + C_{\text{eq}}}{\lambda} \sum_{j=1}^k C_{Ai}^{k-j} W_{\widehat{\mathcal{D}}_j}(A\nabla u + \mathbf{F}). \end{aligned} \quad (7.11)$$

We have thus proved (7.10a) thanks to (7.11), Lemma 6.17 and the limit-conformity of $\widehat{\mathcal{D}}_j$ for $j = 1, \dots, k$ (we recall that $\widehat{\mathcal{D}}_j$ is a WDGGD and hence a sequence of $\widehat{\mathcal{D}}_j$ is limit-conforming). Under the additional assumptions on the data, from Lemmas 6.14 and 6.15 for $\widehat{\mathcal{D}}_k$ we have

$$\begin{aligned} S_{\mathcal{D}_1, J}(u) &\leq C_S h_{\mathcal{M}_1} \|u\|_{H^2(\Omega)}, \\ W_{\widehat{\mathcal{D}}_k}(A\nabla u + \mathbf{F}) &\leq C_W h_{\widehat{\mathcal{M}}_k} \|A\nabla u + \mathbf{F}\|_{H^1(\Omega_k)^d} \end{aligned} \quad (7.12)$$

and so

$$E_{\widehat{\mathcal{D}}_k} \leq C_A C_{Ai}^{k-1} C_S h_{\mathcal{M}_1} \|u\|_{H^2(\Omega)} + \frac{1 + C_{\text{eq}}}{\lambda} \sum_{j=1}^k C_{Ai}^{k-1} C_W h_{\widehat{\mathcal{M}}_j} \|A\nabla u + \mathbf{F}\|_{H^1(\Omega_j)^d},$$

which implies (7.10b) with

$$C_1 := C_A C_{Ai}^{k-1} C_S \|u\|_{H^2(\Omega)} + \frac{1 + C_{\text{eq}}}{\lambda} \sum_{j=1}^k C_{Ai}^{k-1} C_W \|A\nabla u + \mathbf{F}\|_{H^1(\Omega_j)^d}.$$

Let $\xi_k = \operatorname{argmin}_{\xi \in Z_{\mathcal{D}_k}} S_{\widehat{\mathcal{D}}_k, J, \xi}(u)$, it holds

$$\min_{\xi \in Z_{\mathcal{D}_k}} (S_{\widehat{\mathcal{D}}_k, J, \xi}(u) + C_\partial |\kappa_k - \xi|_{\partial \Omega_k^-}) \leq S_{\widehat{\mathcal{D}}_k, J, \xi_k}(u) + C_\partial |\kappa_k - \xi_k|_{\partial \Omega_k^-},$$

using Lemma 7.4 we get $S_{\widehat{\mathcal{D}}_k, J, \xi_k}(u) \leq C_S h_{\widehat{\mathcal{M}}_k} \|u\|_{H^2(\Omega_k)}$ and again from Lemma 7.8 and (7.12) we obtain the bound (7.10c) with

$$C_2 := C_A C_S \|u\|_{H^2(\Omega_k)} + \frac{C_W}{\lambda} \|A\nabla u + \mathbf{F}\|_{H^1(\Omega_k)^d}, \quad (7.13)$$

where $C_3 = C_A C_\partial$. ■

7.2.2 Improved local estimate

Under stronger conditions and using the pointwise error estimates proved in [36] we can further improve the local estimate (7.10c) for $k = 2$.

Let $\mathbf{z} \in \Omega$, the weight function $\sigma_{\mathbf{z},h}(\mathbf{x}) = h/(h + |\mathbf{x} - \mathbf{z}|)$ and $\|\cdot\|_{W_{\mathbf{z},h}^{2,\infty}(\Omega)}$ a weighted Sobolev norm defined by

$$\|v\|_{W_{\mathbf{z},h}^{2,\infty}(\Omega)} = \max_{i=0,1,2} |v|_{W_{\mathbf{z},h}^{i,\infty}}, \quad |v|_{W_{\mathbf{z},h}^{i,\infty}} = \max_{|\alpha|=i} \|\sigma_{\mathbf{z},h} \frac{\partial^\alpha v}{\partial \mathbf{x}^\alpha}\|_{L^\infty(\Omega)}.$$

We will use the following lemma, which is a version of [36, Corollary 5.5].

Lemma 7.11. *Let $A = aI_d$ with $I_d \in \mathbb{R}^{d \times d}$ the identity matrix and $a > 0$. Let $u \in W_0^{1,\infty}(\Omega) \cap W^{2,\infty}(\Omega)$ be solution of (6.2) with $f \in L^2(\Omega)$, $\vartheta_1 \in X_{\mathcal{D}_1}$ solution of (7.4c). Then there is a constant $C_\infty >$ such that for any $\mathbf{z} \in \bar{\Omega}$*

$$|u(\mathbf{z}) - \Pi_{\mathcal{D}_1} \vartheta_1(\mathbf{z})| \leq C_\infty h^2 \log(h^{-1}) \|u\|_{W_{\mathbf{z},h}^{2,\infty}(\Omega)}.$$

Applying Lemma 7.11 to (7.10c) we obtain a better bound on the local error for $k = 2$, as explained in the following theorem.

Theorem 7.12. *Let $u \in W_0^{1,\infty}(\Omega) \cap W^{2,\infty}(\Omega)$ be solution of (6.2) with $A = aI_d$ and $f \in L^2(\Omega)$ as in Lemma 7.11. Let \mathcal{D}_k and $\widehat{\mathcal{D}}_k$ be global and local WDGGD and $((\kappa_k, \widehat{\vartheta}_k))_{k=1}^2$ the sequence defined by the local scheme (7.4a) to (7.4c). Under the assumption that $h \leq C \min_{K \in \mathcal{M}_1} h_K$ with $C > 0$ independent of h , there exists C_4 independent of u and h such that*

$$\begin{aligned} & \|\nabla u - \nabla_{\widehat{\mathcal{D}}_2, \kappa_2} \widehat{\vartheta}_2\|_{L^2(\Omega_2)^d} + |\widehat{\vartheta}_2|_{\widehat{\mathcal{J}}(2), \kappa_2} \\ & \leq C_2 h_{\widehat{\mathcal{M}}_2} + C_4 \left(h_{\widehat{\mathcal{M}}_2} |u|_{H^2(D_2)} + h^{3/2} \log(h^{-1}) \sup_{\mathbf{y} \in \partial\Omega_2 \setminus \partial\Omega} \|u\|_{W_{\mathbf{y},h}^{2,\infty}(\Omega)} \right), \end{aligned} \quad (7.14)$$

where D_2 is a neighborhood of Ω_2 specified below.

Note that (7.14) bounds the error in the local domain Ω_2 and has three terms on the right. From (7.13) we see that the first term depends on u and \mathbf{F} in Ω_2 . The second term depends on u in a small neighborhood of Ω_2 . The last term depends on the regularity of u in the whole domain, but it is of higher order and is measured in a weighted norm whose weight is $\mathcal{O}(1)$ close to the artificial boundary and $\mathcal{O}(h)$ far from it. Hence the error in Ω_2 depends mostly on the regularity of u and \mathbf{F} inside or very close to Ω_2 .

Proof. First we observe that (7.10c) for $k = 2$ is valid with $\xi_2 \in Z_{\mathcal{D}_2}$ such that $\Pi_{\mathcal{D}_2} \xi_2$ is the orthogonal projection of u onto $\Pi_{\mathcal{D}_2} Z_{\mathcal{D}_2}$, indeed even for this choice of ξ_2 we still have $S_{\widehat{\mathcal{D}}_2, J, \xi_2}(u) \leq C_S h_{\widehat{\mathcal{M}}_2} \|u\|_{H^2(\Omega_2)}$. Let $K \in \widehat{\mathcal{M}}_2$, $T \in \mathcal{M}_2 \setminus \widehat{\mathcal{M}}_2$ and $\sigma \in \mathcal{F}_K \cap \mathcal{F}_T$. From Assumption 7.1b) we have $K, T \in \mathcal{M}_1$ and Assumption 7.1d) implies $h_K \leq \rho h_T$. There exists C_Π ([39, Lemma 1.59]) independent of u , T and h_K such that

$$\int_\sigma |\Pi_T \xi_2 - u|(\mathbf{y})^2 d\mathbf{y} \leq C_\Pi h_K^3 |u|_{H^2(T)}^2.$$

Using Assumption 7.1d) we obtain $1/d_{K,\sigma} \leq \rho/h_K$, hence

$$\sum_{\substack{K \in \widehat{\mathcal{M}}_2 \\ T \in \mathcal{M}_2 \setminus \widehat{\mathcal{M}}_2}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K,\sigma}} \int_\sigma |\Pi_T \xi_2 - u|(\mathbf{y})^2 d\mathbf{y} \leq C_\Pi \rho h_{\widehat{\mathcal{M}}_2}^2 |u|_{H^2(D_2)}^2,$$

with $D_2 = \cup_{\{T \in \mathcal{M}_2 \setminus \widehat{\mathcal{M}}_2 : \partial T \cap \partial K \neq \emptyset, K \in \widehat{\mathcal{M}}_2\}} T$. From Lemma 7.11 we have

$$\int_{\sigma} |u - \Pi_{\bar{T}} \kappa_2|(\mathbf{y})^2 d\mathbf{y} = \int_{\sigma} |u - \Pi_{\bar{T}} \vartheta_1|(\mathbf{y})^2 d\mathbf{y} \leq |\sigma| C_{\infty}^2 h^4 \log(h^{-1})^2 \sup_{\mathbf{y} \in \sigma} \|u\|_{W_{\mathbf{y}, h}^{2, \infty}(\Omega)}^2.$$

Since $h \leq C \min_{K \in \mathcal{M}_1} h_K$ it follows that $1/d_{K, \sigma} \leq C\rho/h$ and thus

$$\begin{aligned} & \sum_{\substack{K \in \widehat{\mathcal{M}}_2 \\ T \in \mathcal{M}_2 \setminus \widehat{\mathcal{M}}_2}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{K, \sigma}} \int_{\sigma} |u - \Pi_{\bar{T}} \kappa_2|(\mathbf{y})^2 d\mathbf{y} \\ & \leq |\partial\Omega_2 \setminus \partial\Omega| C_{\infty}^2 C\rho h^3 \log(h^{-1})^2 \sup_{\mathbf{y} \in \partial\Omega_2 \setminus \partial\Omega} \|u\|_{W_{\mathbf{y}, h}^{2, \infty}(\Omega)}^2. \end{aligned}$$

Applying a triangle inequality on $|\kappa_2 - \xi_2|_{\partial\Omega_2^-}$ in (7.10c) we get (7.14). \blacksquare

7.2.3 A priori error analysis for the global solution

We next study the error on the whole domain Ω of the numerical solution $\vartheta_k \in X_{\mathcal{D}_k}$ defined by the local scheme (7.4). The next Lemma 7.13 is the main ingredient for the global error bound.

Lemma 7.13. *Let $u \in H_0^1(\Omega)$ be solution to (6.6) and $(\vartheta_k)_{k=1}^M$ be the sequence defined by the local scheme (7.4a) to (7.4c). Then we have*

$$\begin{aligned} \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} & \leq C_5 (\|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega)^d} + |\vartheta_{k-1}|_{J(k-1)}) \\ & + C_5 (\|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{J}(k), \kappa_k}). \end{aligned}$$

where $C_5 = \sqrt{2}(1 + C_{\eta})(1 + \sqrt{2}C_{\omega, k})$.

Proof. We have

$$\begin{aligned} & \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d}^2 \\ & = \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K, \sigma}} |\nabla u(\mathbf{x}) - \nabla_{\bar{K}} \vartheta_k(\mathbf{x}) - \eta(s) \frac{[\![\vartheta_k]\!]_{K, \sigma}(\mathbf{y})}{d_{K, \sigma}} \mathbf{n}_{K, \sigma}|^2 d\mathbf{x} \\ & = \sum_{T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_T} \int_{D_{T, \sigma}} |\nabla u(\mathbf{x}) - \nabla_{\bar{T}} \vartheta_k(\mathbf{x}) - \eta(s) \frac{[\![\vartheta_k]\!]_{T, \sigma}(\mathbf{y})}{d_{T, \sigma}} \mathbf{n}_{T, \sigma}|^2 d\mathbf{x} \\ & \quad + \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K, \sigma}} |\nabla u(\mathbf{x}) - \nabla_{\bar{K}} \vartheta_k(\mathbf{x}) - \eta(s) \frac{[\![\vartheta_k]\!]_{K, \sigma}(\mathbf{y})}{d_{K, \sigma}} \mathbf{n}_{K, \sigma}|^2 d\mathbf{x} \\ & = I_1 + I_2. \end{aligned}$$

For the first term I_1 , we have the following considerations. Let $T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k$, then $T \in \mathcal{M}_{k-1}$ and $\nabla_{\bar{T}} \vartheta_k = \nabla_{\bar{T}} \vartheta_{k-1}$. Let $\sigma \in \mathcal{F}_T$, if $\sigma \notin \widehat{\mathcal{F}}_{k, b}$ then $[\![\vartheta_k]\!]_{T, \sigma} = [\![\vartheta_{k-1}]\!]_{T, \sigma}$. If $\sigma \in \widehat{\mathcal{F}}_{k, b}$ then $\sigma = \partial K \cap \partial T$ with $K \in \widehat{\mathcal{M}}_k$ and by Assumption 7.1b) $K \in \widehat{\mathcal{M}}_{k-1}$. Using (7.4a) and (7.4b) we have

$$\begin{aligned} [\![\vartheta_k]\!]_{T, \sigma} - [\![\vartheta_{k-1}]\!]_{T, \sigma} & = \omega_{T, \sigma} (\Pi_{\bar{K}} \hat{\vartheta}_k - \Pi_{\bar{T}} \vartheta_{k-1}) - \omega_{T, \sigma} (\Pi_{\bar{K}} \vartheta_{k-1} - \Pi_{\bar{T}} \vartheta_{k-1}) \\ & = \omega_{T, \sigma} (\Pi_{\bar{K}} \hat{\vartheta}_k - \Pi_{\bar{K}} \vartheta_{k-1}). \end{aligned}$$

Next, adding and removing $\llbracket \vartheta_{k-1} \rrbracket_{T,\sigma}$ from $\llbracket \vartheta_k \rrbracket_{T,\sigma}$ we get

$$\begin{aligned} I_1 &\leq 2 \sum_{T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_T} \int_{D_{T,\sigma}} |\nabla u(\mathbf{x}) - \nabla_{\bar{T}} \vartheta_{k-1}(\mathbf{x}) - \eta(s) \frac{\llbracket \vartheta_{k-1} \rrbracket_{T,\sigma}(\mathbf{y})}{d_{T,\sigma}} \mathbf{n}_{T,\sigma}|^2 d\mathbf{x} \\ &\quad + 2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_T \cap \mathcal{F}_K} \int_{D_{T,\sigma}} |\eta(s) \omega_{T,\sigma} \frac{(\Pi_K \hat{\vartheta}_k - \Pi_{\bar{K}} \vartheta_{k-1})(\mathbf{y})}{d_{T,\sigma}}|^2 d\mathbf{x}. \end{aligned}$$

Since $\omega_{T,\sigma} \leq 1$, using a change of variables we have

$$\begin{aligned} I_1 &\leq 2 \|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega \setminus \Omega_k)}^2 \\ &\quad + 2C_\eta^2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_T \cap \mathcal{F}_K} \frac{1}{d_{T,\sigma}} \int_{\sigma} (\Pi_K \hat{\vartheta}_k - \Pi_{\bar{K}} \vartheta_{k-1})(\mathbf{y})^2 d\mathbf{y} \\ &= 2 \|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega \setminus \Omega_k)}^2 + 2C_\eta^2 |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial \Omega_k^+}^2, \end{aligned}$$

where for $\phi_k \in X_{\mathcal{D}_k}$

$$|\phi_k|_{\partial \Omega_k^+}^2 := \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{T,\sigma}} \int_{\sigma} \Pi_K \phi_k(\mathbf{y})^2 d\mathbf{y}.$$

For the second term I_2 we have $\llbracket \vartheta_k \rrbracket_{K,\sigma} = \llbracket \hat{\vartheta}_k \rrbracket_{K,\sigma,\kappa_k}$ if $\sigma \in \widehat{\mathcal{F}}_{k,i}$ or $\sigma \in \widehat{\mathcal{F}}_{k,b} \cap \mathcal{F}_{k,b}$ and $\llbracket \vartheta_k \rrbracket_{K,\sigma} = \omega_{K,\sigma} \llbracket \hat{\vartheta}_k \rrbracket_{K,\sigma,\kappa_k}$ if $\sigma \in \widehat{\mathcal{F}}_{k,b} \setminus \mathcal{F}_{k,b}$. Hence

$$\begin{aligned} I_2 &\leq 2 \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d}^2 \\ &\quad + 2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_T \cap \mathcal{F}_K} \int_{D_{K,\sigma}} |\eta(s) \frac{(1 - \omega_{K,\sigma}) \llbracket \hat{\vartheta}_k \rrbracket_{K,\sigma,\kappa_k}(\mathbf{y})}{d_{K,\sigma}}|^2 d\mathbf{x} \\ &\leq 2 \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d}^2 + 2C_\eta^2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_T \cap \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} \llbracket \hat{\vartheta}_k \rrbracket_{K,\sigma,\kappa_k}(\mathbf{y})^2 d\mathbf{y} \\ &\leq 2 \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d}^2 + 2C_\eta^2 |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k}^2. \end{aligned}$$

We then obtain

$$\begin{aligned} \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d}^2 &\leq 2 \|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega)^d}^2 + 2 \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d}^2 \\ &\quad + 2C_\eta^2 |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial \Omega_k^+}^2 + 2C_\eta^2 |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k}^2. \end{aligned} \tag{7.15}$$

Using similar arguments, we have

$$\begin{aligned}
 |\vartheta_k|_{J(k)}^2 &= \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} [\![\vartheta_k]\!]_{K,\sigma}(\mathbf{y})^2 d\mathbf{y} \\
 &\leq \sum_{K \in \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_K} \frac{1}{d_{K,\sigma}} \int_{\sigma} [\![\hat{\vartheta}_k]\!]_{K,\sigma,\kappa_k}(\mathbf{y})^2 d\mathbf{y} \\
 &\quad + 2 \sum_{T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k} \sum_{\sigma \in \mathcal{F}_T} \frac{1}{d_{T,\sigma}} \int_{\sigma} [\![\vartheta_{k-1}]\!]_{T,\sigma}(\mathbf{y})^2 d\mathbf{y} \\
 &\quad + 2 \sum_{\substack{K \in \widehat{\mathcal{M}}_k \\ T \in \mathcal{M}_k \setminus \widehat{\mathcal{M}}_k}} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_T} \frac{1}{d_{T,\sigma}} \int_{\sigma} \omega_{T,\sigma}^2 (\Pi_{\overline{K}} \hat{\vartheta}_k - \Pi_{\overline{K}} \vartheta_{k-1})(\mathbf{y})^2 d\mathbf{y} \\
 &\leq |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k}^2 + 2 |\vartheta_{k-1}|_{J(k)}^2 + 2 |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial\Omega_k^+}^2.
 \end{aligned} \tag{7.16}$$

Combining (7.15) and (7.16) we get

$$\begin{aligned}
 &\|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} \\
 &\leq \sqrt{2} (\|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega)^d} + \|\nabla u - \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + C_\eta |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial\Omega_k^+} \\
 &\quad + C_\eta |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k}) + \sqrt{2} (|\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k} + |\vartheta_{k-1}|_{J(k)} + |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial\Omega_k^+}) \\
 &\leq \sqrt{2} (\|\nabla u - \nabla_{\mathcal{D}_{k-1}} \vartheta_{k-1}\|_{L^2(\Omega)^d} + |\vartheta_{k-1}|_{J(k)} + \|\nabla u - \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} \\
 &\quad + (1 + C_\eta) |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k}) + \sqrt{2} (1 + C_\eta) |\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial\Omega_k^+}.
 \end{aligned}$$

Since we easily compute

$$|\hat{\vartheta}_k - \vartheta_{k-1}|_{\partial\Omega_k^+} \leq \sqrt{2} C_{\omega,k} (|\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k} + |\vartheta_{k-1}|_{J(k)}),$$

we obtain

$$\begin{aligned}
 &\|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} \\
 &\leq \sqrt{2} (1 + C_\eta) (1 + \sqrt{2} C_{\omega,k}) (\|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_{k-1}\|_{L^2(\Omega)^d} + |\vartheta_{k-1}|_{J(k)}) \\
 &\quad + \sqrt{2} (1 + C_\eta) (1 + \sqrt{2} C_{\omega,k}) (\|\nabla u - \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k}),
 \end{aligned}$$

and the lemma is proved. \blacksquare

Theorem 7.14. Let $u \in H_0^1(\Omega)$ be solution of (6.6) and $(\vartheta_k)_{k=1}^M$ be the sequence defined by the local scheme (7.4a) to (7.4c). Then for $k = 1, \dots, M$

$$\lim_{h \rightarrow 0} \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} = 0. \tag{7.17}$$

Moreover, if $u \in H_0^1(\Omega) \cap H^2(\Omega)$, the coefficients of A are Lipschitz continuous and $\mathbf{F} \in H^1(\Omega)^d$; then, there exists C_6 depending on $\alpha, \ell, d, \rho, C_r, |\Omega|, A, \mathbf{F}$ and u such that

$$\|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} \leq C_6 h. \tag{7.18}$$

Proof. Follows from a recursive argument, Theorem 7.10 and Lemma 7.13. \blacksquare

7.3 The local WDGGD scheme for quasilinear elliptic problems

In this section we analyze the local WDGGD for a class of quasilinear problems satisfying Assumption 6.2. For the sake of simplicity we consider $f \in L^2(\Omega)$, but the algorithm and the results are easily generalized to $f \in H^{-1}(\Omega)$. Under Assumption 6.2 there exists a unique weak solution $u \in H_0^1(\Omega)$ of

$$\int_{\Omega} A(u) \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega). \quad (7.19)$$

The local scheme for problem (7.19) is given as follows. Set $\vartheta_1 \in X_{\mathcal{D}_1}$ a solution of

$$\int_{\Omega} A(\Pi_{\mathcal{D}_1} \vartheta_1) \nabla_{\mathcal{D}_1} \vartheta_1 \cdot \nabla_{\mathcal{D}_1} \phi_1 \, dx = \int_{\Omega} f \Pi_{\mathcal{D}_1} \phi_1 \, dx \quad (7.20a)$$

for all $\phi_1 \in X_{\mathcal{D}_1}$. For $k \geq 2$ we set

$$\vartheta_k = \kappa_k + \hat{\vartheta}_k, \quad (7.20b)$$

where $\kappa_k \in Z_{\mathcal{D}_k}$ is given by

$$\kappa_k = \vartheta_{k-1} \chi_{\Omega \setminus \Omega_k} \quad (7.20c)$$

and $\hat{\vartheta}_k \in Y_{\mathcal{D}_k}$ is solution of

$$\int_{\Omega_k} A(\Pi_{\widehat{\mathcal{D}}_{k-1}} \hat{\vartheta}_{k-1}) \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, dx = \int_{\Omega_k} f \Pi_{\widehat{\mathcal{D}}_k} \varphi_k \, dx \quad (7.20d)$$

for all $\varphi_k \in Y_{\mathcal{D}_k}$.

7.3.1 A priori error analysis for quasilinear problems

We define again a subset $\mathcal{H} \subset \mathbb{R}_+$ with zero as only accumulation point and for each $h \in \mathcal{H}$ a sequence of meshes $(\mathfrak{T}_{h,k})_{k=1}^M$ satisfying Assumption 7.1 with $h = \max_{k=1,\dots,M} h_{\mathcal{M}_k}$. From $(\mathfrak{T}_{h,k})_{k=1}^M$ we define $(\widehat{\mathfrak{T}}_{h,k})_{k=1}^M$ as explained after Assumption 7.1. We consider the weighted gradient discretizations $\mathcal{D}_{h,k}$, $\widehat{\mathcal{D}}_{h,k}$ deriving from $\mathfrak{T}_{h,k}$ and $\widehat{\mathfrak{T}}_{h,k}$, as defined in Section 7.1. We reformulate Theorem 6.12 in Theorem 7.15 here below (see also [41, Theorem 2.35]) using the current notation. It will be used to prove convergence of the local scheme (7.20) in Theorem 7.16.

Theorem 7.15. *For any $h \in \mathcal{H}$ there exists exactly one $\vartheta_{h,1} \in \mathcal{D}_{h,1}$ solution to (7.20a). Moreover, $\Pi_{\mathcal{D}_{h,1}} \vartheta_{h,1}$ converges strongly in $L^2(\Omega)$ to a solution u of (7.19) and $\nabla_{\mathcal{D}_{h,1}} \vartheta_{h,1}$ converges strongly in $L^2(\Omega)^d$ to ∇u as $h \rightarrow 0$.*

We will prove that the same result holds for $\vartheta_{h,k}$ with $k \geq 2$. We start by proving convergence of the local solutions $\hat{\vartheta}_{h,k}$. For simplicity we drop the index h in what follows.

Theorem 7.16. *Let Assumption 6.2 hold, $((\kappa_k, \hat{\vartheta}_k))_{k=1}^M$ be the sequence given by the local scheme (7.20a) to (7.20d) and $u \in H_0^1(\Omega)$ be solution of (7.19). Then for $k = 1, \dots, M$*

$$\lim_{h \rightarrow 0} \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} = 0, \quad (7.21a)$$

$$\lim_{h \rightarrow 0} |\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} = 0, \quad (7.21b)$$

where the limit is taken for $h \in \mathcal{H}$.

Proof. We will prove (7.21) by recursion. For $k = 1$ we easily get (7.21a), indeed $\kappa_1 = 0$, $\hat{\vartheta}_1 = \vartheta_1$ and by Theorem 7.15 we get

$$\lim_{h \rightarrow 0} \|\nabla u - \nabla_{\widehat{\mathcal{D}}_1, \kappa_1} \hat{\vartheta}_1\|_{L^2(\Omega_k)^d} = \lim_{h \rightarrow 0} \|\nabla u - \nabla_{\mathcal{D}_1} \vartheta_1\|_{L^2(\Omega)^d} = 0. \quad (7.22)$$

Let $\phi_1 \in X_{\mathcal{D}_1}$, we have

$$|\hat{\vartheta}_1|_{\widehat{\mathcal{J}}(1), \kappa_1} = |\vartheta_1|_{J(1)} \leq |\vartheta_1 - \phi_1|_{J(1)} + |\phi_1|_{J(1)}.$$

From [52] we infer the existence of a constant C_{eq} depending only on α, ℓ, d such that

$$|\vartheta_1 - \phi_1|_{J(1)} \leq C_{\text{eq}} \|\nabla_{\mathcal{D}_1} \vartheta_1 - \nabla_{\mathcal{D}_1} \phi_1\|_{L^2(\Omega)^d},$$

hence

$$|\hat{\vartheta}_1|_{\widehat{\mathcal{J}}(1), \kappa_1} \leq C_{\text{eq}} \|\nabla_{\mathcal{D}_1} \vartheta_1 - \nabla_{\mathcal{D}_1} \phi_1\|_{L^2(\Omega)^d} + |\phi_1|_{J(1)}.$$

Taking $\phi_1 = \operatorname{argmin}_{\phi \in X_{\mathcal{D}_1}} (\|\Pi_{\mathcal{D}_1} \phi - u\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}_1} \phi - \nabla u\|_{L^2(\Omega)^d} + |\phi|_{J(1)})$ we get (7.21b) for $k = 1$ using the triangle inequality, (7.22) and Lemma 6.17.

Let $k \geq 2$ and suppose that (7.21) holds for $k - 1$. By Lemma 7.6 there exists $\bar{\vartheta}_k \in Y_{\mathcal{D}_k}$ satisfying

$$\begin{aligned} \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k - \nabla_{\widehat{\mathcal{D}}_{k-1}, \kappa_{k-1}} \hat{\vartheta}_{k-1}\|_{L^2(\Omega_k)^d} &\leq C_i |\hat{\vartheta}_{k-1}|_{\widehat{\mathcal{J}}(k-1), \kappa_{k-1}}, \\ |\bar{\vartheta}_k|_{\widehat{\mathcal{J}}(k), \kappa_k} &\leq C_i |\hat{\vartheta}_{k-1}|_{\widehat{\mathcal{J}}(k-1), \kappa_{k-1}}. \end{aligned} \quad (7.23)$$

Let $\tilde{\vartheta}_k \in Y_{\mathcal{D}_k}$ be solution of

$$\int_{\Omega_k} A_{k-1} (\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k + \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k) \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x} = \int_{\Omega_k} f \Pi_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x}$$

for all $\varphi_k \in Y_{\mathcal{D}_k}$, where $A_{k-1} = A(\Pi_{\widehat{\mathcal{D}}_{k-1}} \hat{\vartheta}_{k-1})$. Since $\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} (\bar{\vartheta}_k + \tilde{\vartheta}_k) = \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k + \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k$ it follows that $\hat{\vartheta}_k = \bar{\vartheta}_k + \tilde{\vartheta}_k$. From (7.21) for $k - 1$ and (7.23) it follows that $\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k \rightarrow \nabla u$ strongly in $L^2(\Omega_k)^d$. Thus if $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \rightarrow 0$ strongly in $L^2(\Omega_k)^d$ then $\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k \rightarrow \nabla u$ strongly in $L^2(\Omega_k)^d$ and hence (7.21a) holds for k . From the coercivity of A

$$\begin{aligned} \lambda \|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d}^2 &\leq \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} \\ &= \int_{\Omega_k} f \Pi_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} - \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} \\ &\leq \|f\|_{L^2(\Omega_k)} \|\Pi_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)} + \bar{\lambda} \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k\|_{L^2(\Omega_k)^d} \|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d} \\ &\leq (C_p \|f\|_{L^2(\Omega_k)} + \bar{\lambda} \|\nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \bar{\vartheta}_k\|_{L^2(\Omega_k)^d}) \|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d} \end{aligned}$$

and hence $\|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d}$ is bounded. It follows from the compactness of $\widehat{\mathcal{D}}_k$ and [41, Lemma 2.15] that there exists $w \in H_0^1(\Omega_k)$ and a subsequence \mathcal{H}' of \mathcal{H} such that $\Pi_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \rightarrow w$ strongly in $L^2(\Omega_k)$ and $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \rightharpoonup \nabla w$ weakly in $L^2(\Omega_k)^d$ as $h \rightarrow 0$ with $h \in \mathcal{H}'$. We will show that $w = 0$, that the convergence holds for the whole sequence \mathcal{H} and that $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k$ converges strongly. Let $v \in H_0^1(\Omega_k)$ and

$$\varphi_k = \operatorname{argmin}_{\varphi \in Y_{\mathcal{D}_k}} (\|\Pi_{\widehat{\mathcal{D}}_k} \varphi - v\|_{L^2(\Omega_k)} + \|\nabla_{\widehat{\mathcal{D}}_k} \varphi - \nabla v\|_{L^2(\Omega_k)^d} + |\varphi|_{\widehat{\mathcal{J}}(k), 0}).$$

Since $\widehat{\mathcal{D}}_k$ is a WDGGD, from Lemma 6.17 we have that $\Pi_{\widehat{\mathcal{D}}_k} \varphi_k \rightarrow v$ strongly in $L^2(\Omega_k)$ and $\nabla_{\widehat{\mathcal{D}}_k} \varphi_k \rightarrow \nabla v$ strongly in $L^2(\Omega_k)^d$. From (7.21a) $\nabla_{\widehat{\mathcal{D}}_{k-1}, \kappa_{k-1}} \hat{\vartheta}_{k-1} \rightarrow \nabla u$ strongly in $L^2(\Omega_k)^d$,

furthermore by coercivity and consistency we can show that $\Pi_{\widehat{\mathcal{D}}_{k-1}} \hat{\vartheta}_{k-1} \rightarrow u$ strongly in $L^2(\Omega_k)$ as well. The same holds for $\bar{\vartheta}_k$. Hence by the nonlinear strong convergence lemma [41, section D.4] we obtain

$$\begin{aligned} A(\Pi_{\widehat{\mathcal{D}}_{k-1}} \hat{\vartheta}_{k-1}) \nabla_{\widehat{\mathcal{D}}_k} \varphi_k &\rightarrow A(u) \nabla v \text{ strongly in } L^2(\Omega_k)^d, \\ A(\Pi_{\widehat{\mathcal{D}}_{k-1}} \hat{\vartheta}_{k-1}) \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \bar{\vartheta}_k &\rightarrow A(u) \nabla u \text{ strongly in } L^2(\Omega_k)^d. \end{aligned}$$

It follows from the weak-strong convergence lemma [41, section D.4] and symmetry of A that

$$\int_{\Omega_k} A(u) \nabla w \cdot \nabla v \, d\mathbf{x} = \int_{\Omega_k} \nabla w \cdot A(u) \nabla v \, d\mathbf{x} = \lim_{h \rightarrow 0} \int_{\Omega_k} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x}, \quad (7.24)$$

where the limit is for $h \in \mathcal{H}'$. On the other hand we have

$$\int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x} = \int_{\Omega_k} f \Pi_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x} - \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \bar{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x}$$

and taking the limit we get

$$\lim_{h \rightarrow 0} \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi_k \, d\mathbf{x} = \int_{\Omega_k} f v \, d\mathbf{x} - \int_{\Omega_k} A(u) \nabla u \cdot \nabla v \, d\mathbf{x} = 0. \quad (7.25)$$

Putting together (7.24) and (7.25) and using the symmetry of A_{k-1} we obtain

$$\int_{\Omega_k} A(u) \nabla w \cdot \nabla v \, d\mathbf{x} = 0$$

for all $v \in H_0^1(\Omega_k)$ and so $w = 0$. We can repeat the same reasoning for each subsequence of $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k$ and obtain the same limit $w = 0$, hence $\Pi_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \rightarrow 0$ strongly in $L^2(\Omega_k)$ and $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \rightarrow 0$ weakly in $L^2(\Omega_k)^d$ for the whole sequence \mathcal{H} . Furthermore

$$\int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} = \int_{\Omega_k} f \Pi_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} - \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_{k,\kappa_k}} \bar{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x}$$

and so

$$\lim_{h \rightarrow 0} \int_{\Omega_k} A_{k-1} \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k \, d\mathbf{x} = 0,$$

which shows that $\lim_{h \rightarrow 0} \|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d} = 0$ and hence the strong convergence of $\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k$. It remains to show (7.21b), we have

$$|\hat{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k} \leq |\bar{\vartheta}_k|_{\widehat{\mathcal{J}}(k),\kappa_k} + |\tilde{\vartheta}_k|_{\widehat{\mathcal{J}}(k),0} \leq C_i |\vartheta_{k-1}|_{\widehat{\mathcal{J}}(k-1),\kappa_{k-1}} + C_{\text{eq}} \|\nabla_{\widehat{\mathcal{D}}_k} \tilde{\vartheta}_k\|_{L^2(\Omega_k)^d}$$

and the result follows. ■

The next Theorem can be proved with similar arguments as used in Section 7.2.

Theorem 7.17. *Let Assumption 6.2 hold. Consider $(\vartheta_k)_{k=1}^M$, the sequence given by the local scheme (7.20a) to (7.20d) and $u \in H_0^1(\Omega)$ the solution of (7.19). Then for $k = 1, \dots, M$, we have*

$$\lim_{h \rightarrow 0} \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)} = 0,$$

where the limit is taken for $h \in \mathcal{H}$.

7.4 Numerical experiments

In the following numerical experiments, we will use examples where the subdomains $\{\Omega_k\}_{k=1}^M$ and meshes $\{\mathfrak{T}_k\}_{k=1}^M$ are defined a priori. This might be realistic in applications where the location of the singularities or high contrast of the solution are known. When such knowledge is not available, we should use a posteriori error estimators for detecting the local subdomains. This is developed in Chapter 8.

In what follows $\{\Omega_k\}_{k=1}^M$ will be a sequence of embedded domains but we recall that this is not a requirement. In the examples we consider $f \in L^2(\Omega)$ and denote by $\zeta_k \in X_{\mathcal{D}_k}$ the solution of

$$\int_{\Omega} A(\Pi_{\mathcal{D}_k} \zeta_k) \nabla_{\mathcal{D}_k} \zeta_k \cdot \nabla_{\mathcal{D}_k} \phi_k \, d\mathbf{x} = \int_{\Omega} f \Pi_{\mathcal{D}_k} \phi_k \, d\mathbf{x} \quad \forall \phi_k \in X_{\mathcal{D}_k}, \quad (7.26)$$

we refer to ζ_k as the classical solution, that is, the one obtained by the usual scheme which solves the equations in the whole domain after each mesh refinement. We can write $\zeta_k = \hat{\zeta}_k + \eta_k$ with $\hat{\zeta}_k \in Y_{\mathcal{D}_k}$ and $\eta_k \in Z_{\mathcal{D}_k}$. We will often compare ϑ_k and $\hat{\vartheta}_k$ the solutions of (7.4) or (7.20) against ζ_k and $\hat{\zeta}_k$ respectively.

Computational cost. Since in our setting the meshes are defined a priori, only the most accurate solution ζ_M needs to be computed. For the iterative schemes (7.4) and (7.20) instead it is imperative to compute ϑ_k for $k = 1, \dots, M$. If for example a conjugate gradient method is used to solve the linear systems, then the computational cost of the local scheme can be considerably smaller than the classical scheme due to the smaller problems solved on the fine meshes. For nonlinear problems, the local scheme might be faster for any linear solver, as the nonlinear system is solved only on a coarse mesh (see Section 7.3). This is illustrated in our numerical experiments.

It is useful to define the quantities

$$\begin{aligned} \text{Local Err}(\hat{\vartheta}_k) &:= \|\nabla u - \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k\|_{L^2(\Omega_k)^d} + |\hat{\vartheta}_k|_{J(k), \kappa_k}, \\ \text{Global Err}(\vartheta_k) &:= \|\nabla u - \nabla_{\mathcal{D}_k} \vartheta_k\|_{L^2(\Omega)^d} + |\vartheta_k|_{J(k)}. \end{aligned}$$

Similarly we define Local Err($\hat{\zeta}_k$) and Global Err(ζ_k) for the local and global error of the classical solutions. The local and classical schemes have been implemented with the help of the C++ library **libMesh** [79].

7.4.1 Convergence rates

In this example we want to illustrate the results of Theorem 7.10 ((7.10b) and (7.10c)) and of Theorem 7.14 ((7.18)), hence we consider an example with smooth solution. Let $\Omega = [-1, 1] \times [-1, 1]$, $A = I_2$ the identity matrix and $f \in L^2(\Omega)$ such that the exact solution is

$$u(\mathbf{x}) = e^{-120\|\mathbf{x}\|^2}. \quad (7.27)$$

Let $M = 4$, the local domains are such that $\mathbf{x} \in \Omega_k$ if $\|\mathbf{x}\|_{\infty} < (5 - k)/4$ for $k = 1, \dots, 4$.

In the first experiment we want to illustrate the estimates (7.10b) and (7.18), i.e. the convergence of the local and global errors with respect to the global mesh size. For a fixed h we consider uniform simplicial meshes $\widehat{\mathcal{M}}_k$ on Ω_k with mesh size $h_{\widehat{\mathcal{M}}_k} = h/2^{k-1}$ and apply the local algorithm (7.4), we let $h \rightarrow 0$ and verify the convergence rates. From Figures 7.3(a) and 7.3(b) we see that (7.10b) and (7.18) are verified for the local solution ϑ_4 . We also see that the classical scheme gives results with the same accuracy as the local scheme, both for the local and the global error.

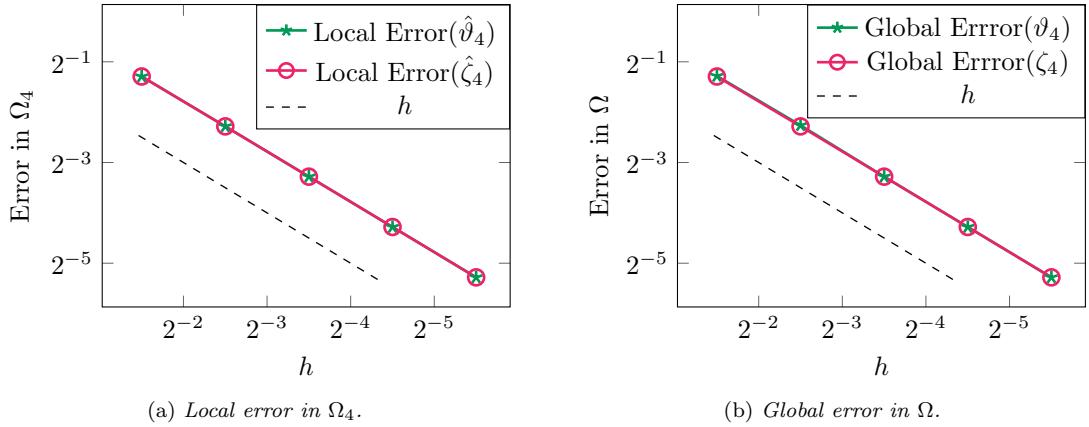


Figure 7.3. Convergence rates. Errors of the local ϑ_4 and classical ζ_4 solutions vs. the mesh size h .

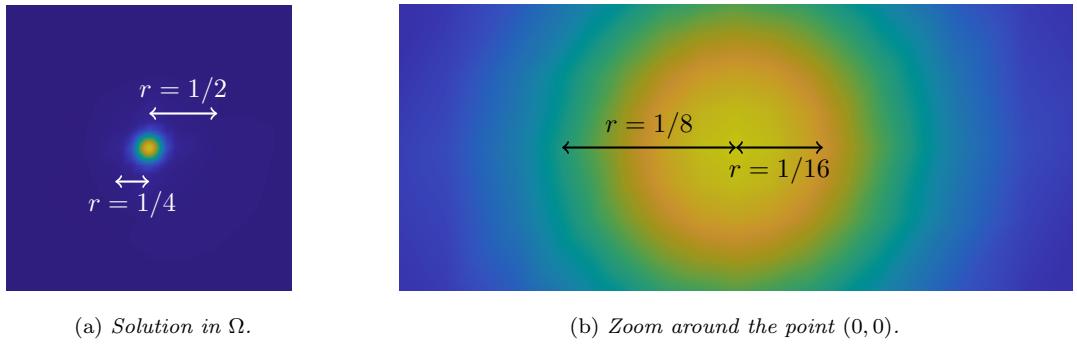


Figure 7.4. Convergence rates. Solution u from (7.27) and size r of local domain Ω_2 .

This example also indicates that if the high gradient regions are localized then there is no need of solving the problem in the whole domain after refinements.

In the next experiment we want to see the influence of the second term (boundary layer term) in the right-hand side of (7.10c) on Local Error($\hat{\vartheta}_M$). Let $r \in]0, 1[$, we set $M = 2$, $\Omega_1 = \Omega$ and $\Omega_2 = [-r, r] \times [-r, r]$. We fix $h_{\mathcal{M}_1} = \sqrt{2}/8$ the mesh size of \mathcal{M}_1 and let $h_{\widehat{\mathcal{M}}_2} \rightarrow 0$. We plot the results for different values of r (an illustration of this numerical experiment is given in Figure 7.4).

In Figure 7.5(a) we see that when r is large enough the local error scales with the local mesh size. If, instead, r is too small to cover the high gradient regions then the local error saturates very quickly. With $r = 1/8$ we get nice convergence up to $h_{\mathcal{M}_1}/h_{\widehat{\mathcal{M}}_2} = 16$ and with $r = 1/4, 1/2$ we do not see any saturation effects. In Figure 7.4 we see that $r = 1/16$ is too small to cover the local variations and indeed the local error does not converge. In Figure 7.5(b) we plot the total error on Ω . We remark that the error saturates for $r = 1/16, 1/8$. It is interesting to compare the results for $r = 1/8$ in Figures 7.5(a) and 7.5(b), in the first one there is a nice convergence while in the second an immediate saturation. This indicates that even if the error outside of Ω_2 is important, it does not propagate quickly into Ω_2 ; corroborating the results of Theorem 7.12, where we show that the error due to artificial boundary conditions is of higher order. Notice that the results displayed in Figure 7.5(b) are not in disagree with (7.18) in Theorem 7.14 since in this experiment $h = h_{\mathcal{M}_1}$ is kept constant.

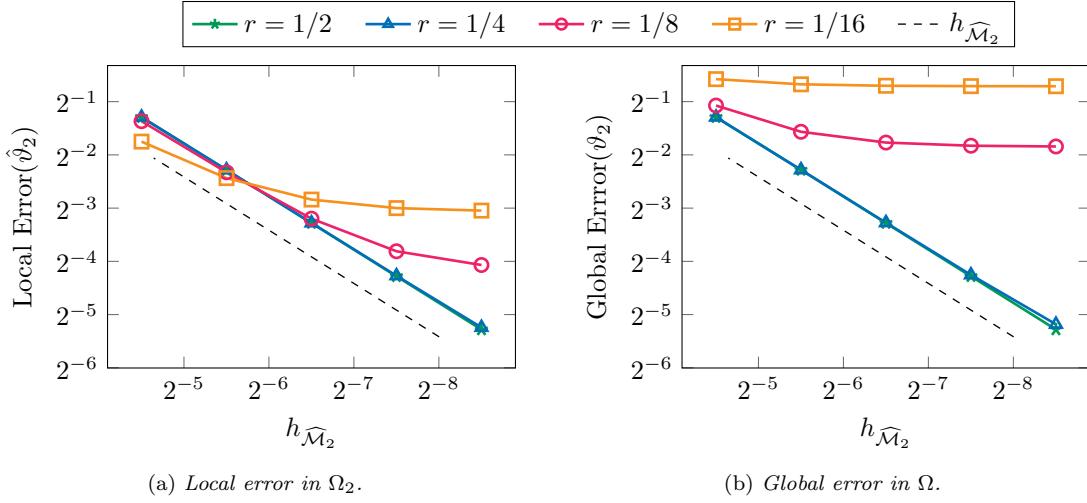


Figure 7.5. Convergence rates. Local solution's ϑ_2 error vs. local mesh size $h_{\widehat{\mathcal{M}}_2}$, for varying local domain Ω_2 diameter r .

In Figure 7.6 we plot the results of the same experiment shown in Figure 7.5 but for ζ_4 instead of ϑ_4 . We see that, again, the classical scheme gives similar results.

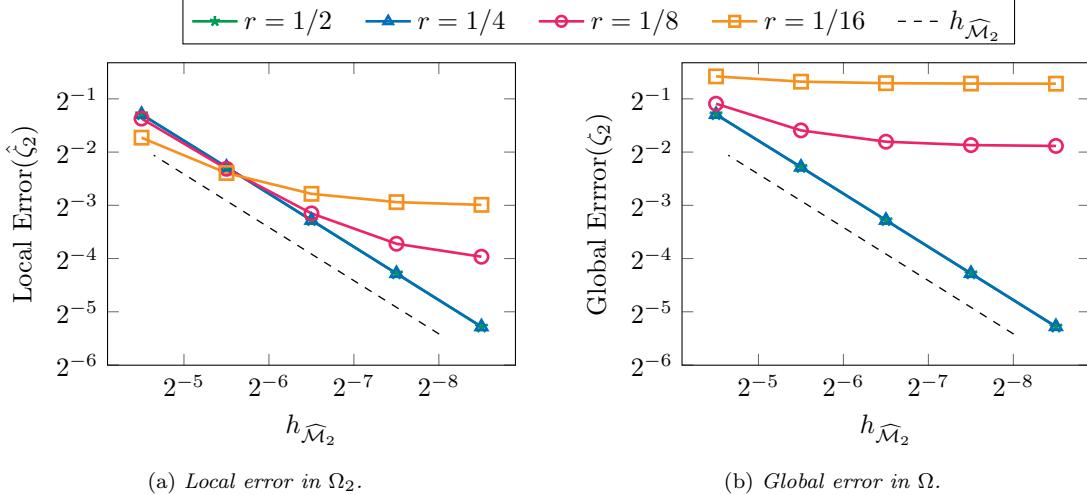


Figure 7.6. Convergence rates. Classical solution's ζ_2 error vs. local mesh size $h_{\widehat{\mathcal{M}}_2}$, for varying local domain Ω_2 diameter r .

In practice it is desirable to avoid the saturation effects seen in this experiment. For this reason, in Chapter 8 we develop a posteriori error estimators for the local scheme; they are capable of detecting the large error regions and thus mitigate saturation errors.

7.4.2 Influence of artificial boundary conditions

The goal of this experiment is to illustrate the result of Theorem 7.12, we want to illustrate numerically that the error due to artificial boundary conditions is of higher order as proved in

estimate (7.14). We consider the same problem as in Section 7.4.1 with $M = 2$, $\Omega_1 = \Omega$ and $\Omega_2 = [-r, r] \times [-r, r]$ with $r = 1/16$. We saw previously that with this choice of r the error originating from the artificial boundary conditions dominates the local error in Ω_2 . We solve (7.4) with different mesh sizes $h = h_{\mathcal{M}_1}$ using $h_{\widehat{\mathcal{M}}_2} = h/2^5$ as local mesh size, with this choice of $h_{\widehat{\mathcal{M}}_2}$ the dominating term in (7.14) is the last one, i.e. the one in $h^{3/2} \log(h^{-1})$. We measure the local errors in Ω_2 and plot the results in Figure 7.7. We see that indeed the local error satisfies (7.14) and converges even slightly faster than predicted.

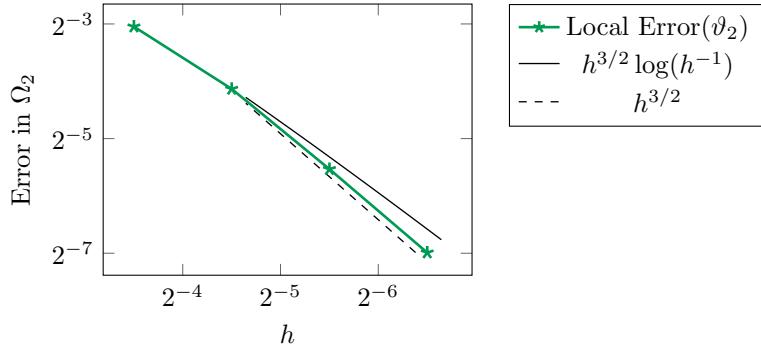


Figure 7.7. Influence of artificial boundary conditions. Convergence order of artificial boundary conditions' error term.

7.4.3 Non regular problem: discontinuous data

We next want to investigate numerically the convergence of the local scheme for a solution only belonging to $H^{1+\varepsilon}(\Omega)$ (for small $\varepsilon > 0$). The convergence is predicted by estimates (7.10a) and (7.17). We consider a problem that has been studied in [76] and [96] (in the context of a posteriori error estimators).

Let $\Omega = [-1, 1] \times [-1, 1]$ and consider problem (6.1) with $f = 0$. We divide the computational domain into four equal parts. Let the tensor be defined as $A(\mathbf{x}) = a_1 I_2$ in the 1st and 3rd quadrants and $A(\mathbf{x}) = a_2 I_2$ in the 2nd and 4th quadrants. The exact solution is given by $u(r, \theta) = r^\gamma \mu(\theta)$, where

$$\mu(\theta) = \begin{cases} \cos((\pi/2 - \sigma)\gamma) \cos((\theta - \pi/2 + \rho)\gamma) & \text{if } 0 \leq \theta \leq \pi/2, \\ \cos(\rho\gamma) \cos((\theta - \pi + \sigma)\gamma) & \text{if } \pi/2 < \theta \leq \pi, \\ \cos(\sigma\gamma) \cos((\theta - \pi - \rho)\gamma) & \text{if } \pi < \theta \leq 3\pi/2, \\ \cos((\pi/2 - \rho)\gamma) \cos((\theta - 3\pi/2 - \sigma)\gamma) & \text{if } 3\pi/2 < \theta < 2\pi. \end{cases}$$

The parameters γ , ρ , σ and $R := a_1/a_2$ satisfy the following non linear equations

$$\begin{aligned} R &= -\tan((\pi/2 - \sigma)\gamma) \cot(\rho\gamma), \\ 1/R &= -\tan(\rho\gamma) \cot(\sigma\rho), \\ R &= -\tan(\rho\gamma) \cot((\pi/2 - \rho)\gamma), \\ \max\{0, \pi\gamma - \pi\} &< 2\gamma\rho < \min\{\pi\gamma, \pi\}, \\ \max\{0, \pi - \pi\gamma\} &< -2\gamma\sigma < \min\{\pi, 2\pi - \pi\gamma\}. \end{aligned}$$

It is known that $u \in H^{1+\gamma-\epsilon}(\Omega)$ for any $\epsilon > 0$. In this example we choose $\gamma = 0.1$, $\sigma = -19\pi/4$, $\rho = \pi/4$ and $R \approx 161$.

In order to investigate the estimates (7.10a) and (7.17), we perform the same experiments as in Section 7.4.1, shown in Figure 7.3. We take $M = 4$ and the same domain and mesh sequences. We let $h \rightarrow 0$ and show the results in Figure 7.8. We find a convergence rate of 0.09, which is consistent with the results of [88] and the fact that u is almost in $H^{1.1}(\Omega)$. As was observed in Section 7.4.1, we see that the two solutions ϑ_4 and ζ_4 have the same errors, both in the local and global domains.

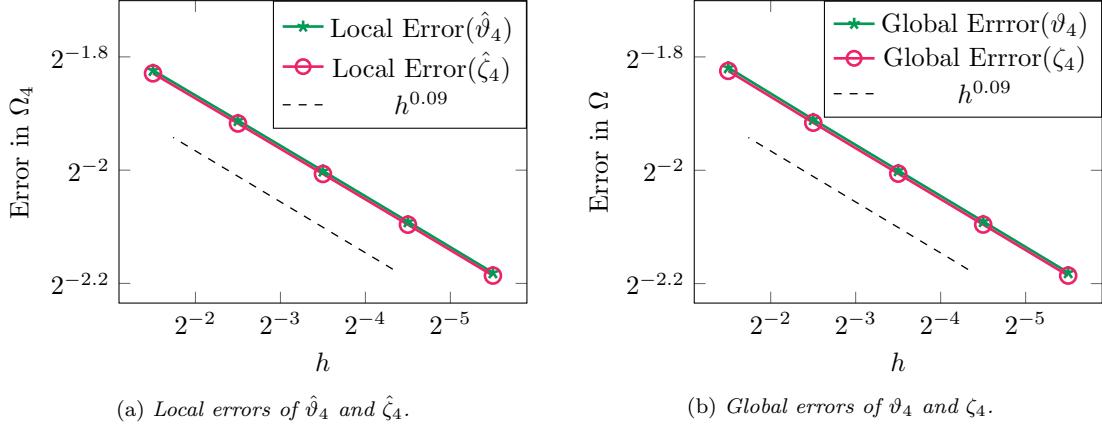


Figure 7.8. Non regular problem. Errors of the local ϑ_4 and classical ζ_4 solutions vs. the mesh size h .

The influence of the term $|\kappa_k - \xi_k|_{\partial\Omega_k^-}$ in (7.6) on Local Error($\hat{\vartheta}_M$) is established next, repeating the experiment of Section 7.4.1, taking Ω_2 depending on $r \in]0, 1[$ and letting $h_{\widehat{\mathcal{M}}_2} \rightarrow 0$. The results for ϑ_2 and ζ_2 are plotted in Figures 7.9 and 7.10 respectively. In contrast to the previous experiment, we do not have any saturation since the error inside the local domain largely dominates.

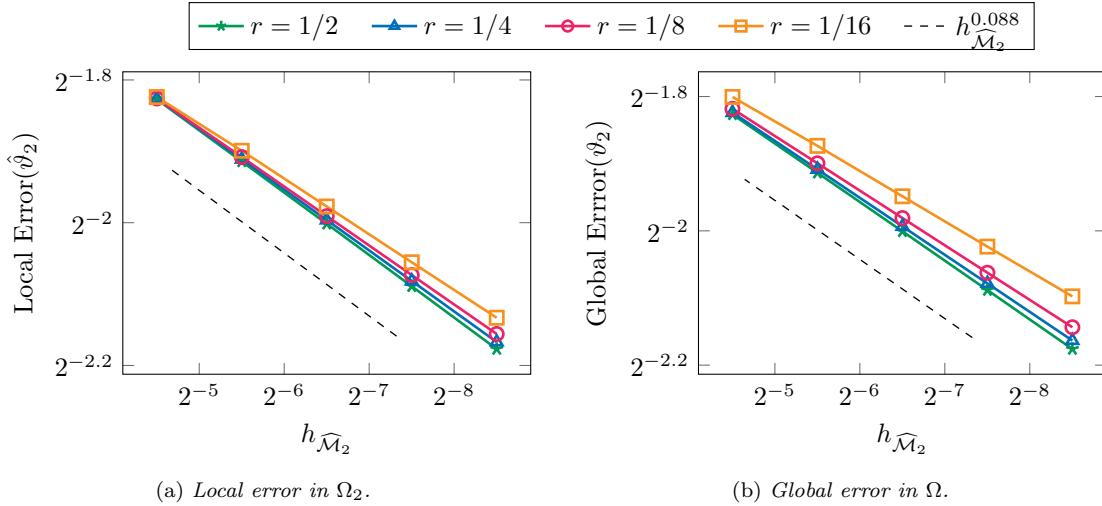


Figure 7.9. Non regular problem. Local solution's ϑ_2 error vs. local mesh size $h_{\widehat{\mathcal{M}}_2}$, for varying local domain Ω_2 diameter r .

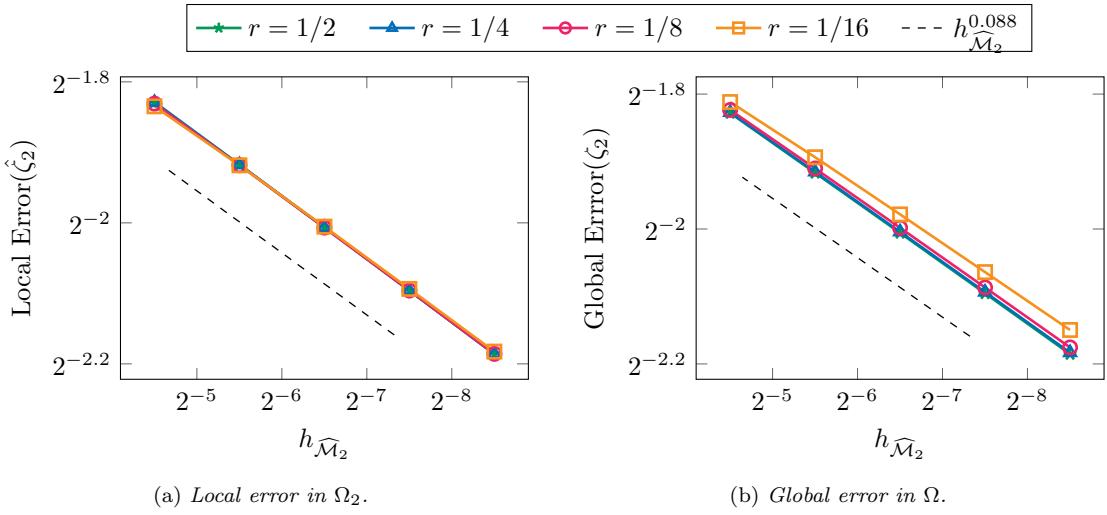


Figure 7.10. *Non regular problem. Classical solution's ζ_2 error vs. local mesh size $h_{\widehat{\mathcal{M}}_2}$, for varying local domain Ω_2 diameter r .*

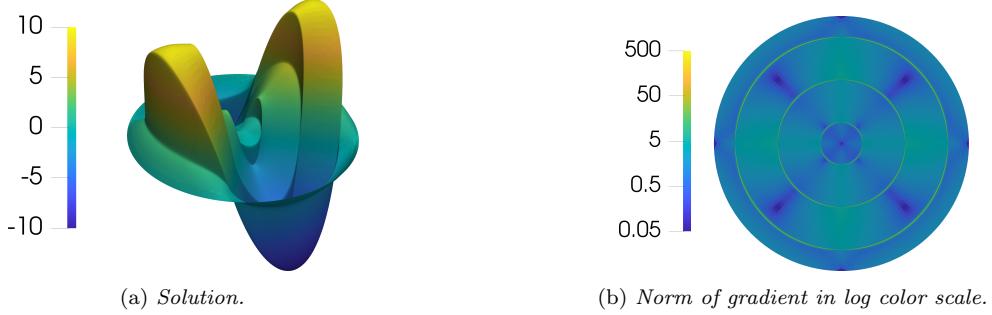


Figure 7.11. *Computational efficiency for a linear equation. Illustration of the solution and the gradient's norm.*

7.4.4 Computational efficiency for a linear equation

In this experiment we want to compare the numerical efficiency of the classical and local schemes on a linear equation, by computing a sequence of solutions with each scheme and plotting the accuracy against the cost.

We consider equation (6.1) with $\Omega = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 < 3\pi\}$, a diffusion tensor $A(\mathbf{x}) = \varepsilon + 1 - \sin(\|\mathbf{x}\|_2)^{100}$ with $\varepsilon = 10^{-3}$ and the force f is 1 if \mathbf{x} is the first or third quadrants and -1 else. An illustration of the solution is given in Figure 7.11. We choose five local domains defined as $\Omega_1 = \Omega$ and Ω_k for $k = 2, \dots, 5$ are neighborhoods of the three bright circles in Figure 7.11(b), more precisely

$$\Omega_k = \bigcup_{j=1}^3 \{\mathbf{x} \in \mathbb{R}^2 : |\|\mathbf{x}\|_2 - (2j-1)\pi/2| < 2^{2-k}\} \quad \text{for } k = 2, \dots, 5.$$

The meshes $\widehat{\mathcal{M}}_k$ are built so that $h_{\widehat{\mathcal{M}}_1} \approx 0.3$ and for $k = 2, \dots, 5$ we have $h_{\widehat{\mathcal{M}}_k} = h_{\widehat{\mathcal{M}}_{k-1}}/2$. We run the local scheme and at each level we compute the full error and cost of ϑ_k . As a measure of the cost for ϑ_k we take the sum of the time spent solving the linear systems up to level k using the conjugate gradient (CG) method with incomplete Cholesky (IC) factorization as preconditioner.

In [54] it is shown that this approach is the most robust and efficient for such problems. Then we run the classical scheme (7.26) on each mesh \mathcal{M}_k and obtain a sequence of solutions ζ_k . For each $k = 1, \dots, 5$ we compute the full error and cost of ζ_k . The cost is given by the time spent for solving the linear system at level k , where we use again CG with IC as preconditioner. Observe that here the cost is not cumulative as in the local method, since the classical scheme does not need ζ_{k-1} in order to compute ζ_k . In Figure 7.12(a) we plot the global error against the cost for both schemes, we see a significant speed-up for the local scheme. In Figure 7.12(b) we plot the speed-up in function of the error, the graph is obtained dividing the two curves seen in Figure 7.12(a).

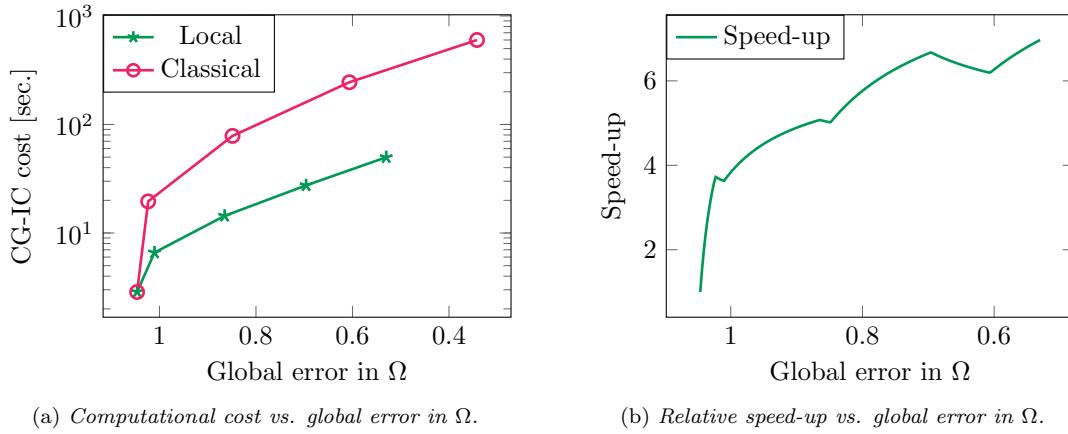


Figure 7.12. Computational efficiency for a linear equation. Computational times of local and classical schemes and relative speed-up..

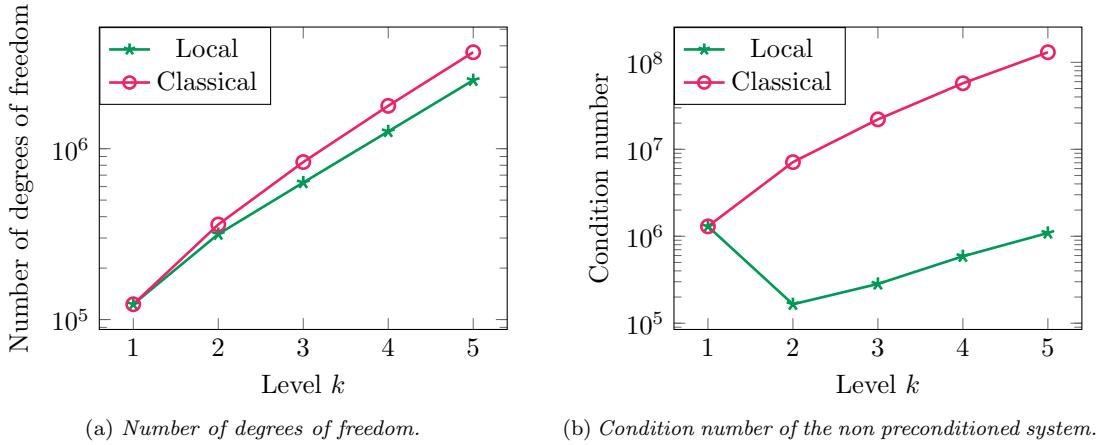
For linear problems such as in this experiment, the reason for the speed-up is not only the reduced number of degrees of freedom but mostly the condition number of the linear system. The classical scheme solves linear systems arising from spatial discretization on the whole domain; hence, the matrix has high variations in its components due to possibly high contrasts in the tensor and different scales of resolution. Instead, the local scheme uses matrices built from local discretizations, hence the tensor has milder variations and the elements of the local mesh have uniform size. This leads to matrices with smaller condition number. We see in Figure 7.13(a) that the number of degrees of freedom of the two schemes is almost the same, while in Figure 7.13(b) it is shown that the condition number of the stiffness matrix is much lower for the local scheme.

7.4.5 Quasilinear equation

In our last numerical experiment we want to compare the efficiency of the local and classical methods when solving a quasilinear equation. We consider the stationary Richards equation in pressure head form, given by

$$-\nabla \cdot (A(\mathbf{x}, h)\nabla(h - x_2)) = 0. \quad (7.28)$$

It describes the movement of a fluid in an unsaturated medium and can be put in the form of (7.19) with the change of variables $u = h - x_2$. We consider $\Omega = [-50, 50] \times [-50, 50]$ and add the Dirichlet condition $g(\mathbf{x}) = 10(50 - x_2) + 3(50 + x_2)$. The diffusion tensor is given by $A(\mathbf{x}, h) = A_s(\mathbf{x})A_r(h)$, where $A_s(\mathbf{x})$ is the conductivity in saturated conditions and $A_r(h)$ is the



(a) *Number of degrees of freedom.* (b) *Condition number of the non preconditioned system.*

Figure 7.13. Computational efficiency for a linear equation. Properties of the linear systems.

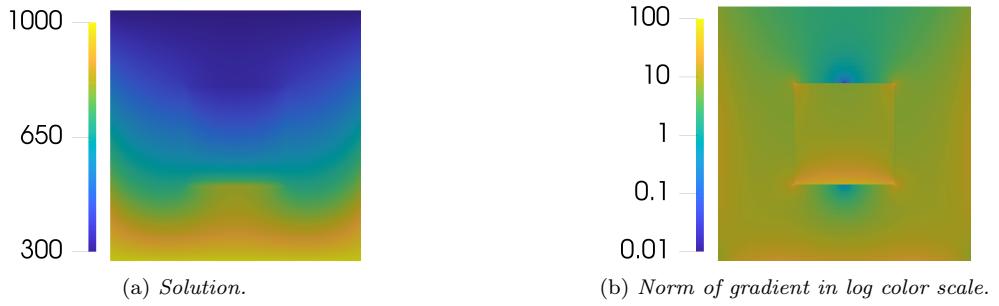


Figure 7.14. Quasilinear equation. Solution and norm of the gradient for the Richards equation.

relative conductivity. These latter quantities are defined by

$$A_s(\boldsymbol{x}) = \begin{cases} 10^{-3} & \text{if } \|\boldsymbol{x}\|_\infty \leq 20, \\ 1 & \text{else,} \end{cases} \quad A_r(h) = \frac{(1 - (ah)^{n-1}(1 + (ah)^n)^{-m})^2}{(1 + (ah)^n)^{m/2}}.$$

The model $A_r(h)$ has been taken from [119], where $m = 1 - 1/n$ is chosen. The parameters a, n are soil dependent: we choose $a = 1/500$ and $n = 2.68$, which is in the range of real case parameters. Remark that the tensor is discontinuous in \boldsymbol{x} and hence does not satisfy Assumption 6.2. We plot in Figure 7.14 the reference solution and the norm of its gradient; we see that the gradient is highly discontinuous.

Let $M = 4$, $\Omega_1 = \Omega$ and Ω_k for $k = 2, 3, 4$ defined by $\boldsymbol{x} \in \Omega_k$ if $\|\boldsymbol{x}\|_\infty \leq 20(1 + 2^{-k})$. First, we fix $h_{\widehat{\mathcal{M}}_k} = 100\sqrt{2}/2^{4+k}$ and compute the local solutions ϑ_k given by the local method (7.20). At the first level $k = 1$ we need to solve a nonlinear problem on a coarse grid using Newton iterations, where the initial guess is an extrapolation of the Dirichlet condition $g(\boldsymbol{x})$ on the whole domain. In the next levels $k > 1$ the local scheme solves a linear system using the Picard iteration step defined in (7.20d). At each level we compute the full error and cost of ϑ_k . As a measure of the cost for ϑ_k we take the sum of the time spent solving the linear and non linear systems up to level k . Since at $k = 1$ we perform a linearization of the system, it is no more symmetric because of the additional term, hence it has to be solved with the GMRES iterative scheme with incomplete LU (ILU) factorization as preconditioner, instead of CG with IC. In the following iterations, i.e. for $k \geq 2$, we solve a linear system and hence the CG scheme with IC is used.

Then we compute similar solutions with the classical method and compare the costs. For the classical solution we need, for each $k = 1, 2, 3, 4$, to solve (7.28) with the Newton method. As initial guess we take again $g(\mathbf{x})$ and the Newton iterations are stopped when the error of the classical solution ζ_k is similar to the one of ϑ_k . In about 3 or 4 Newton iterations we obtained errors differing by only about 1%. To measure the cost of ζ_k we consider the time spent in solving the nonlinear system at level k . The cost here is not cumulative as in the local method but on the other hand the linear systems to solve are not symmetric and the GMRES scheme with ILU preconditioner is used. In Figure 7.15(a) we plot the error against the cost for this experiment. We see that the local scheme performs much better than the classical scheme in terms of computational cost versus accuracy.

Finally, we compare the accuracy and cost of solving the local systems (7.20d) but where we replace $\Pi_{\widehat{\mathcal{D}}_{k-1}} \widehat{\vartheta}_{k-1}$ by $\Pi_{\widehat{\mathcal{D}}_k} \widehat{\vartheta}_k$, i.e., defining nonlinear local problems. These local systems must be solved with the Newton method and GMRES with ILU. We denote by θ_k^1 the solution where we use one Newton iteration and by θ_k^2 the solution with two Newton iterations. In Figure 7.15(b) we plot the error against the cost for ϑ_k , θ_k^1 and θ_k^2 . We see that one Picard iteration gives very similar errors to the one or two Newton iterations but at a smaller cost, thanks to the symmetric linear system.

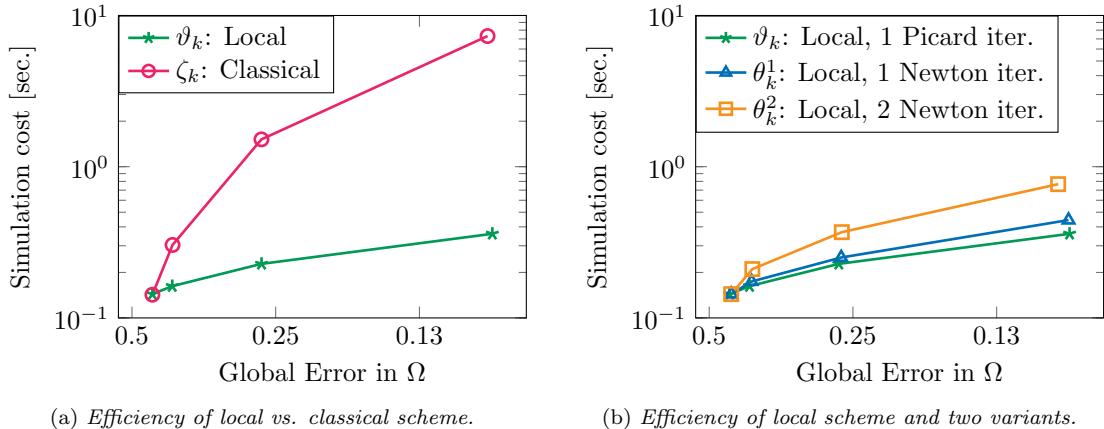


Figure 7.15. Quasilinear equation. Efficiency of classical scheme, local scheme and local scheme with Newton iterations instead of one Picard iteration.

8 A posteriori error analysis of a local adaptive discontinuous Galerkin scheme for advection-diffusion-reaction equations

In this chapter we derive an a posteriori error analysis for a local adaptive Symmetric Weighted Interior Penalty (SWIP) scheme for the advection-diffusion-reaction equation

$$\begin{aligned} -\nabla \cdot (A \nabla u) + \boldsymbol{\beta} \cdot \nabla u + \mu u &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \partial\Omega, \end{aligned} \quad (8.1)$$

where we assume that $\Omega \subset \mathbb{R}^d$ is a polytopal domain with $d \geq 2$, $\boldsymbol{\beta} \in W^{1,\infty}(\Omega)^d$, $\mu \in L^\infty(\Omega)$ and $A \in L^\infty(\Omega)^{d \times d}$, with $A(\mathbf{x})$ a symmetric piecewise constant matrix with eigenvalues in $[\lambda, \bar{\lambda}]$, where $\bar{\lambda} \geq \lambda > 0$. Moreover, we assume that $\mu - \frac{1}{2}\nabla \cdot \boldsymbol{\beta} \geq 0$ a.e. in Ω . This term $\mu - \frac{1}{2}\nabla \cdot \boldsymbol{\beta}$ appears in the symmetric part of the operator $\mathcal{B}(\cdot, \cdot)$ defined in (8.3) and hence the assumption $\mu - \frac{1}{2}\nabla \cdot \boldsymbol{\beta} \geq 0$ is needed for coercivity. Finally, we set $f \in L^2(\Omega)$. Under these assumptions, the unique weak solution $u \in H_0^1(\Omega)$ of (8.1) satisfies

$$\mathcal{B}(u, v) = \int_{\Omega} fv \, d\mathbf{x} \quad \forall v \in H_0^1(\Omega), \quad (8.2)$$

where

$$\mathcal{B}(u, v) = \int_{\Omega} (A \nabla u \cdot \nabla v + (\boldsymbol{\beta} \cdot \nabla u)v + \mu uv) \, d\mathbf{x}. \quad (8.3)$$

For the pure diffusion case, the method presented in this chapter is equivalent to the local WDGGD scheme of Chapter 7; therefore, it is proved to converge. Nonetheless, in Chapter 7 the subdomains Ω_k were defined upon an a priori knowledge of the high gradient regions, which is not always available. In this chapter we will derive a posteriori error indicators that can be used to identify the subdomains Ω_k and to bound the numerical error. The a posteriori error analysis is mainly based on [48], where a posteriori error estimators free of undetermined constants are derived for the SWIP method by means of consistent diffusive and advective flux reconstructions. Due to the local nature of our scheme, here we must relax the regularity requirements on the reconstructed fluxes and allow for jumps at subdomain boundaries. Still, the reconstructed fluxes will satisfy a conservation property.

The chapter is structured as follows. We first define the local adaptive SWIP method in Section 8.1, the a posteriori error estimators and the error bounds are given in Section 8.2. In Section 8.3 we define the flux reconstructions and prove the main results. Finally, in Section 8.4 we present various numerical examples illustrating the efficiency and versatility of the method.

8.1 The local adaptive discontinuous Galerkin scheme

In this section we introduce the local algorithm based on the discontinuous Galerkin method. We start with some preliminary definitions and assumptions on the mesh, mainly the same as in Chapter 7 and repeated here for completeness. Then we define the numerical scheme.

8.1.1 Preliminary definitions

We start by collecting some notations related to the geometry and the mesh of the subdomains, before recalling the definition of the discontinuous Galerkin finite element method.

Subdomains and meshes

Let $M \in \mathbb{N}^*$ and $\{\Omega_k\}_{k=1}^M$ be a sequence of open subdomains of Ω with $\Omega_1 = \Omega$. The domains Ω_k for $k \geq 2$ can be any subset of Ω , in practice they will be chosen by the a posteriori error estimators (see Section 8.1.2). We consider $\{\mathcal{M}_k\}_{k=1}^M$ a sequence of simplicial meshes on Ω and $\mathcal{F}_k = \mathcal{F}_{k,b} \cup \mathcal{F}_{k,i}$ is the set of boundary and internal faces of \mathcal{M}_k . The assumption below ensures that \mathcal{M}_{k+1} is a refinement of \mathcal{M}_k inside the subdomain Ω_{k+1} .

Assumption 8.1.

- a) For each $k = 1, \dots, M$, $\overline{\Omega}_k = \cup_{K \in \mathcal{M}_k, K \subset \Omega_k} \overline{K}$.
- b) For $k = 1, \dots, M - 1$,
 - i) $\{K \in \mathcal{M}_{k+1} : K \subset \Omega \setminus \Omega_{k+1}\} = \{K \in \mathcal{M}_k : K \subset \Omega \setminus \Omega_{k+1}\}$,
 - ii) if $K, T \in \mathcal{M}_k$ with $K \subset \Omega_{k+1}$, $T \subset \Omega \setminus \Omega_{k+1}$ and $\partial K \cap \partial T \neq \emptyset$ then $K \in \mathcal{M}_{k+1}$,
 - iii) if $K \in \mathcal{M}_k$ and $K \subset \Omega_{k+1}$, either $K \in \mathcal{M}_{k+1}$ or K is a union of elements in \mathcal{M}_{k+1} .

Let $\widehat{\mathcal{M}}_k = \{K \in \mathcal{M}_k : K \subset \Omega_k\}$ and $\widehat{\mathcal{F}}_k = \widehat{\mathcal{F}}_{k,b} \cup \widehat{\mathcal{F}}_{k,i}$ the set of faces of $\widehat{\mathcal{M}}_k$, with $\widehat{\mathcal{F}}_{k,b}$ and $\widehat{\mathcal{F}}_{k,i}$ the boundary and internal faces, respectively. Condition a) in Assumption 8.1 ensures that $\widehat{\mathcal{M}}_k$ is a simplicial mesh on Ω_k . Condition b) guarantees that in $\Omega \setminus \Omega_{k+1}$ and in the neighborhood of $\partial\Omega_{k+1} \setminus \partial\Omega$ the meshes \mathcal{M}_k and \mathcal{M}_{k+1} are equal and that \mathcal{M}_{k+1} is a refinement of \mathcal{M}_k in Ω_{k+1} . An example of domains and meshes satisfying Assumption 8.1 is illustrated in Figure 7.1. Note that Assumption 8.1 corresponds to items a) and b) of Assumption 7.1.

Discontinuous Galerkin finite element method

The local adaptive discontinuous Galerkin method will solve local elliptic problems in Ω_k by using a discontinuous Galerkin scheme introduced in [49], which we recall here. In what follows, $\mathfrak{T} = (D, \mathcal{M}, \mathcal{F})$ denotes a tuple defined by a domain D , a simplicial mesh \mathcal{M} on D and its set of faces $\mathcal{F} = \mathcal{F}_b \cup \mathcal{F}_i$. In practice we will consider $\mathfrak{T}_k = (\Omega, \mathcal{M}_k, \mathcal{F}_k)$ or $\widehat{\mathfrak{T}}_k = (\Omega_k, \widehat{\mathcal{M}}_k, \widehat{\mathcal{F}}_k)$. For $\mathfrak{T} = (D, \mathcal{M}, \mathcal{F})$ we define

$$V(\mathfrak{T}) = \{v \in L^2(D) : v|_K \in \mathbb{P}_\ell(K), \forall K \in \mathcal{M}\}, \quad (8.4)$$

where $\mathbb{P}_\ell(K)$ is the set of polynomials in K of total degree ℓ . As usual for such discontinuous Galerkin methods we need to define appropriate means, jumps, weights and penalization parameters. For $K \in \mathcal{M}$ we denote \mathbf{n}_K the unit normal outward to K and $\mathcal{F}_K = \{\sigma \in \mathcal{F} : \sigma \subset \partial K\}$. Let $\sigma \in \mathcal{F}_i$ and $K, T \in \mathcal{M}$ with $\sigma = \partial K \cap \partial T$, then $\mathbf{n}_\sigma = \mathbf{n}_K$ and

$$\delta_{K,\sigma} = \mathbf{n}_\sigma^\top A|_K \mathbf{n}_\sigma, \quad \delta_{T,\sigma} = \mathbf{n}_\sigma^\top A|_T \mathbf{n}_\sigma.$$

The weights are defined by

$$\omega_{K,\sigma} = \frac{\delta_{T,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}, \quad \omega_{T,\sigma} = \frac{\delta_{K,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}$$

and the penalization parameters by

$$\gamma_\sigma = \frac{2\delta_{K,\sigma}\delta_{T,\sigma}}{\delta_{K,\sigma} + \delta_{T,\sigma}}, \quad \nu_\sigma = \frac{1}{2}|\boldsymbol{\beta} \cdot \mathbf{n}_\sigma|.$$

If $\sigma \in \mathcal{F}_b$ and $K \in \mathcal{M}$ with $\sigma = \partial K \cap \partial D$ then \mathbf{n}_σ is \mathbf{n}_D the unit outward normal to ∂D and

$$\delta_{K,\sigma} = \mathbf{n}_\sigma^\top A|_K \mathbf{n}_\sigma, \quad \omega_{K,\sigma} = 1, \quad \gamma_\sigma = \delta_{K,\sigma}, \quad \nu_\sigma = \frac{1}{2}|\boldsymbol{\beta} \cdot \mathbf{n}_\sigma|.$$

Let $g \in L^2(\partial D)$, we define the means and jumps of $v \in V(\mathfrak{T})$ as follows. For $\sigma \in \mathcal{F}_b$ with $\sigma = \partial K \cap \partial D$ we set

$$\{v\}_{\omega,\sigma} = v|_K, \quad \{v\}_{g,\sigma} = \frac{1}{2}(v|_K + g), \quad [v]_{g,\sigma} = v|_K - g$$

and for $\sigma \in \mathcal{F}_i$ with $\sigma = \partial K \cap \partial T$

$$\{v\}_{\omega,\sigma} = \omega_{K,\sigma}v|_K + \omega_{T,\sigma}v|_T, \quad \{v\}_{g,\sigma} = \frac{1}{2}(v|_K + v|_T), \quad [v]_{g,\sigma} = v|_K - v|_T.$$

We define $\{\cdot\}_\sigma := \{\cdot\}_{0,\sigma}$ and $\{\cdot\}_\sigma := \{\cdot\}_{0,\sigma}$. A similar notation holds for vector valued functions and whenever no confusion can arise the subscript σ is omitted. Let h_σ be the diameter of σ and $\eta_\sigma > 0$ a user parameter, for $u, v \in V(\mathfrak{T})$ we define the bilinear form

$$\begin{aligned} \mathcal{B}(u, v, \mathfrak{T}, g) = & \int_D (A \nabla u \cdot \nabla v + (\mu - \nabla \cdot \boldsymbol{\beta})uv - u\boldsymbol{\beta} \cdot \nabla v) \, d\mathbf{x} \\ & - \sum_{\sigma \in \mathcal{F}} \int_\sigma ([v]\{A \nabla u\}_\omega \cdot \mathbf{n}_\sigma + [u]_g \{A \nabla v\}_\omega \cdot \mathbf{n}_\sigma) \, d\mathbf{y} \\ & + \sum_{\sigma \in \mathcal{F}} \int_\sigma ((\eta_\sigma \frac{\gamma_\sigma}{h_\sigma} + \nu_\sigma)[u]_g[v] + \boldsymbol{\beta} \cdot \mathbf{n}_\sigma \{u\}_g[v]) \, d\mathbf{y}, \end{aligned} \quad (8.5)$$

where the gradients are taken element wise. The bilinear form $\mathcal{B}(\cdot, \cdot, \mathfrak{T}, g)$ will be used to approximate elliptic problems in D with Dirichlet condition g . This scheme is known as the Symmetric Weighted Interior Penalty (SWIP) scheme [49]. The SWIP method is an improvement of the Symmetric Interior Penalty scheme (SIP) [23], where the weights are defined as $\omega_{K,\sigma} = \omega_{T,\sigma} = 1/2$. The use of diffusivity-dependent averages increases the robustness of the method for problems with strong diffusion discontinuities. The bilinear form defined in (8.5) is mathematically equivalent to other formulations where $v\boldsymbol{\beta} \cdot \nabla u$ or $\nabla \cdot (\boldsymbol{\beta}u)v$ appear instead of $u\boldsymbol{\beta} \cdot \nabla v$ (see [49] and [39, Section 4.6.2]). Our choice of formulation is convenient to express local conservation laws (see [39, Section 2.2.3]).

8.1.2 The local adaptive algorithm

In this section we present the local scheme. In order to facilitate the comprehension of the method, we start with an informal description and then provide a pseudo-code for the algorithm. We denote u_k the global solutions on Ω and \hat{u}_k the local solutions on Ω_k , which are used to correct the global solutions.

Given a discretization $\mathfrak{T}_1 = (\Omega, \mathcal{M}_1, \mathcal{F}_1)$ on Ω the local scheme computes a first approximate solution $u_1 \in V(\mathfrak{T}_1)$ to (8.2). The algorithm then performs the following steps for $k = 2, \dots, M$.

- i) Given the current solution u_{k-1} , identify the region Ω_k where the error is large and define a new refined mesh \mathcal{M}_k satisfying Assumption 8.1 by iterating the following steps.
 - a) For each element $K \in \mathcal{M}_{k-1}$ compute an error indicator $\eta_{M,K}$ (defined in (8.11)) and mark the local domain Ω_k using the fixed energy fraction marking strategy [40, Section 4.2]. Hence, Ω_k is defined as the union of the elements with largest error indicator $\eta_{M,K}$ and it is such that the error committed inside of Ω_k is at least a prescribed fraction of the total error.
 - b) Define the new mesh \mathcal{M}_k by refining the elements $K \in \mathcal{M}_{k-1}$ with $K \subset \Omega_k$.
 - c) Enlarge the local domain Ω_k defined at step a) by adding a one element wide boundary layer (i.e. in order to satisfy item 2b of Assumption 8.1).
 - d) Define the local mesh $\widehat{\mathcal{M}}_k$ by the elements of \mathcal{M}_k inside of Ω_k .
- ii) Solve a local elliptic problem in Ω_k on the refined mesh $\widehat{\mathcal{M}}_k$ using u_{k-1} as artificial boundary conditions on $\partial\Omega_k \setminus \partial\Omega$. The solution is denoted $\hat{u}_k \in V(\widehat{\mathfrak{T}}_k)$, where $\widehat{\mathfrak{T}}_k = (\Omega_k, \widehat{\mathcal{M}}_k, \widehat{\mathcal{F}}_k)$.
- iii) The local solution \hat{u}_k is used to correct the previous solution u_{k-1} inside of Ω_k and obtain the new global solution u_k .

The pseudo-code of the local scheme is given in Algorithm 2, where $\chi_{\Omega \setminus \Omega_k}$ is the indicator function of $\Omega \setminus \Omega_k$ and $(\cdot, \cdot)_k$ is the inner product in $L^2(\Omega_k)$. The function LocalDomain(u_k, \mathfrak{T}_k) used in Algorithm 2 performs steps a)-d) of i), it is given in Algorithm 3 for completeness.

Algorithm 2 LocalScheme(\mathfrak{T}_1)

```

Find  $u_1 \in V(\mathfrak{T}_1)$  solution to  $\mathcal{B}(u_1, v_1, \mathfrak{T}_1, 0) = (f, v_1)_1$  for all  $v_1 \in V(\mathfrak{T}_1)$ .
for  $k = 2, \dots, M$  do
     $(\mathfrak{T}_k, \widehat{\mathfrak{T}}_k) = \text{LocalDomain}(u_{k-1}, \mathfrak{T}_{k-1})$ .
     $g_k = u_{k-1} \chi_{\Omega \setminus \Omega_k} \in V(\mathfrak{T}_k)$ .
    Find  $\hat{u}_k \in V(\widehat{\mathfrak{T}}_k)$  solution to  $\mathcal{B}(\hat{u}_k, v_k, \widehat{\mathfrak{T}}_k, g_k) = (f, v_k)_k$  for all  $v_k \in V(\widehat{\mathfrak{T}}_k)$ .
     $u_k = g_k + \hat{u}_k \in V(\mathfrak{T}_k)$ .
end for

```

Algorithm 3 LocalDomain(u_k, \mathfrak{T}_k)

```

Compute the marking estimator  $\eta_{M,K}$  (see (8.11)) for all  $K \in \mathcal{M}_k$ .
 $\eta_{M,\infty} = \max_{K \in \mathcal{M}_k} \eta_{M,K}$ 
 $\eta_{M,2} = \left( \sum_{K \in \mathcal{M}_k} \eta_{M,K}^2 \right)^{1/2}$ 
 $\theta = 0.5, \beta = 0.7, \gamma = 0.4$ 
do
     $\overline{\Omega}_{k+1} = \bigcup_{\{K \in \mathcal{M}_k : \eta_{M,K} > \theta \eta_{M,\infty}\}} \overline{K}$ 
     $\tilde{\eta}_{M,2} = \left( \sum_{\{K \in \mathcal{M}_k : \eta_{M,K} > \theta \eta_{M,\infty}\}} \eta_{M,K}^2 \right)^{1/2}$ 
     $\theta \leftarrow \beta \theta$ 
while  $\tilde{\eta}_{M,2} < \gamma \eta_{M,2}$ 
 $\mathcal{M}_{k+1}$  is defined by refining the elements  $K \in \mathcal{M}_k$  with  $K \subset \Omega_{k+1}$ .
 $\overline{L}_{k+1} = \bigcup_{\{K \in \mathcal{M}_{k+1} : K \subset \Omega \setminus \Omega_{k+1}, \overline{K} \cap \overline{\Omega}_{k+1} \neq \emptyset\}} \overline{K}$ .
 $\Omega_{k+1} \leftarrow \Omega_{k+1} \cup L_{k+1}$ 
 $\widehat{\mathcal{M}}_{k+1} = \{K \in \mathcal{M}_{k+1} : K \subset \Omega_{k+1}\}$ 
return  $\mathfrak{T}_k = (\Omega, \mathcal{M}_k, \mathcal{F}_k)$  and  $\widehat{\mathfrak{T}}_k = (\Omega_{k+1}, \widehat{\mathcal{M}}_{k+1}, \widehat{\mathcal{F}}_{k+1})$ .

```

We end this section showing that for purely diffusive problems the local scheme introduced here

and the local scheme of Section 7.2 are equivalent.

Theorem 8.2. *Let $\beta = \mathbf{0}$, $\mu = 0$ and $A_K := A|_K$ the restriction of A to an element $K \in \mathcal{M}_1$ be constant, let as well $\{\Omega_k\}_{k=1}^M$ be a fixed sequence of local subdomains. Then $u_k = \Pi_{\mathcal{D}_k} \vartheta_k$, where u_k is the solution provided by Algorithm 2 and ϑ_k the solution given by (7.4).*

Proof. From (7.4) and $f \in L^2(\Omega)$ we have

$$\int_{\Omega_k} A \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi \, d\mathbf{x} = (f, \Pi_{\widehat{\mathcal{D}}_k} \varphi)_k \quad \forall \varphi \in Y_{\mathcal{D}_k}$$

and following Section 6.2.3 we obtain

$$\begin{aligned} & \int_{\Omega_k} A \nabla_{\widehat{\mathcal{D}}_k, \kappa_k} \hat{\vartheta}_k \cdot \nabla_{\widehat{\mathcal{D}}_k} \varphi \, d\mathbf{x} \\ &= \int_{\Omega_k} A \nabla \Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k \cdot \nabla \Pi_{\widehat{\mathcal{D}}_k} \varphi \, d\mathbf{x} \\ & \quad - \sum_{\sigma \in \widehat{\mathcal{F}}_k} \int_{\sigma} ([\Pi_{\widehat{\mathcal{D}}_k} \varphi] \{A \nabla \Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k\}_{\omega} + [\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k]_{\Pi_{\mathcal{D}_k} \kappa_k} \{A \nabla \Pi_{\widehat{\mathcal{D}}_k} \varphi\}_{\omega}) \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \\ & \quad + \sum_{\sigma \in \widehat{\mathcal{F}}_k} \eta_{\sigma} \frac{\gamma_{\sigma}}{h_{\sigma}} \int_{\sigma} [\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k]_{\Pi_{\mathcal{D}_k} \kappa_k} [\Pi_{\widehat{\mathcal{D}}_k} \varphi] \, d\mathbf{y} \\ &= \mathcal{B}(\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k, \Pi_{\widehat{\mathcal{D}}_k} \varphi, \widehat{\mathfrak{T}}_k, \Pi_{\mathcal{D}_k} \kappa_k). \end{aligned}$$

Hence,

$$\mathcal{B}(\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k, \Pi_{\widehat{\mathcal{D}}_k} \varphi, \widehat{\mathfrak{T}}_k, \Pi_{\mathcal{D}_k} \kappa_k) = (f, \Pi_{\widehat{\mathcal{D}}_k} \varphi)_k \quad \forall \varphi \in Y_{\mathcal{D}_k}. \quad (8.6)$$

In contrast, \hat{u}_k satisfies

$$\mathcal{B}(\hat{u}_k, v_k, \widehat{\mathfrak{T}}_k, g_k) = (f, v_k)_k \quad \forall v_k \in V(\widehat{\mathfrak{T}}_k)$$

and as $\Pi_{\widehat{\mathcal{D}}_k} Y_{\mathcal{D}_k} = V(\widehat{\mathfrak{T}}_k)$, in (8.6) we replace $\Pi_{\widehat{\mathcal{D}}_k} \varphi$ by v_k , from which follows

$$\mathcal{B}(\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k, v_k, \widehat{\mathfrak{T}}_k, \Pi_{\mathcal{D}_k} \kappa_k) = \mathcal{B}(\hat{u}_k, v_k, \widehat{\mathfrak{T}}_k, g_k) \quad \forall v_k \in V(\widehat{\mathfrak{T}}_k).$$

Next, we have the following recursive implications: if $\Pi_{\mathcal{D}_k} \kappa_k = g_k$ then $\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k = \hat{u}_k$, as $\Pi_{\widehat{\mathcal{D}}_k} \hat{\vartheta}_k, \hat{u}_k \in \Pi_{\widehat{\mathcal{D}}_k} Y_{\mathcal{D}_k}$, and therefore $\Pi_{\mathcal{D}_k} \vartheta_k = u_k$ which implies $\Pi_{\mathcal{D}_{k+1}} \kappa_{k+1} = g_{k+1}$. Since $\kappa_1 = 0$ and $g_1 = 0$ the theorem is proved. ■

8.2 A posteriori error estimators via flux and potential reconstructions

The error estimators used to mark the local domains Ω_k and to provide error bounds on the numerical solution u_k are introduced here.

In the framework of selfadjoint elliptic problems, the equilibrated fluxes method [18, 26] is a technique largely used to derive a posteriori error estimators free of undetermined constants and is based on the definition of local fluxes which satisfy a local conservation property. Since local fluxes and conservation properties are intrinsic to the discontinuous Galerkin formulation, this discretization is well suited for the equilibrated fluxes method [17, 37]. In [46, 78] the Raviart-Thomas-Nédélec space is used to build an $H_{\text{div}}(\Omega)$ conforming reconstruction \mathbf{t}_h of the discrete diffusive flux $-A \nabla u_h$. A diffusive flux \mathbf{t}_h with optimal divergence, in the sense that it coincides with the orthogonal projection of the right-hand side f onto the discontinuous Galerkin

space, is obtained. In [48] the authors extend this approach to convection-diffusion-reaction equations by defining an $H_{\text{div}}(\Omega)$ conforming convective flux \mathbf{q}_h approximating βu_h and satisfying a conservation property.

We follow a similar strategy and define in the next section error estimators as functions of the diffusive and convective flux reconstructions $\mathbf{t}_k, \mathbf{q}_k$ for the local scheme, as well as an $H_0^1(\Omega)$ conforming potential reconstruction s_k of the solution u_k .

8.2.1 A posteriori error estimators

The error estimators as functions of the potential reconstruction s_k approximating the solution u_k , the diffusive and convective fluxes \mathbf{t}_k and \mathbf{q}_k approximating $-A\nabla u_k$ and βu_k , respectively, are defined in this section.

Following the iterative and local nature of our scheme, we define the diffusive and convective flux reconstructions as

$$\mathbf{t}_k = \mathbf{t}_{k-1} \chi_{\Omega \setminus \Omega_k} + \hat{\mathbf{t}}_k, \quad \mathbf{q}_k = \mathbf{q}_{k-1} \chi_{\Omega \setminus \Omega_k} + \hat{\mathbf{q}}_k, \quad (8.7)$$

where $\mathbf{t}_0 = \mathbf{q}_0 = 0$ and $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k$ are $H_{\text{div}}(\Omega_k)$ conforming flux reconstructions of $-A\nabla \hat{u}_k, \beta \hat{u}_k$, respectively, and where \hat{u}_k is the local solution. They satisfy a local conservation property and are defined in Section 8.3.1. We readily see that this definition allows for flux jumps at the subdomains boundaries, while giving enough freedom to define $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k$ in a way that a conservation property is satisfied. The flux reconstructions are used to measure the non conformity of the numerical fluxes. In the same spirit we define a potential reconstruction $s_k \in H_0^1(\Omega)$ used to measure the non conformity of the numerical solution. It is defined recursively as

$$s_k = s_{k-1} \chi_{\Omega \setminus \Omega_k} + \hat{s}_k, \quad (8.8)$$

where $s_0 = 0$ and $\hat{s}_k \in H^1(\Omega_k)$ is such that $s_k \in H_0^1(\Omega)$. More details about the definitions of $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k$ and \hat{s}_k will be given in Section 8.3.1, for the time being we will define the error estimators.

Let $K \in \mathcal{M}_k$, $v \in H^1(K)$,

$$\|v\|_K^2 = \|A^{1/2} \nabla v\|_{L^2(K)^d}^2 + \|(\mu - \frac{1}{2} \nabla \cdot \boldsymbol{\beta})^{1/2} v\|_{L^2(K)}^2 \quad (8.9)$$

and $m_K, \tilde{m}_K, m_\sigma, D_{t,K,\sigma}, c_{\beta,\mu,K} > 0$ some known constants which will be defined in Section 8.3.2. The non conformity of the numerical solution u_k is measured by the estimator

$$\eta_{NC,K} = \|u_k - s_k\|_K. \quad (8.10a)$$

The residual estimator is

$$\eta_{R,K} = m_K \|f - \nabla \cdot \mathbf{t}_k - \nabla \cdot \mathbf{q}_k - (\mu - \nabla \cdot \boldsymbol{\beta}) u_k\|_{L^2(K)}, \quad (8.10b)$$

which can be seen as the residual of (8.2) where we first replace u by u_k , then $-A\nabla u_k$ by \mathbf{t}_k , βu_k by \mathbf{q}_k and finally use the Green theorem. The error estimators defined in (8.10c) to (8.10j) measure the error introduced by these substitutions and the error introduced when applying the Green theorem to $\mathbf{t}_k, \mathbf{q}_k$, which are not in $H_{\text{div}}(\Omega)$.

The diffusive flux estimator measures the difference between $-A\nabla u_k$ and \mathbf{t}_k . It is given by

$\eta_{DF,K} = \min\{\eta_{DF,K}^1, \eta_{DF,K}^2\}$, where

$$\begin{aligned}\eta_{DF,K}^1 &= \|A^{1/2}\nabla u_k + A^{-1/2}\mathbf{t}_k\|_{L^2(K)^d}, \\ \eta_{DF,K}^2 &= m_K \|(\mathcal{I} - \pi_0)(\nabla \cdot (A\nabla u_k + \mathbf{t}_k))\|_{L^2(K)} \\ &\quad + \tilde{m}_K^{1/2} \sum_{\sigma \in \mathcal{F}_K} C_{t,K,\sigma}^{1/2} \|(A\nabla u_k + \mathbf{t}_k) \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)},\end{aligned}\tag{8.10c}$$

π_0 is the L^2 -orthogonal projector onto $\mathbb{P}_0(K)$ and \mathcal{I} is the identity operator. Let $\sigma \in \mathcal{F}_k$ and $\pi_{0,\sigma}$ be the L^2 -orthogonal projector onto $\mathbb{P}_0(\sigma)$. The convection and upwinding estimators measure the difference between βu_k , βs_k and \mathbf{q}_k and are defined by

$$\eta_{C,1,K} = m_K \|(\mathcal{I} - \pi_0)(\nabla \cdot (\mathbf{q}_k - \beta s_k))\|_{L^2(K)},\tag{8.10d}$$

$$\eta_{C,2,K} = \frac{1}{2} c_{\beta,\mu,K}^{-1/2} \|(\nabla \cdot \beta)(u_k - s_k)\|_{L^2(K)},\tag{8.10e}$$

$$\tilde{\eta}_{C,1,K} = m_K \|(\mathcal{I} - \pi_0)(\nabla \cdot (\mathbf{q}_k - \beta u_k))\|_{L^2(K)},\tag{8.10f}$$

$$\eta_{U,K} = \sum_{\sigma \in \mathcal{F}_K} \chi_\sigma m_\sigma \|\pi_{0,\sigma} \llbracket \mathbf{q}_k - \beta s_k \rrbracket \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)},\tag{8.10g}$$

$$\tilde{\eta}_{U,K} = \sum_{\sigma \in \mathcal{F}_K} \chi_\sigma m_\sigma \|\pi_{0,\sigma} \llbracket \mathbf{q}_k - \beta u_k \rrbracket \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)},\tag{8.10h}$$

where $\chi_\sigma = 2$ if $\sigma \in \mathcal{F}_{k,b}$ and $\chi_\sigma = 1$ if $\sigma \in \mathcal{F}_{k,i}$. Finally, we introduce the jump estimators coming from the application of the Green theorem to \mathbf{t}_k and \mathbf{q}_k (see Lemma 8.8). Those are defined by

$$\eta_{\Gamma,1,K} = \frac{1}{2} (|K| c_{\beta,\mu,K})^{-1/2} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \|\pi_{0,\sigma} \llbracket \mathbf{q}_k \rrbracket \cdot \mathbf{n}_\sigma\|_{L^1(\sigma)},\tag{8.10i}$$

$$\eta_{\Gamma,2,K} = \frac{1}{2} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} D_{t,K,\sigma} \|\llbracket \mathbf{t}_k \rrbracket \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)}.\tag{8.10j}$$

We end the section defining the marking error estimator $\eta_{M,K}$ employed to mark Ω_k in the LocalDomain routine of in Algorithm 2 (see also Algorithm 3), let

$$\begin{aligned}\eta_{M,K} &= \eta_{NC,K} + \eta_{R,K} + \alpha \eta_{DF,K} + \eta_{C,1,K} + \eta_{C,2,K} + \alpha \eta_U \\ &\quad + \eta_{\Gamma,1,K} + \eta_{\Gamma,2,K} + \tilde{\eta}_{C,1,K} + \tilde{\eta}_U.\end{aligned}\tag{8.11}$$

The weight α appearing in (8.11) is due to the fact that $\eta_{DF,K}$ and $\eta_{U,K}$ are the principal error indicators. In the numerical experiments we use $\alpha = 5$.

8.2.2 Main results

We state here our main results related to the a posteriori analysis of the local scheme, in particular we will provide error bounds on the numerical solution u_k which are robust in singularly perturbed regimes and free of undetermined constants.

We start defining the norms for which we provide the error bounds, the same norms are used in [48]. The operator \mathcal{B} defined in (8.3) can be written $\mathcal{B} = \mathcal{B}_S + \mathcal{B}_A$, where \mathcal{B}_S and \mathcal{B}_A are symmetric and skew-symmetric operators defined by

$$\begin{aligned}\mathcal{B}_S(u, v) &= \int_{\Omega} (A\nabla u \cdot \nabla v + (\mu - \frac{1}{2}\nabla \cdot \beta)uv) \, dx, \\ \mathcal{B}_A(u, v) &= \int_{\Omega} (\beta \cdot \nabla u + \frac{1}{2}(\nabla \cdot \beta)u)v \, dx,\end{aligned}$$

for $u, v \in H^1(\mathcal{M}_k)$. The energy norm is defined by the symmetric operator as

$$\|v\|^2 = \mathcal{B}_S(v, v) = \|A^{1/2} \nabla v\|_{L^2(\Omega)^d}^2 + \|(\mu - \frac{1}{2} \nabla \cdot \beta)^{1/2} v\|_{L^2(\Omega)}^2,$$

observe that $\|v\|^2 = \sum_{K \in \mathcal{M}_k} \|v\|_K^2$ with $\|\cdot\|_K$ as in (8.9). Since the norm $\|\cdot\|$ is defined by the symmetric operator, it is well suited to study problems with dominant diffusion or reaction. On the other hand, it is inappropriate for convection dominated problems since it lacks a term measuring the error along the velocity direction. For this kind of problems we use the augmented norm

$$\|v\|_{\oplus} = \|v\| + \sup_{\substack{w \in H_0^1(\Omega) \\ \|w\|=1}} (\mathcal{B}_A(v, w) + \mathcal{B}_J(v, w)),$$

where

$$\mathcal{B}_J(v, w) = - \sum_{\sigma \in \mathcal{F}_{k,i}} \int_{\sigma} [\beta v] \cdot \mathbf{n}_{\sigma} \{ \pi_0 w \} \, dy$$

is a term needed to sharpen the error bounds. The next two theorems give a bound on the error of the local scheme, measured in the energy or the augmented norm.

Theorem 8.3. *Let $u \in H_0^1(\Omega)$ be the solution to (8.2), $u_k \in V(\mathfrak{T}_k)$ given by Algorithm 2, $s_k \in V(\mathfrak{T}_k) \cap H_0^1(\Omega)$ from (8.8) and (8.16) and $\mathbf{t}_k, \mathbf{q}_k \in \mathbf{RTN}_{\mathfrak{s}}(\mathcal{M}_k)$ be defined by (8.7) and (8.13). Then, the error measured in the energy norm is bounded as*

$$\|u - u_k\| \leq \eta = \left(\sum_{K \in \mathcal{M}_k} \eta_{NC,K}^2 \right)^{1/2} + \left(\sum_{K \in \mathcal{M}_k} \eta_{1,K}^2 \right)^{1/2},$$

where $\eta_{1,K} = \eta_{R,K} + \eta_{DF,K} + \eta_{C,1,K} + \eta_{C,2,K} + \eta_{U,K} + \eta_{\Gamma,1,K} + \eta_{\Gamma,2,K}$.

Theorem 8.4. *Under the same assumptions of Theorem 8.3, the error measured in the augmented norm is bounded as*

$$\|u - u_k\|_{\oplus} \leq \tilde{\eta} = 2\eta + \left(\sum_{K \in \mathcal{M}_k} \eta_{2,K}^2 \right)^{1/2},$$

with η from Theorem 8.3 and $\eta_{2,K} = \eta_{R,K} + \eta_{DF,K} + \tilde{\eta}_{C,1,K} + \tilde{\eta}_{U,K} + \eta_{\Gamma,1,K} + \eta_{\Gamma,2,K}$.

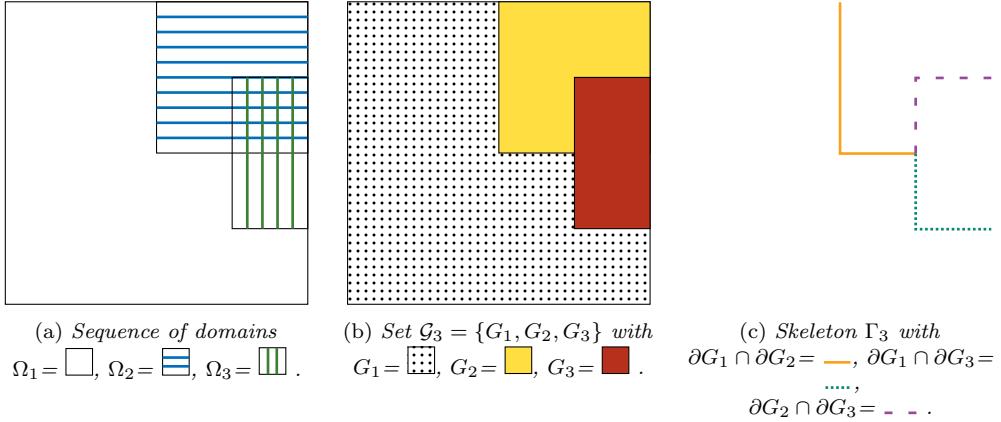
The error estimators of Theorems 8.3 and 8.4 are free of undetermined constants, indeed they depend on the numerical solution, the smallest eigenvalues of the diffusion tensor, on the essential minimum of $\mu - \frac{1}{2} \nabla \cdot \beta$, the mesh size and known geometric constants.

8.3 Potential and flux reconstructions, proofs of the main results

In this section, we will define the potential, diffusion and advection reconstructions, define the geometric constants appearing in the error estimators defined in (8.10a) to (8.10j) and finally prove Theorems 8.3 and 8.4.

8.3.1 Potential and flux reconstruction via the equilibrated flux method

We define here the flux reconstructions $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k$ of (8.7) and the potential reconstruction \hat{s}_k of (8.8). In what follows we assume that \mathcal{M}_k has hanging nodes only on the interface $\partial\Omega_k \setminus \partial\Omega$ since it simplifies the analysis, however we can follow [48, Appendix] to drop this requirement.

Figure 8.1. Example of sequence of domains $\Omega_1, \Omega_2, \Omega_3$, set G_3 and skeleton Γ_3 .

We start defining some broken Sobolev spaces and then the potential and flux reconstructions. For $k = 1, \dots, M$ let $\mathcal{G}_k = \{G_j \mid j = 1, \dots, k\}$, where $G_k = \Omega_k$ and

$$G_j = \Omega_j \setminus \cup_{i=j+1}^k \bar{\Omega}_i \quad \text{for } j = 1, \dots, k-1.$$

In Figures 8.1(a) and 8.1(b) we give an example of a sequence of domains Ω_k and the corresponding set \mathcal{G}_k . We define the broken spaces

$$\begin{aligned} H_{\text{div}}(\mathcal{G}_k) &= \{\mathbf{v} \in L^2(\Omega)^d : \mathbf{v}|_G \in H_{\text{div}}(G) \text{ for all } G \in \mathcal{G}_k\}, \\ H^1(\mathcal{M}_k) &= \{v_k \in L^2(\Omega) : v_k|_K \in H^1(K) \text{ for all } K \in \mathcal{M}_k\}, \end{aligned}$$

the divergence and gradient operators in $H_{\text{div}}(\mathcal{G}_k)$ and $H^1(\mathcal{M}_k)$ are taken element wise. We extend the jump operator $[\![\cdot]\!]_\sigma$ to the broken space $H^1(\mathcal{M}_k)$. We call Γ_k the internal skeleton of \mathcal{G}_k , that is

$$\Gamma_k = \{\partial G_i \cap \partial G_j \mid G_i, G_j \in \mathcal{G}_k, i \neq j\},$$

an example of Γ_k is given in Figure 8.1(c). For each $\gamma \in \Gamma_k$ we define $\mathcal{F}_\gamma = \{\sigma \in \mathcal{F}_{k,i} \mid \sigma \subset \gamma\}$ and set \mathbf{n}_γ , the normal to γ , as $\mathbf{n}_\gamma|_\sigma = \mathbf{n}_\sigma$. The jump $[\![\cdot]\!]_\gamma$ on γ is defined by $[\![\cdot]\!]_\gamma|_\sigma = [\![\cdot]\!]_\sigma$.

In [48] the reconstructed fluxes live in $H_{\text{div}}(\Omega)$. For the local algorithm we need to build such fluxes using the recursive relation (8.7). This leads to fluxes having jumps across the boundaries of the subdomains, i.e. $\gamma \in \Gamma_k$, hence they lie in the broken space $H_{\text{div}}(\mathcal{G}_k)$. In the rest of this section we explain how to build fluxes which are in an approximation space of $H_{\text{div}}(\mathcal{G}_k)$ and satisfy a local conservation property. We start by introducing a broken version of the usual Raviart-Thomas-Nédélec spaces [97, 100], which we define as

$$\mathbf{RTN}_z(\mathcal{M}_k) := \{\mathbf{v}_k \in H_{\text{div}}(\mathcal{G}_k) : \mathbf{v}_k|_K \in \mathbf{RTN}_z(K) \text{ for all } K \in \mathcal{M}_k\}, \quad (8.12)$$

where $z \in \{\ell-1, \ell\}$ and $\mathbf{RTN}_z(K) = \mathbb{P}_z(K)^d + \mathbf{x} \mathbb{P}_z(K)$. In order to build functions in $\mathbf{RTN}_z(\mathcal{M}_k)$ we need a characterization of this space. Let $\mathbf{v}_k \in L^2(\Omega)^d$ such that $\mathbf{v}_k|_K \in \mathbf{RTN}_z(K)$ for each $K \in \mathcal{M}_k$, it is known that $\mathbf{v}_k \in H_{\text{div}}(\Omega)$ if and only if $[\![\mathbf{v}_k]\!]_\sigma \cdot \mathbf{n}_\sigma = 0$ for all $\sigma \in \mathcal{F}_{k,i}$ (see [39, Lemma 1.24]). Since we search for fluxes \mathbf{v}_k in $H_{\text{div}}(\mathcal{G}_k)$, we relax this condition and allow $[\![\mathbf{v}_k]\!]_\gamma \cdot \mathbf{n}_\gamma \neq 0$ for $\gamma \in \Gamma_k$.

Lemma 8.5. *Let $\mathbf{v}_k \in L^2(\Omega)^d$ be such that $\mathbf{v}_k|_K \in \mathbf{RTN}_z(K)$ for each $K \in \mathcal{M}_k$, then $\mathbf{v}_k \in \mathbf{RTN}_z(\mathcal{M}_k)$ if and only if $[\![\mathbf{v}_k]\!]_\sigma \cdot \mathbf{n}_\sigma = 0$ for all $\sigma \notin \cup_{\gamma \in \Gamma_k} \mathcal{F}_\gamma$.*

Proof. Following the lines of [39, Lemma 1.24]. ■

The diffusive and convective fluxes $\mathbf{t}_k, \mathbf{q}_k \in \mathbf{RTN}_\varepsilon(\mathcal{M}_k)$ are defined recursively as in (8.7), where $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k \in \mathbf{RTN}_\varepsilon(\widehat{\mathcal{M}}_k)$, with

$$\mathbf{RTN}_\varepsilon(\widehat{\mathcal{M}}_k) := \{\mathbf{v}_k \in H_{\text{div}}(\Omega_k) : \mathbf{v}_k \in \mathbf{RTN}_\varepsilon(K) \text{ for all } K \in \widehat{\mathcal{M}}_k\},$$

are given by the relations

$$\begin{aligned} \int_\sigma \hat{\mathbf{t}}_k \cdot \mathbf{n}_\sigma p_k \, d\mathbf{y} &= \int_\sigma (-\{A \nabla \hat{u}_k\}_\omega \cdot \mathbf{n}_\sigma + \eta_\sigma \frac{\gamma_\sigma}{h_\sigma} [\hat{u}_k]_{g_k}) p_k \, d\mathbf{y}, \\ \int_\sigma \hat{\mathbf{q}}_k \cdot \mathbf{n}_\sigma p_k \, d\mathbf{y} &= \int_\sigma (\boldsymbol{\beta} \cdot \mathbf{n}_\sigma \{ \hat{u}_k \}_{g_k} + \nu_\sigma [\hat{u}_k]_{g_k}) p_k \, d\mathbf{y} \end{aligned} \quad (8.13a)$$

for all $\sigma \in \widehat{\mathcal{F}}_k$ and $p_k \in \mathbb{P}_\varepsilon(\sigma)$ and

$$\begin{aligned} \int_K \hat{\mathbf{t}}_k \cdot \hat{\mathbf{r}}_k \, d\mathbf{x} &= - \int_K A \nabla \hat{u}_k \cdot \hat{\mathbf{r}}_k \, d\mathbf{x} + \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \omega_{K,\sigma} [\hat{u}_k]_{g_k} A|_K \hat{\mathbf{r}}_k \cdot \mathbf{n}_\sigma \, d\mathbf{y}, \\ \int_K \hat{\mathbf{q}}_k \cdot \hat{\mathbf{r}}_k \, d\mathbf{x} &= \int_K \hat{u}_k \boldsymbol{\beta} \cdot \hat{\mathbf{r}}_k \, d\mathbf{x} \end{aligned} \quad (8.13b)$$

for all $K \in \widehat{\mathcal{M}}_k$ and $\hat{\mathbf{r}}_k \in \mathbb{P}_{\varepsilon-1}(K)^d$. Since $\hat{\mathbf{t}}_k|_K \cdot \mathbf{n}_\sigma, \hat{\mathbf{q}}_k|_K \cdot \mathbf{n}_\sigma \in \mathbb{P}_\varepsilon(\sigma)$ (see [30, Proposition 3.2]) then (8.13a) defines $\hat{\mathbf{t}}_k|_K \cdot \mathbf{n}_\sigma, \hat{\mathbf{q}}_k|_K \cdot \mathbf{n}_\sigma$ on σ . The remaining degrees of freedom are fixed by (8.13b) [30, Proposition 3.3]. Thanks to (8.13a) we have $[\hat{\mathbf{t}}_k] \cdot \mathbf{n}_\sigma = 0$ and $[\hat{\mathbf{q}}_k] \cdot \mathbf{n}_\sigma = 0$ for $\sigma \in \widehat{\mathcal{F}}_{k,i}$ and hence $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k \in \mathbf{RTN}_\varepsilon(\widehat{\mathcal{M}}_k)$. By construction it follows $\mathbf{t}_k, \mathbf{q}_k \in \mathbf{RTN}_\varepsilon(\mathcal{M}_k)$.

Let $K \in \mathcal{M}_k$ and π_ε be the L^2 -orthogonal projector onto $\mathbb{P}_\varepsilon(K)$, the following lemma states a local conservation property of the reconstructed fluxes. The proof follows the lines of [48, Lemma 2.1]

Lemma 8.6. *Let $u_k \in V(\mathfrak{T}_k)$ be given by Algorithm 2 and $\mathbf{t}_k, \mathbf{q}_k \in H_{\text{div}}(\mathcal{G}_k)$ defined by (8.7) and (8.13). For all $K \in \mathcal{M}_k$ it holds*

$$(\nabla \cdot \mathbf{t}_k + \nabla \cdot \mathbf{q}_k + \pi_\varepsilon((\mu - \nabla \cdot \boldsymbol{\beta}) u_k))|_K = \pi_\varepsilon f|_K.$$

Proof. Let $K \in \mathcal{M}_k$ and $j = \max\{j = 1, \dots, k : K \subset \Omega_j\}$, then $K \in \widehat{\mathcal{M}}_j$, $\mathbf{t}_k|_K = \hat{\mathbf{t}}_j|_K$, $\mathbf{q}_k|_K = \hat{\mathbf{q}}_j|_K$ and $u_k|_K = \hat{u}_j|_K$. Let $v_j \in \mathbb{P}_\varepsilon(K)$, by the Green theorem we have

$$\int_K (\nabla \cdot \hat{\mathbf{t}}_j + \nabla \cdot \hat{\mathbf{q}}_j) v_j \, d\mathbf{x} = - \int_K (\hat{\mathbf{t}}_j + \hat{\mathbf{q}}_j) \cdot \nabla v_j \, d\mathbf{x} + \sum_{\sigma \in \mathcal{F}_K} \int_\sigma v_j (\hat{\mathbf{t}}_j + \hat{\mathbf{q}}_j) \cdot \mathbf{n}_K \, d\mathbf{y} \quad (8.14)$$

and using $\mathcal{B}(\hat{u}_j, v_j, \widehat{\mathfrak{T}}_j, g_j) = (f, v_j)_j$ it follows

$$\begin{aligned} \int_K f v_j \, d\mathbf{x} &= \int_K (A \nabla \hat{u}_j \cdot \nabla v_j + (\mu - \nabla \cdot \boldsymbol{\beta}) \hat{u}_j v_j - \hat{u}_j \boldsymbol{\beta} \cdot \nabla v_j) \, d\mathbf{x} \\ &\quad - \sum_{\sigma \in \mathcal{F}_K} \int_\sigma ([v_j] \{ A \nabla \hat{u}_j \}_\omega \cdot \mathbf{n}_\sigma + [\hat{u}_j]_{g_j} \{ A \nabla v_j \}_\omega \cdot \mathbf{n}_\sigma) \, d\mathbf{y} \\ &\quad + \sum_{\sigma \in \mathcal{F}_K} \int_\sigma ((\eta_\sigma \frac{\gamma_\sigma}{h_\sigma} + \nu_\sigma) [\hat{u}_j]_{g_j} [v_j] + \boldsymbol{\beta} \cdot \mathbf{n}_\sigma \{ \hat{u}_j \}_{g_j} [v_j]) \, d\mathbf{y}. \end{aligned}$$

Since $\{ A \nabla v_j \}_\omega = \omega_{K,\sigma} A|_K \nabla v_j$ and $[v_j] \mathbf{n}_\sigma = v_j|_K \mathbf{n}_K$, using (8.13) and (8.14), we obtain

$$\int_K f v_j \, d\mathbf{x} = \int_K (\nabla \cdot \hat{\mathbf{t}}_j + \nabla \cdot \hat{\mathbf{q}}_j + (\mu - \nabla \cdot \boldsymbol{\beta}) \hat{u}_j) v_j \, d\mathbf{x} \quad (8.15)$$

and the result follows from $\nabla \cdot \hat{\mathbf{t}}_j, \nabla \cdot \hat{\mathbf{q}}_j \in \mathbb{P}_\varepsilon(K)$, $\mathbf{t}_k|_K = \hat{\mathbf{t}}_j|_K$, $\mathbf{q}_k|_K = \hat{\mathbf{q}}_j|_K$ and $u_k|_K = \hat{u}_j|_K$. ■

In order to define the $H_0^1(\Omega)$ conforming approximation s_k of u_k we will need the so-called Oswald operator already considered in [75] for a posteriori estimates. Let $\mathfrak{T} = (D, \mathcal{M}, \mathcal{F})$, $g \in C^0(\partial D)$ and consider $\mathcal{O}_{\mathfrak{T},g} : V(\mathfrak{T}) \rightarrow V(\mathfrak{T}) \cap H^1(D)$, for a function $v \in V(\mathfrak{T})$ the value of $\mathcal{O}_{\mathfrak{T},g}v$ is prescribed at the Lagrange interpolation nodes p of the conforming finite element space $V(\mathfrak{T}) \cap H^1(D)$. Let $p \in \overline{D}$ be a Lagrange node, if $p \notin \partial D$ we set

$$\mathcal{O}_{\mathfrak{T},g}v(p) = \frac{1}{\#\mathcal{M}_p} \sum_{K \in \mathcal{M}_p} v|_K(p),$$

where $\mathcal{M}_p = \{K \in \mathcal{M} : p \in \overline{K}\}$. If instead $p \in \partial D$ then $\mathcal{O}_{\mathfrak{T},g}v(p) = g(p)$, where g is the Dirichlet condition at ∂D . The reconstructed potential $s_k \in V(\mathfrak{T}_k) \cap H_0^1(\Omega)$ is built as in (8.8), where

$$\hat{s}_k = \mathcal{O}_{\widehat{\mathfrak{T}}_k, s_{k-1}} \hat{u}_k. \quad (8.16)$$

8.3.2 Constant definitions and preliminary results

Here we define the constants appearing in (8.10a) to (8.10j) and derive preliminary results needed to prove Theorems 8.3 and 8.4.

Let $K \in \mathcal{M}_k$ and $\sigma \in \mathcal{F}_K$, we recall that $|K|$ is the measure of K and $|\sigma|$ the $d - 1$ dimensional measure of σ . We denote by $c_{A,K}$ the minimal eigenvalue of $A|_K$. Next, we denote by $c_{\beta,\mu,K}$ the essential minimum of $\mu - \frac{1}{2}\nabla \cdot \beta \geq 0$ on K . In what follows we will assume that $\mu - \frac{1}{2}\nabla \cdot \beta > 0$ a.e. in Ω , hence $c_{\beta,\mu,K} > 0$ for all $K \in \mathcal{M}_k$, and provide error estimators under this assumption. We explain in Section 8.3.4 how to overcome this limitation slightly modifying the proofs and error estimators.

The cutoff functions m_K, \tilde{m}_K and m_σ are defined by

$$\begin{aligned} m_K &= \min\{C_p^{1/2}h_K c_{A,K}^{-1/2}, c_{\beta,\mu,K}^{-1/2}\}, \\ \tilde{m}_K &= \min\{(C_p + C_p^{1/2})h_K c_{A,K}^{-1}, h_K^{-1}c_{\beta,\mu,K}^{-1} + c_{\beta,\mu,K}^{-1/2}c_{A,K}^{-1/2}/2\}, \\ m_\sigma^2 &= \min\{\max_{K \in \mathcal{M}_\sigma} \{3d|\sigma|h_K^2|K|^{-1}c_{A,K}^{-1}\}, \max_{K \in \mathcal{M}_\sigma} \{|\sigma||K|^{-1}c_{\beta,\mu,K}^{-1}\}\}, \end{aligned} \quad (8.17)$$

where $C_p = 1/\pi^2$ is an optimal Poincaré constant for convex domains [98]. We next state the following bounds

$$\|v - \pi_0 v\|_{L^2(K)} \leq m_K \|v\|_K \quad \forall K \in \mathcal{M}_k, \quad (8.18a)$$

$$\|v - \pi_0 v|_K\|_{L^2(\sigma)} \leq C_{t,K,\sigma}^{1/2} \tilde{m}_K^{1/2} \|v\|_K \quad \forall \sigma \in \mathcal{F}_k \text{ and } K \in \mathcal{M}_\sigma, \quad (8.18b)$$

$$\|\llbracket \pi_0 v \rrbracket\|_{L^2(\sigma)} \leq m_\sigma \sum_{K \in \mathcal{M}_\sigma} \|v\|_K \quad \forall \sigma \in \mathcal{F}_k, \quad (8.18c)$$

where $\mathcal{M}_\sigma = \{K \in \mathcal{M}_k : \sigma \subset \partial K\}$ and $C_{t,K,\sigma}$ is the constant of the trace inequality

$$\|v|_K\|_{L^2(\sigma)}^2 \leq C_{t,K,\sigma} (h_K^{-1} \|v\|_{L^2(K)}^2 + \|v\|_{L^2(K)} \|\nabla v\|_{L^2(K)^d}). \quad (8.19)$$

It has been proved in [112, Lemma 3.12] that for a simplex it holds $C_{t,K,\sigma} = |\sigma|h_K/|K|$.

Let us briefly explain the role of constants (8.17) and how the bounds (8.18) are obtained. We observe that for each bound in (8.18) the cut off functions take the minimum between two possible values, allowing for robust error estimation in singularly perturbed regimes. For (8.18a), using the Poincaré inequality [98, equation 3.2] we have

$$\begin{aligned} \|v - \pi_0 v\|_{L^2(K)} &\leq C_p^{1/2}h_K \|\nabla v\|_{L^2(K)^d} \\ &\leq C_p^{1/2}h_K c_{A,K}^{-1/2} \|A^{1/2} \nabla v\|_{L^2(K)^d} \leq C_p^{1/2}h_K c_{A,K}^{-1/2} \|v\|_K. \end{aligned} \quad (8.20a)$$

Denoting $(\cdot, \cdot)_K$ the $L^2(K)$ inner product, it holds

$$\|v - \pi_0 v\|_{L^2(K)}^2 = (v - \pi_0 v, v - \pi_0 v)_K = (v - \pi_0 v, v)_K \leq \|v - \pi_0 v\|_{L^2(K)} \|v\|_{L^2(K)},$$

hence

$$\|v - \pi_0 v\|_{L^2(K)} \leq \|v\|_{L^2(K)} \leq c_{\beta, \mu, K}^{-1/2} \|(\mu - \frac{1}{2} \nabla \cdot \beta)^{1/2} v\|_{L^2(K)} \leq c_{\beta, \mu, K}^{-1/2} \|v\|_K \quad (8.20b)$$

and (8.18a) follows. The choice between bounds (8.20a) and (8.20b) depends on whether the problem is singularly perturbed or not. Bounds (8.18b) and (8.18c) are obtained similarly, see [35, Lemma 4.2] and [126, Lemma 4.5]. Finally, for $K \in \mathcal{M}_k$ and $\sigma \in \mathcal{F}_K$ we define

$$D_{t, K, \sigma} = \left(\frac{C_{t, K, \sigma}}{2h_K c_{\beta, \mu, K}} \left(1 + \sqrt{1 + h_K^2 \frac{c_{\beta, \mu, K}}{c_{A, K}}} \right) \right)^{1/2}, \quad (8.21)$$

which is used to bound $\|v|_K\|_{L^2(\sigma)}$ in terms of $\|v\|_K$ in the next lemma.

Lemma 8.7. *Let $v_k \in H^1(\mathcal{M}_k)$, for each $K \in \mathcal{M}_k$ and $\sigma \in \mathcal{F}_K$ it holds*

$$\|v_k|_K\|_{L^2(\sigma)} \leq D_{t, K, \sigma} \|v_k\|_K.$$

Proof. Let $v_k \in H^1(\mathcal{M}_k)$ and $\epsilon > 0$. Applying Hölder's inequality to the trace inequality (8.19) we get

$$\|v_k|_K\|_{L^2(\sigma)}^2 \leq C_{t, K, \sigma} ((h_K^{-1} + \frac{1}{2\epsilon}) \|v_k\|_{L^2(K)}^2 + \frac{\epsilon}{2} \|\nabla v_k\|_{L^2(K)^d}^2).$$

Hence, if there exists $D_{t, K, \sigma} > 0$ independent of v_k such that

$$\begin{aligned} C_{t, K, \sigma} ((h_K^{-1} + \frac{1}{2\epsilon}) \|v_k\|_{L^2(K)}^2 + \frac{\epsilon}{2} \|\nabla v_k\|_{L^2(K)^d}^2) \\ \leq D_{t, K, \sigma}^2 (c_{A, K} \|\nabla v_k\|_{L^2(K)^d}^2 + c_{\beta, \mu, K} \|v_k\|_{L^2(K)}^2) \end{aligned} \quad (8.22)$$

then $\|v_k|_K\|_{L^2(\sigma)}^2 \leq D_{t, K, \sigma}^2 \|v_k\|_K^2$ and the result holds. Relation (8.22) holds if

$$C_{t, K, \sigma} (h_K^{-1} + \frac{1}{2\epsilon}) \leq D_{t, K, \sigma}^2 c_{\beta, \mu, K}, \quad C_{t, K, \sigma} \frac{\epsilon}{2} \leq D_{t, K, \sigma}^2 c_{A, K}$$

and hence $D_{t, K, \sigma}^2 = \max\{C_{t, K, \sigma} (h_K^{-1} + \frac{1}{2\epsilon}) c_{\beta, \mu, K}^{-1}, C_{t, K, \sigma} \frac{\epsilon}{2} c_{A, K}^{-1}\}$. Taking ϵ such that the maximum is minimized we get $D_{t, K, \sigma}$ as in (8.21). ■

The proof of the following Lemma is inspired from [48, Theorem 3.1], the main difference is that we take into account the weaker regularity of the reconstructed fluxes.

Lemma 8.8. *Let $u \in H_0^1(\Omega)$ be the solution to (8.2), $u_k \in V(\mathfrak{T}_k)$ given by Algorithm 2, $s_k \in H_0^1(\Omega)$ from (8.8) and (8.16), $\mathbf{t}_k, \mathbf{q}_k \in H_{\text{div}}(\mathcal{G}_k)$ defined by (8.7) and (8.13) and $v \in H_0^1(\Omega)$. Then*

$$|\mathcal{B}(u - u_k, v) + \mathcal{B}_A(u_k - s_k, v)| \leq \left(\sum_{K \in \mathcal{M}_k} \eta_{1, K}^2 \right)^{1/2} \|v\|,$$

with $\eta_{1, K} = \eta_{R, K} + \eta_{DF, K} + \eta_{C, 1, K} + \eta_{C, 2, K} + \eta_{U, K} + \eta_{\Gamma, 1, K} + \eta_{\Gamma, 2, K}$.

Proof. Since u satisfies (8.2), using the definition of \mathcal{B} and \mathcal{B}_A

$$\begin{aligned} \mathcal{B}(u - u_k, v) + \mathcal{B}_A(u_k - s_k, v) &= \int_{\Omega} (f - (\mu - \nabla \cdot \beta) u_k) v \, dx - \int_{\Omega} A \nabla u_k \cdot \nabla v \, dx \\ &\quad - \int_{\Omega} \frac{1}{2} (\nabla \cdot \beta) (u_k - s_k) v \, dx - \int_{\Omega} \nabla \cdot (\beta s_k) v \, dx. \end{aligned}$$

Using $v\mathbf{t}_k \in H_{\text{div}}(\mathcal{G}_k)$, from the divergence theorem we have

$$\begin{aligned}\int_{\Omega} (v \nabla \cdot \mathbf{t}_k + \nabla v \cdot \mathbf{t}_k) d\mathbf{x} &= \sum_{G \in \mathcal{G}_k} \int_G \nabla \cdot (v \mathbf{t}_k) d\mathbf{x} = \sum_{G \in \mathcal{G}_k} \int_{\partial G} v \mathbf{t}_k \cdot \mathbf{n}_{\partial G} dy \\ &= \sum_{\gamma \in \Gamma_k} \int_{\gamma} [\![v \mathbf{t}_k]\!] \cdot \mathbf{n}_{\gamma} dy = \sum_{\gamma \in \Gamma_k} \int_{\gamma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\gamma} v dy\end{aligned}$$

and hence

$$\begin{aligned}\mathcal{B}(u - u_k, v) + \mathcal{B}_A(u_k - s_k, v) &= \int_{\Omega} (f - \nabla \cdot \mathbf{t}_k - \nabla \cdot \mathbf{q}_k - (\mu - \nabla \cdot \boldsymbol{\beta}) u_k) v d\mathbf{x} \\ &\quad - \int_{\Omega} \frac{1}{2} (\nabla \cdot \boldsymbol{\beta}) (u_k - s_k) v d\mathbf{x} + \int_{\Omega} \nabla \cdot (\mathbf{q}_k - \boldsymbol{\beta} s_k) v d\mathbf{x} \\ &\quad - \int_{\Omega} (A \nabla u_k + \mathbf{t}_k) \cdot \nabla v d\mathbf{x} + \sum_{\gamma \in \Gamma_k} \int_{\gamma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\gamma} v dy.\end{aligned}\quad (8.23)$$

From Lemma 8.6 we deduce

$$\begin{aligned}&\left| \int_{\Omega} (f - \nabla \cdot \mathbf{t}_k - \nabla \cdot \mathbf{q}_k - (\mu - \nabla \cdot \boldsymbol{\beta}) u_k) v d\mathbf{x} \right| \\ &= \left| \int_{\Omega} (f - \nabla \cdot \mathbf{t}_k - \nabla \cdot \mathbf{q}_k - (\mu - \nabla \cdot \boldsymbol{\beta}) u_k) (v - \pi_0 v) d\mathbf{x} \right| \\ &\leq \sum_{K \in \mathcal{M}_k} \eta_{R,K} \|v\|_K.\end{aligned}\quad (8.24a)$$

Similarly, we get

$$\begin{aligned}&\left| \int_{\Omega} (A \nabla u_k + \mathbf{t}_k) \cdot \nabla v d\mathbf{x} \right| \leq \sum_{K \in \mathcal{M}_k} \eta_{DF,K} \|v\|_K, \\ &\left| \int_{\Omega} \frac{1}{2} (\nabla \cdot \boldsymbol{\beta}) (u_k - s_k) v d\mathbf{x} \right| \leq \sum_{K \in \mathcal{M}_k} \eta_{C,2,K} \|v\|_K.\end{aligned}\quad (8.24b)$$

Since $[\![\mathbf{t}_k]\!]_{\sigma} = 0$ for $\sigma \in \mathcal{F}_{k,i} \setminus \cup_{\gamma \in \Gamma_k} \mathcal{F}_{\gamma}$, it holds

$$\sum_{\gamma \in \Gamma_k} \int_{\gamma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\gamma} v dy = \sum_{\sigma \in \mathcal{F}_{k,i}} \int_{\sigma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\sigma} v dy = \frac{1}{2} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \int_{\sigma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\sigma} v dy.$$

Using Lemma 8.7 we obtain

$$\begin{aligned}&\left| \sum_{\gamma \in \Gamma_k} \int_{\gamma} [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\gamma} v dy \right| \leq \frac{1}{2} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \|[\![\mathbf{t}_k]\!] \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \|v\|_{L^2(\sigma)} \\ &\leq \sum_{K \in \mathcal{M}_k} \eta_{\Gamma,2,K} \|v\|_K.\end{aligned}\quad (8.24c)$$

It remains to estimate $\int_{\Omega} \nabla \cdot (\mathbf{q}_k - \boldsymbol{\beta} s_k) v d\mathbf{x}$. For that, we use

$$\begin{aligned}\int_{\Omega} \nabla \cdot (\mathbf{q}_k - \boldsymbol{\beta} s_k) v d\mathbf{x} &= \sum_{K \in \mathcal{M}_k} \int_K (\mathcal{I} - \pi_0) \nabla \cdot (\mathbf{q}_k - \boldsymbol{\beta} s_k) (v - \pi_0 v) d\mathbf{x} \\ &\quad + \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} (\mathbf{q}_k - \boldsymbol{\beta} s_k) \cdot \mathbf{n}_K \pi_0 v dy\end{aligned}$$

and from [48] we get

$$\left| \sum_{K \in \mathcal{M}_k} \int_K (\mathcal{I} - \pi_0) \nabla \cdot (\mathbf{q}_k - \boldsymbol{\beta} s_k) (v - \pi_0 v) \, d\mathbf{x} \right| \leq \sum_{K \in \mathcal{M}_k} \eta_{C,1,K} \|v\|_K. \quad (8.24d)$$

For the second term we write

$$\begin{aligned} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} (\mathbf{q}_k - \boldsymbol{\beta} s_k) \cdot \mathbf{n}_K \pi_0 v \, d\mathbf{y} &= \sum_{\sigma \in \mathcal{F}_k} \int_{\sigma} [\pi_{0,\sigma}(\mathbf{q}_k - \boldsymbol{\beta} s_k) \pi_0 v] \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \\ &= \sum_{\sigma \in \mathcal{F}_{k,i}} \int_{\sigma} \{\pi_0 v\} [\pi_{0,\sigma}(\mathbf{q}_k - \boldsymbol{\beta} s_k)] \cdot \mathbf{n}_{\sigma} + [\pi_0 v] \{\pi_{0,\sigma}(\mathbf{q}_k - \boldsymbol{\beta} s_k)\} \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \\ &\quad + \sum_{\sigma \in \mathcal{F}_{k,b}} \int_{\sigma} \pi_0 v \pi_{0,\sigma}(\mathbf{q}_k - \boldsymbol{\beta} s_k) \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} = I_1 + I_2 + I_3 \end{aligned}$$

and we easily obtain, since $\|\boldsymbol{\beta} s_k\| = 0$,

$$I_1 = \frac{1}{2} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \int_{\sigma} \pi_0 v |_{K} [\pi_{0,\sigma} \mathbf{q}_k] \cdot \mathbf{n}_{\sigma} \, d\mathbf{y}.$$

Using $|\pi_0 v|_K = |K|^{-1/2} \|\pi_0 v\|_{L^2(K)} \leq |K|^{-1/2} \|v\|_{L^2(K)} \leq (|K| c_{\boldsymbol{\beta},\mu,K})^{-1/2} \|v\|_K$ we get

$$I_1 \leq \frac{1}{2} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} (|K| c_{\boldsymbol{\beta},\mu,K})^{-1/2} \|[\pi_{0,\sigma} \mathbf{q}_k] \cdot \mathbf{n}_{\sigma}\|_{L^1(\sigma)} \|v\|_K = \sum_{K \in \mathcal{M}_k} \eta_{\Gamma,1,K} \|v\|_K. \quad (8.24e)$$

Let $\mathcal{M}_{\sigma} = \{K \in \mathcal{M}_k : \sigma \subset \partial K\}$, using (8.18c) for the second term we have

$$\begin{aligned} I_2 &\leq \sum_{\sigma \in \mathcal{F}_{k,i}} m_{\sigma} \|\pi_{0,\sigma} \{\mathbf{q}_k - \boldsymbol{\beta} s_k\} \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \sum_{K \in \mathcal{M}_{\sigma}} \|v\|_K \\ &= \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} m_{\sigma} \|\pi_{0,\sigma} \{\mathbf{q}_k - \boldsymbol{\beta} s_k\} \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \|v\|_K. \end{aligned}$$

For the last term we similarly obtain

$$I_3 \leq \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,b}} m_{\sigma} \|\pi_{0,\sigma}(\mathbf{q}_k - \boldsymbol{\beta} s_k) \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \|v\|_K$$

and hence

$$I_2 + I_3 \leq \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \chi_{\sigma} m_{\sigma} \|\pi_{0,\sigma} \{\mathbf{q}_k - \boldsymbol{\beta} s_k\} \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \|v\|_K = \sum_{K \in \mathcal{M}_k} \eta_{U,K} \|v\|_K, \quad (8.24f)$$

where $\chi_{\sigma} = 2$ if $\sigma \in \mathcal{F}_{k,b}$ and $\chi_{\sigma} = 1$ if $\sigma \in \mathcal{F}_{k,i}$. Plugging relations (8.24a) to (8.24f) into (8.23) we get the result. \blacksquare

In Lemma 8.8 we use Lemma 8.6 to deduce that

$$\int_K (\nabla \cdot \mathbf{t}_k + \nabla \cdot \mathbf{q}_k + (\mu - \nabla \cdot \boldsymbol{\beta}) u_k) \, d\mathbf{x} = \int_K f \, d\mathbf{x} \quad (8.25)$$

and hence (8.24a). However, when the mesh has hanging nodes inside of the local domains Lemma 8.6 is not valid. Indeed, if $\widehat{\mathcal{M}}_k$ has hanging nodes, the fluxes $\hat{\mathbf{t}}_k, \hat{\mathbf{q}}_k$ must be constructed on a refined submesh $\overline{\mathcal{M}}_k$ of $\widehat{\mathcal{M}}_k$ free of hanging nodes, otherwise they may fail to be in $H_{\text{div}}(\Omega_k)$. The constructed fluxes will satisfy relation (8.15), but since $\nabla \cdot \hat{\mathbf{t}}_k, \nabla \cdot \hat{\mathbf{q}}_k \in \mathbb{P}_{\epsilon}(K')$ for $K' \in \overline{\mathcal{M}}_k$ and $\overline{\mathcal{M}}_k$ is finer than $\widehat{\mathcal{M}}_k$, then we cannot conclude as we did in Lemma 8.6. Nonetheless, (8.15) still implies (8.25), which is enough to prove Lemma 8.8.

8.3.3 Proof of the theorems

Here we prove Theorems 8.3 and 8.4. We will consider $\mathcal{B} : H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$ defined in (8.3) for functions in $H^1(\mathcal{M}_k)$.

Proof of Theorem 8.3. It has been proved in [47, Lemma 3.1] that for any $u_k \in V(\mathfrak{T}_k)$ and $u, s \in H_0^1(\Omega)$ it holds

$$\|u - u_k\| \leq \|u_k - s\| + |\mathcal{B}(u - u_k, v) + \mathcal{B}_A(u_k - s, v)|,$$

with $v = (u - s)/\|u - s\|$. Choosing u as the exact solution to (8.2), u_k given by Algorithm 2, $s = s_k$ from (8.8) and using Lemma 8.8 gives the result. ■

Proof of Theorem 8.4. Since $u \in H_0^1(\Omega)$ it holds $\mathcal{B}_J(u, w) = 0$ for all $w \in H_0^1(\Omega)$, using $\mathcal{B}_A \leq \mathcal{B} + |\mathcal{B}_S|$ we get

$$\|u - u_k\|_{\oplus} \leq 2\|u - u_k\| + \sup_{\substack{w \in H_0^1(\Omega) \\ \|w\|=1}} (\mathcal{B}(u - u_k, w) - \mathcal{B}_J(u_k, w)).$$

To conclude the proof we show that

$$\sup_{\substack{w \in H_0^1(\Omega) \\ \|w\|=1}} (\mathcal{B}(u - u_k, w) - \mathcal{B}_J(u_k, w)) \leq \left(\sum_{K \in \mathcal{M}_k} \eta_{2,K}^2 \right)^{1/2}.$$

Following Lemma 8.8, we easily get

$$\begin{aligned} \mathcal{B}(u - u_k, w) - \mathcal{B}_J(u_k, w) &\leq \sum_{K \in \mathcal{M}_k} (\eta_{R,K} + \eta_{DF,K} + \tilde{\eta}_{C,1,K} + \eta_{\Gamma,2,K}) \|w\|_K \\ &\quad + \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K} \int_{\sigma} \pi_0 w (\mathbf{q}_k - \boldsymbol{\beta} u_k) \cdot \mathbf{n}_K \, d\mathbf{y} - \mathcal{B}_J(u_k, w). \end{aligned}$$

The two last terms satisfy

$$\begin{aligned} &\sum_{\sigma \in \mathcal{F}_k} \int_{\sigma} [\pi_0 w (\mathbf{q}_k - \boldsymbol{\beta} u_k)] \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} - \mathcal{B}_J(u_k, w) \\ &= \sum_{\sigma \in \mathcal{F}_k} \chi_{\sigma} \int_{\sigma} [\pi_0 w] \pi_{0,\sigma} \{ \mathbf{q}_k - \boldsymbol{\beta} u_k \} \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} + \sum_{\sigma \in \mathcal{F}_{k,i}} \int_{\sigma} \{ \pi_0 w \} [\pi_{0,\sigma} \mathbf{q}_k] \cdot \mathbf{n}_{\sigma} \, d\mathbf{y} \\ &\leq \sum_{K \in \mathcal{M}_k} (\tilde{\eta}_{U,K} + \eta_{\Gamma,1,K}) \|w\|_K, \end{aligned}$$

where in the last step we followed again Lemma 8.8. ■

8.3.4 Alternative error bounds

Our aim here is to explain how to avoid the assumption $c_{\boldsymbol{\beta}, \mu, K} > 0$ for all $K \in \mathcal{M}_k$ made in Sections 8.2.1 and 8.3.2. This assumption is needed to define $\eta_{\Gamma,1,K}$, $\eta_{\Gamma,2,K}$ but can be avoided if (8.24c) and (8.24e) are estimated differently. For (8.24c), using the trace inequality (8.19) we get

$$\begin{aligned} \left| \sum_{\gamma \in \Gamma_k} \int_{\gamma} [\mathbf{t}_k] \cdot \mathbf{n}_{\gamma} v \, d\mathbf{y} \right| &\leq \frac{1}{2} \sum_{K \in \mathcal{M}_k} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \|[\mathbf{t}_k] \cdot \mathbf{n}_{\sigma}\|_{L^2(\sigma)} \|v|_K\|_{L^2(\sigma)} \\ &\leq \sum_{K \in \mathcal{M}_k} \tilde{\eta}_{\Gamma,2,K} (\|v\|_{L^2(K)}^2 + h_K \|v\|_{L^2(K)} \|\nabla v\|_{L^2(K)})^{1/2}, \end{aligned}$$

where

$$\tilde{\eta}_{\Gamma,2,K} = \frac{1}{2} \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} h_K^{-1/2} C_{t,K,\sigma}^{1/2} \|[\![\mathbf{t}_k]\!] \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)}.$$

Setting $\tilde{\eta}_{\Gamma,2}^2 = \sum_{K \in \mathcal{M}_k} \tilde{\eta}_{\Gamma,2,K}^2$, it yields

$$\begin{aligned} \left| \sum_{\gamma \in \Gamma_k} \int_\gamma [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_\gamma v \, dy \right| &\leq \tilde{\eta}_{\Gamma,2} \left(\sum_{K \in \mathcal{M}_k} \|v\|_{L^2(K)}^2 + h_K \|v\|_{L^2(K)} \|\nabla v\|_{L^2(K)^d} \right)^{1/2} \\ &\leq \tilde{\eta}_{\Gamma,2} \left(\|v\|_{L^2(\Omega)}^2 + h_{\mathcal{M}_k} \|v\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)^d} \right)^{1/2}. \end{aligned}$$

Using the Poincaré inequality $\|v\|_{L^2(\Omega)} \leq d_\Omega \|\nabla v\|_{L^2(\Omega)^d}$, where d_Ω is the diameter of Ω , we get

$$\left| \sum_{\gamma \in \Gamma_k} \int_\gamma [\![\mathbf{t}_k]\!] \cdot \mathbf{n}_\gamma v \, dy \right| \leq \tilde{\eta}_{\Gamma,2} (d_\Omega^2 + h_{\mathcal{M}_k} d_\Omega)^{1/2} \|\nabla v\|_{L^2(\Omega)^d} \leq \tilde{\eta}_{\Gamma,2} c_A^{-1/2} (d_\Omega^2 + h_{\mathcal{M}_k} d_\Omega)^{1/2} \|v\|,$$

where c_A is the minimal eigenvalue of $A(\mathbf{x})$ over Ω . The same procedure can be used to replace (8.24e) by a relation avoiding the term $c_{\beta,\mu,K}^{-1/2}$. The new bounds can be used to modify the results of Theorems 8.3 and 8.4 and obtain error estimators when $\mu - \frac{1}{2} \nabla \cdot \beta > 0$ is not satisfied.

8.4 Numerical experiments

In order to study the properties and illustrate the performance of the local scheme we consider here several numerical examples. First, we look at the convergence rates of the error estimators, focusing on the errors introduced by solving only local problems. Second, we compute the effectivity indices of the error estimators and investigate the efficiency of the new local algorithm. To do so, we compare the local scheme against a classical adaptive method, where after each mesh refinement the problem is solved again on the whole domain. The classical method we refer to is given by Algorithm 4.

Algorithm 4 ClassicalScheme(\mathfrak{T}_1)

```

Find  $\bar{u}_1 \in V(\mathfrak{T}_1)$  solution to  $\mathcal{B}(\bar{u}_1, v_1, \mathfrak{T}_1, 0) = (f, v_1)_1$  for all  $v_1 \in V(\mathfrak{T}_1)$ .
for  $k = 2, \dots, M$  do
     $(\mathfrak{T}_k, \widehat{\mathfrak{T}}_k) = \text{LocalDomain}(\bar{u}_{k-1}, \mathfrak{T}_{k-1})$ .
    Find  $\bar{u}_k \in V(\mathfrak{T}_k)$  solution to  $\mathcal{B}(\bar{u}_k, v_k, \mathfrak{T}_k, 0) = (f, v_k)_1$  for all  $v_k \in V(\mathfrak{T}_k)$ .
end for

```

In all the experiments we use \mathbb{P}_1 elements ($\ell = 1$ in (8.4)) on a simplicial mesh with penalization parameter $\eta_\sigma = 10$, the diffusive and convective fluxes $\mathbf{t}_k, \mathbf{q}_k$ are computed with $\mathbf{z} = 0$ (see (8.12)). Furthermore, β is always such that $\nabla \cdot \beta = 0$. These choices give $\eta_{C,1,K} = \eta_{C,2,K} = \tilde{\eta}_{C,1,K} = 0$. For an estimator $\eta_{*,K}$ we define $\eta_*^2 = \sum_{K \in \mathcal{M}_k} \eta_{*,K}^2$. Similarly to [48], if $A = \varepsilon I_2$ and β is constant then for $v_k \in H^1(\mathcal{M}_k)$ the augmented norm is well estimated by

$$\begin{aligned} \|v_k\|_{\oplus} \leq \|v_k\|_{\oplus'} &:= \|v_k\| + \varepsilon^{-1/2} \|\beta\|_2 \|v_k\|_{L^2(\Omega)} \\ &+ \frac{1}{2} \left(\sum_{K \in \mathcal{M}_k} \left(\sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{k,i}} \tilde{m}_K^{1/2} C_{t,K,\sigma}^{1/2} \|[\![v_k]\!] \beta \cdot \mathbf{n}_\sigma\|_{L^2(\sigma)} \right)^2 \right)^{1/2}. \end{aligned}$$

Hence, in the numerical experiments we consider the computable norm $\|\cdot\|_{\oplus'}$. The effectivity indices of the error estimators η and $\tilde{\eta}$ from Theorems 8.3 and 8.4 are defined as

$$\frac{\eta}{\|u - u_k\|} \quad \text{and} \quad \frac{\tilde{\eta}}{\|u - u_k\|_{\oplus'}}, \quad (8.26)$$

respectively. For the solution \bar{u}_k of the classical algorithm we use the error estimators η and $\tilde{\eta}$ from [48]. They are equivalent to the estimators presented in this paper except that for \bar{u}_k we have $\eta_{\Gamma,1,K} = \eta_{\Gamma,2,K} = 0$, as in this case the reconstructed fluxes are in $H_{\text{div}}(\Omega)$. The effectivity indices for \bar{u}_k are as in (8.26) but with u_k replaced by \bar{u}_k . The numerical experiments have been performed with the help of the C++ library `libMesh` [79].

8.4.1 Error estimators rate of convergence

In this first example, taken from [48], we solve (8.1) in $\Omega = [0, 1] \times [0, 1]$ with $A = \varepsilon I_2$, $\beta = (1, 0)^\top$ and $\mu = 1$. The force term f is chosen so that the exact solution reads

$$u(\mathbf{x}) = \frac{1}{2}x_1(x_1 - 1)x_2(x_2 - 1)(1 - \tanh(10 - 20x_1)), \quad (8.27)$$

see Figure 8.2(a). The purpose of the current experiment is to investigate the convergence rate of the error estimators for different values of $\varepsilon \in \{1, 10^{-2}, 10^{-4}\}$, i.e. for problems ranging from diffusion to advection dominated. In this example the local domains are fixed a priori, we define three domains $\Omega_1, \Omega_2, \Omega_3$ as follows: $\Omega_1 = \Omega$, $\mathbf{x} \in \Omega_2$ if $x_1 \in [0.25, 1]$ and $\mathbf{x} \in \Omega_3$ if $x_1 \in [0.375, 0.75]$, see Figure 8.2(b).

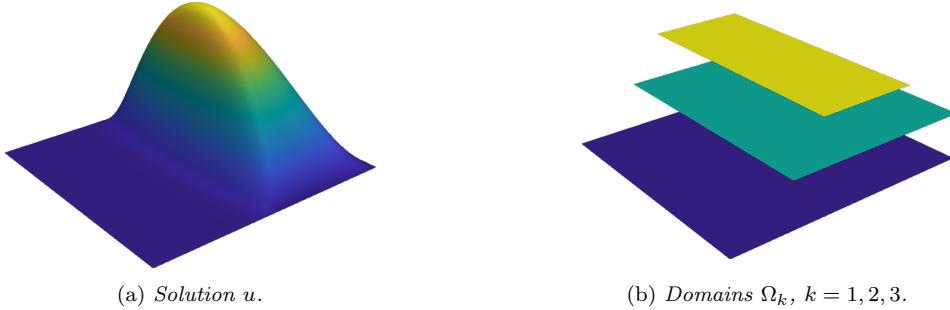


Figure 8.2. Error estimators rate of convergence. Solution $u(\mathbf{x})$ in (8.27) and local domains.

Let h be the grid size of $\widehat{\mathcal{M}}_1$, then the grid sizes of $\widehat{\mathcal{M}}_2$ and $\widehat{\mathcal{M}}_3$ are $h/2$ and $h/4$, respectively. For different choices of h we run Algorithm 2 without calling `LocalDomain`, since the local domains and meshes are chosen beforehand. After the third iteration we compute the exact energy error and the error estimators. The results are reported in Tables 8.1 to 8.3 for $\varepsilon = 1$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-4}$, respectively. We recall that η_{NC} measures the non conformity of u_k , η_R measures the error in the energy conservation, η_{DF} the difference between $-A\nabla u_k$ and the reconstructed diffusive flux \mathbf{t}_k , $\eta_U, \eta_{\bar{U}}$ are upwind errors and $\eta_{\Gamma,1}, \eta_{\Gamma,2}$ measure the jumps of $\mathbf{t}_k, \mathbf{q}_k$ across subdomain boundaries.

We see that the energy error converges with order one, as predicted by the a priori error analysis of [12]. On the other hand, the error estimators $\eta_{\Gamma,1}$ and $\eta_{\Gamma,2}$ measuring the reconstructed fluxes' jumps across subdomain boundaries have a rate of convergence of 0.5, of lower order than the other estimators and the true error. Hence, the local domains must be chosen so that the jumps

h	$\ \ u - u_k\ \ $	η_{NC}	η_R	η_{DF}	η_U	$\tilde{\eta}_U$	$\eta_{\Gamma,1}$	$\eta_{\Gamma,2}$
2^{-4}	1e-2	2.6e-3	1.9e-3	1.3e-2	3.9e-4	3.6e-4	5.8e-3	3.5e-2
2^{-5}	5e-3	1.3e-3	4.8e-4	6.7e-3	1.3e-4	1.3e-4	4.1e-3	2.4e-2
2^{-6}	2.5e-3	6.3e-4	1.2e-4	3.4e-3	4.6e-5	4.6e-5	2.9e-3	1.7e-2
2^{-7}	1.3e-3	3.1e-4	3e-5	1.7e-3	1.6e-5	1.6e-5	2.1e-3	1.2e-2
Order	1	1	2	1	1.5	1.5	0.5	0.5

 Table 8.1. Error estimators rate of convergence. Diffusion $\varepsilon = 1$.

h	$\ \ u - u_k\ \ $	η_{NC}	η_R	η_{DF}	η_U	$\tilde{\eta}_U$	$\eta_{\Gamma,1}$	$\eta_{\Gamma,2}$
2^{-4}	1e-3	2.7e-4	6.8e-4	1.4e-3	4e-3	3.6e-3	5.8e-3	4.8e-4
2^{-5}	5.1e-4	1.3e-4	1.7e-4	6.8e-4	1.3e-3	1.3e-3	4.1e-3	2.7e-4
2^{-6}	2.5e-4	6.2e-5	4.2e-5	3.4e-4	4.7e-4	4.6e-4	2.9e-3	1.8e-4
2^{-7}	1.3e-4	3.1e-5	1.1e-5	1.7e-4	1.6e-4	1.6e-4	2.1e-3	1.2e-4
Order	1	1	2	1	1.5	1.5	0.5	0.5

 Table 8.2. Error estimators rate of convergence. Diffusion $\varepsilon = 10^{-2}$.

h	$\ \ u - u_k\ \ $	η_{NC}	η_R	η_{DF}	η_U	$\tilde{\eta}_U$	$\eta_{\Gamma,1}$	$\eta_{\Gamma,2}$
2^{-4}	1.1e-4	7.1e-5	6.1e-3	2.5e-4	9.1e-3	7.4e-3	5.9e-3	1.9e-5
2^{-5}	5.2e-5	3.1e-5	1.6e-3	1.2e-4	5.6e-3	5e-3	4.1e-3	9.3e-6
2^{-6}	2.6e-5	1.4e-5	4.1e-4	5.6e-5	3.6e-3	3.5e-3	2.9e-3	5.6e-6
2^{-7}	1.3e-5	5.8e-6	1e-4	2.6e-5	1.7e-3	1.6e-3	2.1e-3	3.5e-6
Order	1	1	2	1	1.5	1.5	0.5	0.5

 Table 8.3. Error estimators rate of convergence. Diffusion $\varepsilon = 10^{-4}$.

at their interfaces are small and thus $\eta_{\Gamma,1}, \eta_{\Gamma,2}$ are negligible compared to the other estimators. This is guaranteed taking subdomains covering the large error regions.

8.4.2 Reaction dominated problem

In our next example we consider a symmetric problem and want to compare the local and classical schemes (Algorithms 2 and 4) in a singularly perturbed regime. We investigate the efficiency measured as the computational cost and analyze their effectivity indices. The setting is as follows: we solve (8.1) in $\Omega = [0, 1] \times [0, 1]$ with $\varepsilon = 10^{-6}$, $A = \varepsilon I_2$, $\beta = (0, 0)^\top$, $\mu = 1$ and we choose f such that the exact solution is given by

$$u(\mathbf{x}) = e^{x_1+x_2} \left(x_1 - \frac{1 - e^{-\zeta x_1}}{1 - e^{-\zeta}} \right) \left(x_2 - \frac{1 - e^{-\zeta x_2}}{1 - e^{-\zeta}} \right), \quad (8.28)$$

where $\zeta = 10^4$. The solution is illustrated in Figure 8.3(a).

Since the problem is symmetric we have $\|\cdot\| = \|\cdot\|_\oplus$, but their related error estimators η and $\tilde{\eta}$, respectively, satisfy $\tilde{\eta} > \eta$ and hence the effectivity index of η will be lower (see Theorems 8.3 and 8.4).

Starting from a coarse mesh (128 elements), we let the two algorithms run for $k = 1, \dots, 20$. In Figure 8.3(b) we show the first four subdomains Ω_k chosen by the local scheme. The first iterations are needed to capture the boundary layer and reach the convergence regime, hence we

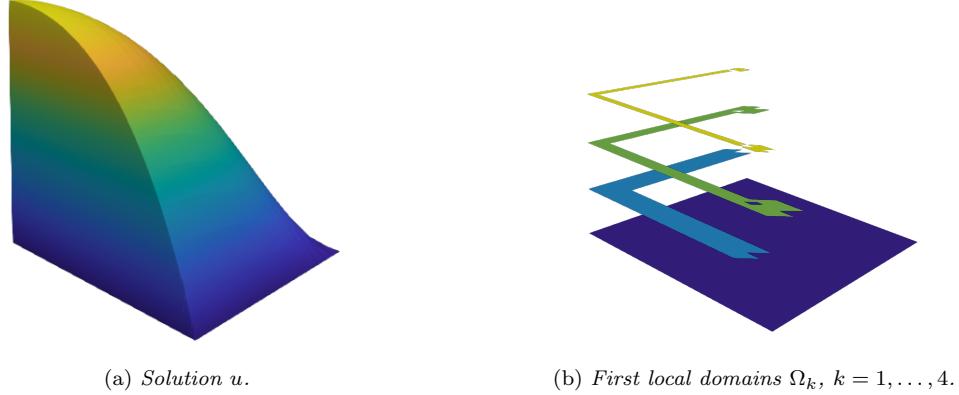


Figure 8.3. Reaction dominated problem. Solution $u(\mathbf{x})$ in (8.28) and first local domains chosen by the error estimators.

will display the results for $k \geq 7$. The most expensive part of the code is the solution of linear systems by means of the conjugate gradient (CG) method preconditioned with the incomplete Cholesky factorization, followed by the computation of the potential and flux reconstruction and then by the evaluation of the error estimators. In the local scheme, the time spent doing these tasks is proportional to the number of elements inside each subdomain Ω_k . For the classical scheme, the cost of these tasks depends on the total number of elements in the mesh. Since the CG routine is the most expensive part, we take the time spent in it as an indicator for the computational cost.

In Figure 8.4(a), we display the simulation cost against the error estimator η , for both the local and classical algorithms. Each circle or star in the figure represents an iteration k . We observe that the local scheme provides similar error bounds but at a smaller cost. The effectivity index of η at each iteration k is shown in Figure 8.4(b), we can observe that the local scheme has an effectivity index similar to the classical scheme.

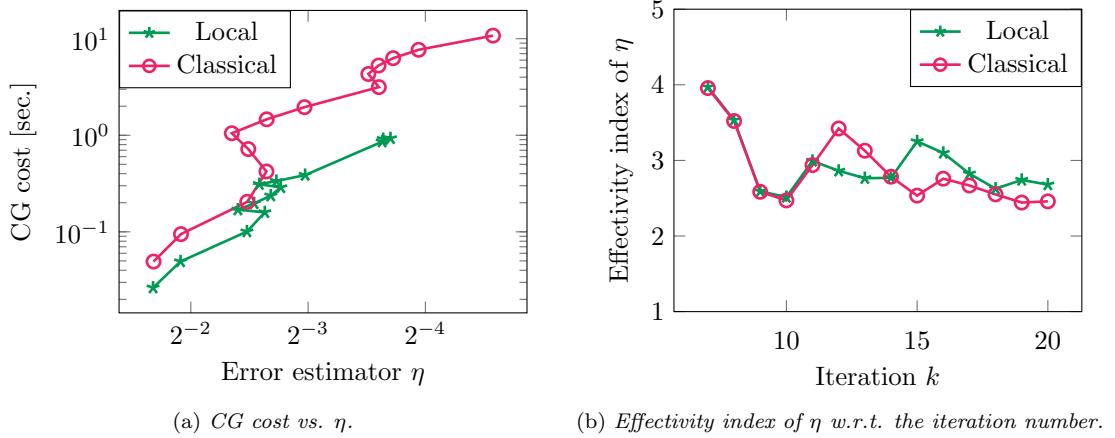


Figure 8.4. Reaction dominated problem. Computational cost vs. error estimator η and effectivity index.

In Figure 8.5(a) we exhibit the cost against the exact energy error and we notice that for some

values of k the mesh is refined but the error stays almost constant. This phenomenon significantly increases the simulation cost of the classical scheme without improving the solution. In contrast, the cost of the local scheme increases only marginally. Dividing the two curves in Figure 8.5(a) we obtain the relative speed-up, which is shown in Figure 8.5(b). We note that as the error decreases the local scheme becomes faster than the classical scheme. In Figure 8.6 we display the

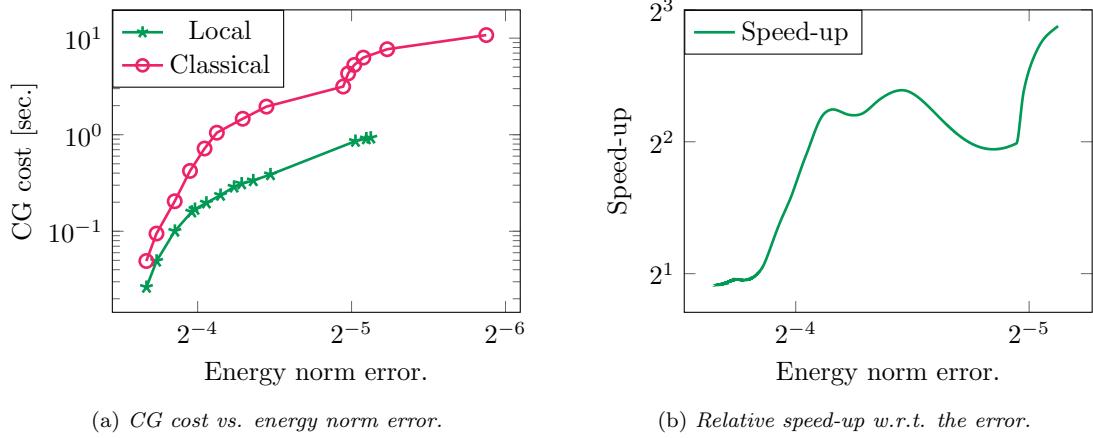


Figure 8.5. Reaction dominated problem. Computational cost vs. energy norm error and relative speed-up.

effectivity index of $\tilde{\eta}$. As expected, for this symmetric problem, it is worse than the effectivity of η .

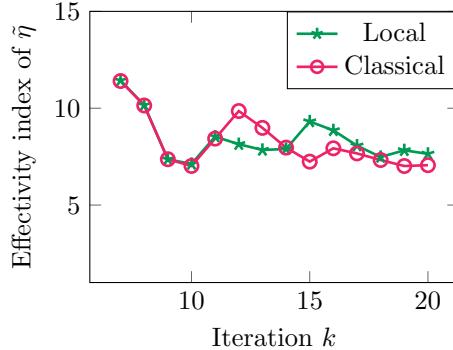


Figure 8.6. Reaction dominated problem. Effectivity index of $\tilde{\eta}$ w.r.t. the iteration number.

8.4.3 Convection dominated problem

In this section we perform the same experiment as in Section 8.4.2 but instead of choosing $\beta = (0, 0)^\top$ we set $\beta = -(1, 1)^\top$, hence we solve a nonsymmetric singularly perturbed problem. The linear systems are solved with the GMRES method preconditioned with the incomplete LU factorization. As in Section 8.4.2, we investigate the effectivity indices and efficiency of the local and classical schemes.

For convection dominated problems, the norm $\|\cdot\|_{\oplus}$ is more appropriate than $\|\cdot\|$ since it measures also the error in the advective direction. In Figure 8.7(a), we display the simulation

cost versus the error estimator $\tilde{\eta}$, we remark that again the local scheme provides similar error bounds at smaller cost. The effectivity index of $\tilde{\eta}$ is displayed in Figure 8.7(b), we note that the local and classical schemes have again similar effectivity indices.

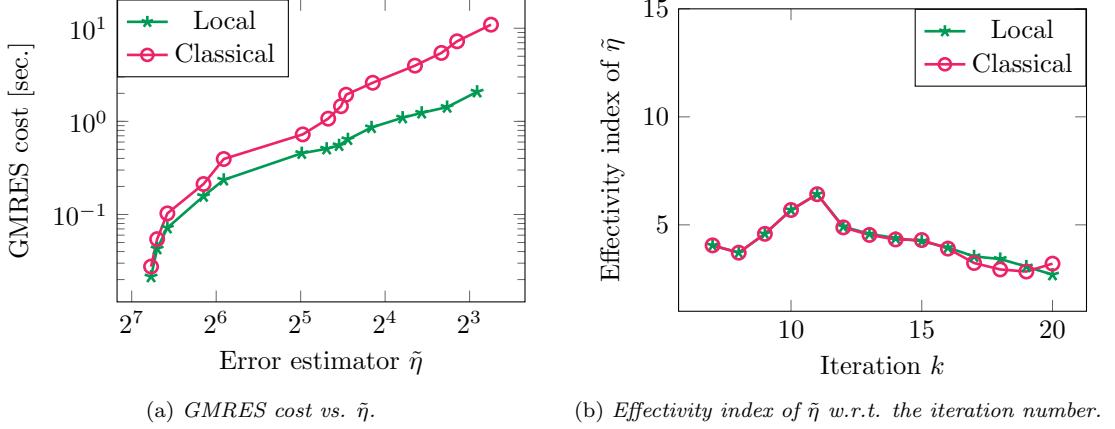


Figure 8.7. Convection dominated problem. Computational cost vs. $\tilde{\eta}$ and effectivity index.

In Figure 8.8 we show the simulation cost versus the error in the augmented norm $\|\cdot\|_{\oplus}$ and the relative speed-up. We again observe that the local scheme is faster.

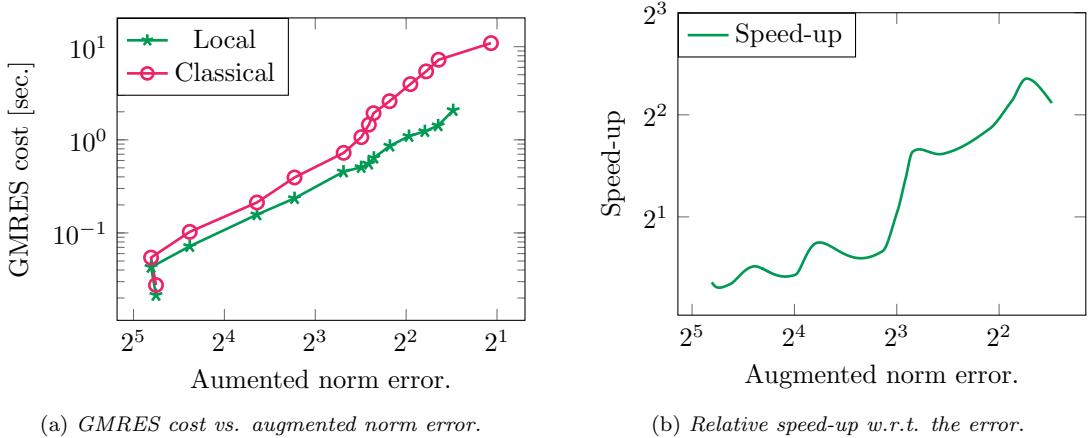
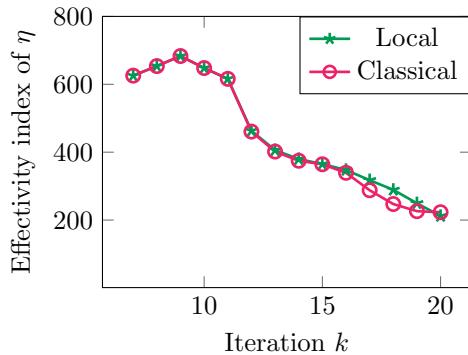
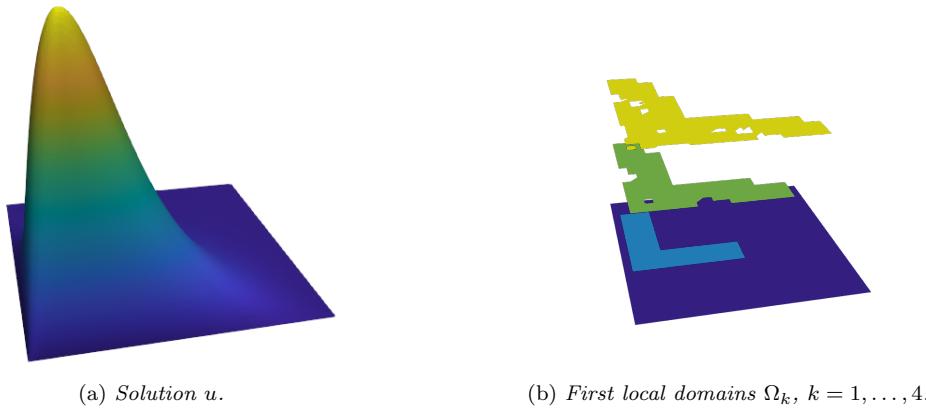


Figure 8.8. Convection dominated problem. Computational cost vs. augmented norm error and relative speed-up.

For completeness, we plot in Figure 8.9 the effectivity index of η . We see that it is completely off. This illustrates that this estimator does not capture the convective error and is hence not appropriate for convection dominated problems.

8.4.4 A smooth problem

In our last example, we want to apply the local scheme to a smooth problem. We solve (8.1) with $\Omega = [0, 1] \times [0, 1]$, $A = I_2$, $\beta = -(1, 1)^T$ and $\mu = 1$. The forcing term f is chosen such that the

Figure 8.9. Convection dominated problem. Effectivity index of η w.r.t. the iteration number.Figure 8.10. Smooth problem. Solution $u(\mathbf{x})$ in (8.29) and first local domains chosen by the error estimators.

exact solution is given by

$$u(\mathbf{x}) = e^{-\kappa \|\mathbf{x}\|_2} \left(x_1 - \frac{1 - e^{-\kappa x_1}}{1 - e^{-\kappa}} \right) \left(x_2 - \frac{1 - e^{-\kappa x_2}}{1 - e^{-\kappa}} \right) \quad (8.29)$$

with $\kappa = 10$. An illustration of the exact solution is given in Figure 8.10(a). We run the local and classical schemes for $k = 1, \dots, 15$ starting with a uniform mesh of 128 elements. The first four subdomains chosen by the local scheme are shown in Figure 8.10(b). For this problem, the error estimators $\eta_{\Gamma,1}$, $\eta_{\Gamma,2}$ measuring the reconstructed flux jumps dominate the other estimators and the effectivity index of the local scheme is larger than the index for the classical scheme (that approaches 1.5 for $k = 15$). However, the error estimators of the local scheme are still efficient in choosing the appropriate regions to be refined. In Figure 8.11(a) we show the computational cost in function of the energy errors. We observe that the local method achieves a similar accuracy at a smaller cost. In Figure 8.11(b) we highlight the relative speed-up of the local scheme and observe that it gets faster as the error decreases. We deduce that the local scheme can be employed also for smooth problems and if a tight estimation of the errors is needed then a full solve at the end of the iteration can be performed.

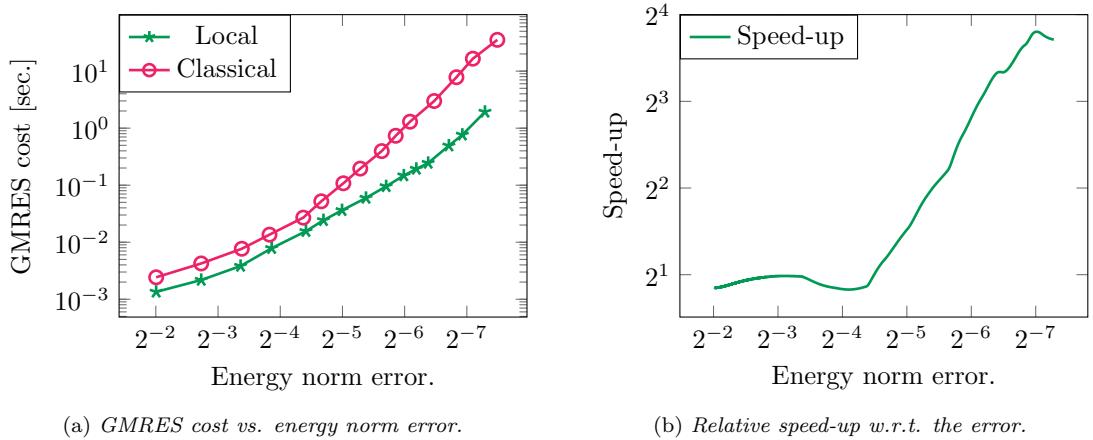


Figure 8.11. Smooth problem. Computational cost vs. error and relative speed-up.

9 Conclusion of Part II

In this part of the thesis, we introduced a local adaptive discontinuous Galerkin scheme for linear and quasilinear elliptic equations. The scheme relies on a coarse solution which is successively improved by solving a sequence of localized elliptic problems in confined subdomains, where the mesh is refined.

Chapter 6 introduces the foundations for proving convergence of the local scheme. We first define the gradient discretization method, a general framework for proving the convergence of discretization schemes for diffusive equations. Then we present and analyze the Weighted Discontinuous Galerkin Gradient Discretization (WDGGD), which is the space discretization method used in the local scheme. It is also shown that for pure diffusion problems the WDGGD scheme is equivalent to the Symmetric Weighted Interior Penalty (SWIP) method.

In Chapter 7 we define the local WDGGD scheme and perform the a priori error analysis. Thanks to the gradient discretization framework we show in Theorems 7.14 and 7.17 that, under minimal regularity assumptions, the local scheme converges to the exact solution of linear and quasilinear elliptic equations. Furthermore, thanks to pointwise estimates the error due to artificial boundary conditions is shown to be of higher order and shown to depend only locally on the regularity of the solution, see Theorem 7.12. Numerical experiments have shown the efficiency of the scheme when applied to equations with localized high gradient regions.

In Chapter 7, the location of high error regions is supposed to be known, which is usually not true in practice. Therefore, in Chapter 8 we provide a posteriori error estimators for the local method, which can be used to identify the subdomains to be refined. Starting from error estimators for the symmetric weighted interior penalty scheme based on conforming potential and flux reconstructions, we allow for flux jumps across the subdomains boundaries and derive new estimators for the local method in Theorems 8.3 and 8.4. Two important properties of the original estimators (for non local schemes) are conserved: the absence of unknown constants and the robustness in singularly perturbed regimes. Numerical experiments confirm the error estimators' robustness for convection-reaction dominated problems and illustrate the efficiency of the local scheme when compared to a classical adaptive algorithm, where at each iteration the solution on the whole computational domain must be recomputed.

Bibliography

- [1] A. ABDULLE, *Fourth order Chebyshev methods with recurrence relation*, SIAM Journal on Scientific Computing, 23 (2002), pp. 2041–2054.
- [2] A. ABDULLE, I. ALMUSLIMANI, AND G. VILMART, *Optimal explicit stabilized integrator of weak order one for stiff and ergodic stochastic differential equations*, Siam Journal on Uncertainty Quantification, 6 (2018), pp. 937–964.
- [3] A. ABDULLE AND S. CIRILLI, *Stabilized methods for stiff stochastic systems*, C. R. Math. Acad. Sci. Paris, 345 (2007), pp. 593–598.
- [4] ———, *S-ROCK: Chebyshev methods for stiff stochastic differential equations*, SIAM Journal on Scientific Computing, 30 (2008), pp. 997–1014.
- [5] A. ABDULLE, L. GANDER, AND G. ROSILHO DE SOUZA, *Optimal stabilized explicit integrators for stiff discrete noise stochastic differential equations*, Technical Report, EPFL, (2020).
- [6] A. ABDULLE, M. J. GROTE, AND G. ROSILHO DE SOUZA, *Stabilized explicit multirate methods for stiff stochastic differential equations*, Preprint submitted for publication, (2020), arXiv:2006.00744 [math.NA].
- [7] A. ABDULLE, Y. HU, AND T. LI, *Chebyshev Methods with Discrete Noise: the τ -ROCK Methods*, Journal of Computational Mathematics, 28 (2010), pp. 195–217.
- [8] A. ABDULLE AND T. LI, *S-ROCK methods for stiff Itô SDEs*, Communications in Mathematical Sciences, 6 (2008), pp. 845–868.
- [9] A. ABDULLE AND A. A. MEDOVIKOV, *Second order Chebyshev methods based on orthogonal polynomials*, Numerische Mathematik, 18 (2001), pp. 1–18.
- [10] A. ABDULLE AND G. A. PAVLIOTIS, *Numerical methods for stochastic partial differential equations with multiple scales*, Journal of Computational Physics, 231 (2012), pp. 2482–2497.
- [11] A. ABDULLE, G. A. PAVLIOTIS, AND U. VAES, *Spectral methods for multiscale stochastic differential equations*, SIAM-ASA Journal on Uncertainty Quantification, 5 (2017), pp. 720–761.
- [12] A. ABDULLE AND G. ROSILHO DE SOUZA, *A local discontinuous Galerkin gradient discretization method for linear and quasilinear elliptic equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 53 (2019), pp. 1269–1303.
- [13] ———, *A posteriori error analysis of a local adaptive discontinuous Galerkin method for convection-difusion-reaction equations*, Preprint submitted for publication, (2020), arXiv:2004.07148 [math.NA].

- [14] ——, *Instabilities and order reduction phenomenon of an interpolation based multirate Runge–Kutta–Chebyshev method*, Technical Report, EPFL, (2020), arXiv:2003.03154 [math.NA].
- [15] ——, *Stabilized explicit multirate methods for stiff stochastic differential equations*, Technical Report, EPFL, (2020).
- [16] A. ABDULLE, G. VILMART, AND K. C. ZYGALAKIS, *Weak second order explicit stabilized methods for stiff stochastic differential equations*, SIAM Journal on Scientific Computing, 35 (2013), pp. A1792–A1814.
- [17] M. AINSWORTH, *A synthesis of a posteriori error estimation techniques for conforming, non-conforming and discontinuous Galerkin finite element methods*, in Recent advances in adaptive computation, vol. 383 of Contemporary Mathematics, Amer. Math. Soc., Providence, RI, 2005, pp. 1–14.
- [18] M. AINSWORTH AND J. T. ODEN, *A unified approach to a posteriori error estimation using element residual methods*, Numerische Mathematik, 65 (1993), pp. 23–50.
- [19] ——, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics, John Wiley and Sons, New York, 2000.
- [20] J. F. ANDRUS, *Numerical solution of systems of ordinary differential equations into subsystems*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 605–611.
- [21] ——, *A Runge–Kutta Method with Stepsize Control for Separated Systems of First-Order ODEs*, Applied Mathematics and Computation, 59 (1993), pp. 193–214.
- [22] ——, *Stability of a multi-rate method for numerical integration of ODE’s*, Computers and Mathematics with Applications, 25 (1993), pp. 3–14.
- [23] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM Journal on Numerical Analysis, 19 (1982), pp. 742–760.
- [24] M. BAKKER, *Analytische aspekten van een minimaxprobleem*, Tech. Rep. TN 62/71, Stichting Mathematisch Centrum, Amsterdam, 1971.
- [25] W. BANGERTH AND R. RANNACHER, *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2003.
- [26] R. E. BANK AND A. WEISER, *Some a posteriori error estimators for elliptic partial differential equations*, Mathematics of Computation, 44 (1985), pp. 283–301.
- [27] L. BARBIÉ, I. RAMIÈRE, AND F. LEBON, *An automatic multilevel refinement technique based on nested local meshes for nonlinear mechanics*, Computers and Structures, 147 (2015), pp. 14–25.
- [28] M. J. BERGER AND J. OLIGER, *Adaptive mesh refinement for hyperbolic partial differential equations*, Journal of Computational Physics, 53 (1984), pp. 484–512.
- [29] A. BRANDT, *Multi-level adaptive solutions to boundary-value problems*, Mathematics of Computation, 31 (1977), pp. 333–390.
- [30] F. BREZZI AND M. FORTIN, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.
- [31] R. BUNDSCUH, F. HAYOT, AND C. JAYAPRAKASH, *The role of dimerization in noise reduction of simple genetic networks*, Journal of Theoretical Biology, 220 (2003), pp. 261–269.

- [32] Y. CAO, D. T. GILLESPIE, AND L. R. PETZOLD, *Adaptive explicit-implicit tau-leaping method with automatic tau selection*, Journal of Chemical Physics, 126 (2007).
- [33] Y. CAO, H. LI, AND L. R. PETZOLD, *Efficient formulation of the stochastic simulation algorithm for chemically reacting systems*, Journal of Chemical Physics, 121 (2004), pp. 4059–4067.
- [34] Y. CAO, L. R. PETZOLD, M. RATHINAM, AND D. T. GILLESPIE, *The numerical stability of leaping methods for stochastic simulation of chemically reacting systems*, Journal of Chemical Physics, 121 (2004), pp. 12169–12178.
- [35] I. CHEDDADI, R. FUČÍK, M. I. PRIETO, AND M. VOHRALÍK, *Guaranteed and robust a posteriori error estimates for singularly perturbed reaction-diffusion problems*, ESAIM: Mathematical Modelling and Numerical Analysis, 43 (2009), pp. 867–888.
- [36] Z. CHEN AND H. CHEN, *Pointwise error estimates of discontinuous Galerkin methods with penalty for second-order elliptic problems*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 1146–1166.
- [37] S. COCHEZ-DHONDT AND S. NICHAISE, *Equilibrated error estimators for discontinuous Galerkin methods*, Numerical Methods for Partial Differential Equations, 24 (2008), pp. 1236–1252.
- [38] C. N. DAWSON, D. QIANG, AND T. F. DUPONT, *A finite difference domain decomposition algorithm for numerical solution of the heat equation*, Mathematics of computation, 57 (1991), pp. 63–71.
- [39] D. A. DI PIETRO AND A. ERN, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69 of Mathématiques et Applications, Springer, Berlin and Heidelberg, 2012.
- [40] W. DÖRFLER, *A convergent adaptive algorithm for Poisson’s equation*, SIAM Journal on Numerical Analysis, 33 (1996), pp. 1106–1124.
- [41] J. DRONIU, R. EYMARD, T. GALLOUËT, C. GUICHARD, AND R. HERBIN, *The Gradient Discretisation Method*, vol. 82, Springer International Publishing, 2018.
- [42] W. E, *Analysis of the heterogeneous multiscale method for ordinary differential equations*, Communications in Mathematical Sciences, 1 (2003), pp. 423–436.
- [43] W. E, D. LIU, AND E. VANDEN-EIJNDEN, *Analysis of multiscale methods for stochastic differential equations*, Communications on Pure and Applied Mathematics, 58 (2005), pp. 1544–1585.
- [44] C. ENGSTLER AND C. LUBICH, *Multirate extrapolation methods for differential equations with different time scales*, Computing, 58 (1997), pp. 173–185.
- [45] W. H. ENRIGHT, T. E. HULL, AND B. LINDBERG, *Comparing numerical methods for stiff systems of O.D.E:s*, BIT Numerical Mathematics, 15 (1975), pp. 10–48.
- [46] A. ERN, S. NICHAISE, AND M. VOHRALÍK, *An accurate $H(\text{div})$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems*, Comptes Rendus Mathematique, 345 (2007), pp. 709–712.
- [47] A. ERN AND A. F. STEPHANSEN, *A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods*, Journal of Computational Mathematics, 26 (2008), pp. 488–510.

- [48] A. ERN, A. F. STEPHANSEN, AND M. VOHRALÍK, *Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems*, Journal of Computational and Applied Mathematics, 234 (2010), pp. 114–130.
- [49] A. ERN, A. F. STEPHANSEN, AND P. ZUNINO, *A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity*, IMA Journal of Numerical Analysis, 29 (2009), pp. 235–256.
- [50] R. E. EWING, R. D. LAZAROV, AND A. VASSILEV, *Finite difference scheme for parabolic problems on composite grids with refinement in time and space*, SIAM Journal on Numerical Analysis, 31 (1994), pp. 1605–1622.
- [51] R. E. EWING, R. D. LAZAROV, AND P. S. VASSILEVSKI, *Finite difference schemes on grids with local refinement in time and space for parabolic problems I. Derivation, stability, and error analysis*, Computing, 45 (1990), pp. 193–215.
- [52] R. EYMARD AND C. GUICHARD, *Discontinuous Galerkin gradient discretisations for the approximation of second-order differential operators in divergence form*, Computational and Applied Mathematics, 37 (2017), pp. 4023–4054.
- [53] P. FERKET, *Solving boundary value problems on composite grids with an application to combustion*, PhD thesis, Eindhoven: Technische Universiteit Eindhoven, 1996.
- [54] L. A. FREITAG AND C. OLLIVIER-GOOCH, *A Cost/Benefit Analysis of Simplicial Mesh Improvement Techniques as Measured by Solution Efficiency*, International Journal of Computational Geometry and Applications, 10 (2000), pp. 361–382.
- [55] L. GANDER, *Optimized Chebyshev methods for discrete stochastic simulations*, master thesis, EPFL, 2019.
- [56] M. J. GANDER AND L. HALPERN, *Techniques for locally adaptive time stepping developed over the last two decades*, Lecture Notes in Computational Science and Engineering, 91 (2013), pp. 377–385.
- [57] C. W. GEAR, G. IOANNIS, AND G. KEVREKIDIS, *Projective methods for stiff differential equations: problems with gaps in their eigenvalue spectrum*, SIAM Journal on Scientific Computing, 24 (2003), pp. 1091–1106.
- [58] C. W. GEAR AND D. R. WELLS, *Multirate linear multistep methods*, BIT Numerical Mathematics, 24 (1984), pp. 484–502.
- [59] D. T. GILLESPIE, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, Journal of Computational Physics, 22 (1976), pp. 403–434.
- [60] D. T. GILLESPIE, *Exact stochastic simulation of coupled chemical reactions*, Journal of Physical Chemistry, 81 (1977), pp. 2340–2361.
- [61] D. T. GILLESPIE, *Approximate accelerated stochastic simulation of chemically reacting systems*, Journal of Chemical Physics, 115 (2001), pp. 1716–1733.
- [62] D. GIVON, I. G. KEVREKIDIS, AND R. KUPFERMAN, *Strong convergence of projective integration schemes for singularly perturbed stochastic differential systems*, Communications in Mathematical Sciences, 4 (2006), pp. 707–729.
- [63] A. GUILLOU AND B. LAGO, *Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles, à pas séparés et à pas liés. Recherche de formules à grand rayon de stabilité.*, in 1er Congr. Ass. Fran. Calcul., AFCAL, Grenoble, 1960, pp. 43–56.

-
- [64] M. GÜNTHER, A. KVÆRNØ, AND P. RENTROP, *Multirate partitioned Runge–Kutta methods*, BIT Numerical Mathematics, 41 (2001), pp. 504–514.
- [65] M. GÜNTHER AND P. RENTROP, *Multirate ROW methods and latency of electric circuits*, Applied Numerical Mathematics, 13 (1993), pp. 83–102.
- [66] K. GUSTAFSSON, *Control Theoretic Techniques for Time Step Selection in Implicit Runge–Kutta Methods*, ACM Trans. Math. Softw., 20 (1994), pp. 496–517.
- [67] W. HACKBUSCH, *Local defect correction method and domain decomposition techniques*, in Defect Correction Methods, K. Böhmer and H. Stetter, eds., Computing Supplementa, Springer, Wien, 1984, pp. 89–113.
- [68] E. HAIRER, S. P. NÖRSETT, AND G. WANNER, *Solving ordinary differential equations I*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2008.
- [69] E. HAIRER AND G. WANNER, *Solving ordinary differential equations II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2002.
- [70] D. J. HIGHAM., *Modeling and simulating chemical reactions*, Siam Review, 50 (2008), pp. 347–368.
- [71] N. J. HIGHAM AND A. H. AL-MOHY, *Computing matrix functions*, Acta Numerica, 19 (2010), pp. 159–208.
- [72] M. HOCHBRUCK AND A. OSTERMANN, *Exponential integrators*, Acta Numerica, 19 (2010), pp. 209–286.
- [73] E. HOFER, *A partially implicit method for large stiff systems of ODEs with only few equations introducing small time-constants*, SIAM Journal on Numerical Analysis, 13 (1976), pp. 645–663.
- [74] Y. HU, A. ABDULLE, AND T. LI, *Boosted hybrid method for solving chemical reaction systems with multiple scales in time and population size*, Communications in Computational Physics, 12 (2012), pp. 981–1005.
- [75] O. A. KARAKASHIAN AND F. PASCAL, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM Journal on Numerical Analysis, 41 (2003), pp. 2374–2399.
- [76] R. B. KELLOG, *On the poisson equation with intersecting interfaces*, Applicable Analysis: An International Journal, 4 (1975), pp. 101–129.
- [77] I. G. KEVREKIDIS AND A. PAPAVASILIOU, *Variance reduction for the equation-free simulation of multiscale stochastic systems*, Multiscale Modeling and Simulation, 6 (2007), pp. 70–89.
- [78] K. Y. KIM, *A posteriori error estimators for locally conservative methods of nonlinear elliptic problems*, Applied Numerical Mathematics, 57 (2007), pp. 1065–1080.
- [79] B. S. KIRK, J. W. PETERSON, R. H. STOGNER, AND G. F. CAREY, *libMesh : a C++ library for parallel adaptive mesh refinement/coarsening simulations*, Engineering with Computers, 22 (2006), pp. 237–254.
- [80] H. KURATA, H. EL-SAMAD, T. YI, M. KHAMMASH, AND J. DOYLE, *Feedback regulation of the heat shock response in *E. coli**, in Proceedings of the 40th IEEE Conference on Decision and Control, vol. 1, 2001, pp. 837–842.

- [81] A. KVÆRNØ, *Stability of multirate Runge–Kutta schemes*, in Proc. of the 10th Coll. on Differential Equations, vol. 1A, 1999, pp. 97–105.
- [82] V. I. LEBEDEV, *Explicit difference schemes with time-variable steps for solving stiff systems of equations*, Sov. J. Numer. Anal. Math. Modelling, 4 (1989), pp. 111–135.
- [83] ———, *How to solve stiff systems of differential equations by explicit methods*, in Numerical methods and applications, CRC, Boca Raton, FL, 1994, pp. 45–80.
- [84] V. I. LEBEDEV AND A. A. MEDOVIKOV, *Explicit methods of second order for the solution of stiff systems of ODEs*, Russian Academy of Science, (1994).
- [85] T. LI, A. ABDULLE, AND W. E, *Effectiveness of implicit methods for stiff stochastic differential equations*, Communications in Computational Physics, 3 (2008), pp. 295–307.
- [86] B. LINDBERG, *IMPEX: a program package for solution of systems of stiff differential equations*, tech. rep., Dept. of Information Processing, Royal Inst. of Tech., Stockholm, 1972.
- [87] D. LIU, *Analysis of multiscale methods for stochastic dynamical systems with multiple time scales*, Multiscale Modeling and Simulation, 8 (2010), pp. 944–964.
- [88] J. LIU, L. MU, X. YE, AND R. JARI, *Convergence of the discontinuous finite volume method for elliptic problems with minimal regularity*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 4537–4546.
- [89] A. MARKOFF, *On a certain problem of D. I. Mendeleieff*, Utcheniya Zapiski Imperatorskoi Akademii Nauk (Russia), 62 (1889), pp. 1–24.
- [90] S. MCCORMICK AND J. THOMAS, *The Fast Adaptive Composite grid (FAC) method for elliptic equations*, Mathematics of Computation, 46 (1986), pp. 439–456.
- [91] S. MCCORMICK AND R. ULRICH, *A finite volume convergence theory for the fast adaptive composite grid methods*, Applied Numerical Mathematics, 14 (1994), pp. 91–103.
- [92] A. A. MEDOVIKOV, *High order explicit methods for parabolic equations*, BIT Numerical Mathematics, 38 (1998), pp. 372–390.
- [93] G. MILSHTEIN AND M. TRETYAKOV, *Stochastic Numerics for Mathematical Physics*, Springer, 2003.
- [94] R. MINERO, M. J. H. ANTHONISSEN, AND R. M. M. MATTHEIJ, *A local defect correction technique for time-dependent problems*, Numerical Methods for Partial Differential Equations, 22 (2006), pp. 128–144.
- [95] T. MIRZAKHANIAN, *Multi-rate Runge–Kutta–Chebyshev time stepping for parabolic equations on adaptively refined meshes*, master thesis, Boise State University, 2017. doi:10.18122/B2V715.
- [96] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Convergence of Adaptive Finite Element Methods*, SIAM Review, 44 (2002), pp. 631–658.
- [97] J. NÉDÉLEC, *Mixed finite elements in \mathbb{R}^3* , Numerische Mathematik, 35 (1980), pp. 315–341.
- [98] L. PAYNE AND H. WEINBERGER, *An optimal Poincaré inequality for convex domains*, Archive for Rational Mechanics and Analysis, 5 (1960), pp. 286–292.
- [99] M. RATHINAM, L. R. PETZOLD, Y. CAO, AND D. T. GILLESPIE, *Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method*, Journal of Chemical Physics, 119 (2003), pp. 12784–12794.

- [100] P. RAVIART AND J. THOMAS, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), E. Magenes and I. Galligani, eds., vol. 606 of Lecture Notes in Mathematics, New York, 1977, Springer-Verlag., pp. 292–315.
- [101] S. REPIN, *A Posteriori Estimates for Partial Differential Equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH and Co, Berlin, 2008.
- [102] J. R. RICE, *Split Runge–Kutta method for simultaneous equations*, Journal of research, National Bureau of Standards. Section B, Mathematics and mathematical physics, 64B (1960), pp. 151–170.
- [103] W. RIHA, *Optimal stability polynomials*, Computing, 9 (1972), pp. 37–43.
- [104] V. SAUL’EV, *Integration of parabolic type equations with the method of nets*, Pergamon Press, 1964.
- [105] V. SAVCENCO, W. HUNSDORFER, AND J. VERWER, *A multirate time stepping strategy for stiff ordinary differential equations*, BIT Numerical Mathematics, 47 (2007), pp. 137–155.
- [106] V. SAVCENCO AND R. M. M. MATTHEIJ, *Multirate numerical integration for stiff ODEs*, in Progress in industrial mathematics at ECMI 2008, vol. 15, Springer, Heidelberg, 2010, pp. 327–332.
- [107] G. I. SHISHKIN AND P. N. VABISHCHEVICH, *Interpolation finite difference schemes on grids locally refined in time*, Computer Methods in Applied Mechanics and Engineering, 190 (2000), pp. 889–901.
- [108] S. SKELBOE AND P. U. ANDERSEN, *Stability properties of backward Euler multirate formulas*, SIAM Journal on Scientific and Statistical Computing, 10 (1989), pp. 1000–1009.
- [109] G. SÖDERLIND, *Automatic control and adaptive time-stepping*, Numerical Algorithms, 31 (2002), pp. 281–310.
- [110] B. P. SOMMEIJER, L. SHAMPINE, AND J. G. VERWER, *RKC: An explicit solver for parabolic PDEs*, Journal of Computational and Applied Mathematics, 88 (1998), pp. 315–326.
- [111] B. P. SOMMEIJER AND J. G. VERWER, *Performance evaluation of a class of Runge–Kutta–Chebyshev methods for solving semi-discrete parabolic differential equations*, Tech. Rep. NW91/80, Stichting Mathematisch Centrum, Amsterdam, 1980.
- [112] A. F. STEPHANSEN, *Méthodes de Galerkine discontinues et analyse d’erreur a posteriori pour les problèmes de diffusion hétérogène*, PhD thesis, Ecole Nationale des Ponts et Chaussées, 2007.
- [113] R. TROMPERT AND J. VERWER, *A static-regridding method for two-dimensional parabolic partial differential equations*, Applied Numerical Mathematics, 8 (1991), pp. 65–90.
- [114] ———, *Analysis of local uniform grid refinement*, Applied Numerical Mathematics, 13 (1993), pp. 251–270.
- [115] ———, *Analysis of the implicit Euler local uniform grid refinement method*, SIAM Journal on Scientific Computing, 14 (1993), pp. 259–278.
- [116] ———, *Runge–Kutta methods and local uniform grid refinement*, Mathematics of Computation, 60 (1993), pp. 591–616.

- [117] P. J. VAN DER HOUWEN, *Construction of integration formulas for initial value problems*, vol. 19, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1977.
- [118] P. J. VAN DER HOUWEN AND B. P. SOMMEIJER, *On the internal stability of explicit, m -stage Runge–Kutta methods for large m -values*, Z. Angew. Math. Mech., 60 (1980), pp. 479–485.
- [119] M. T. VAN GENUCHTEN, *A closed-form equation for Predicting the Hydraulic Conductivity of Unsaturated Soils*, Soil Science Society of America, 44 (1980), pp. 892–898.
- [120] E. VANDEN-EIJNDEN, *Numerical techniques for multi-scale dynamical systems with stochastic effects*, Communications in Mathematical Sciences, 1 (2003), pp. 385–391.
- [121] ———, *On HMM-like integrators and projective integration methods for systems with multiple time scales*, Communications in Mathematical Sciences, 5 (2007), pp. 495–505.
- [122] R. VERFÜRTH, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Wiley-Teubner, New York, 1996.
- [123] J. G. VERWER, *An implementation of a class of stabilized explicit methods for the time integration of parabolic equations*, ACM Transactions on Mathematical Software (TOMS), 6 (1980), pp. 188–205.
- [124] ———, *Explicit Runge–Kutta methods for parabolic partial differential equations*, Applied Numerical Mathematics, 22 (1996), pp. 359–379.
- [125] J. G. VERWER, W. HUNDSDORFER, AND B. P. SOMMEIJER, *Convergence properties of the Runge–Kutta–Chebyshev method*, Numerische Mathematik, 57 (1990), pp. 157–178.
- [126] M. VOHRALÍK, *Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods*, Numerische Mathematik, 111 (2008), pp. 121–158.
- [127] J. WAPPLER, *Die lokale Defektkorrekturmethode zur adaptiven Diskretisierung el- liptischer Differentialgleichungen mit finiten Elementen.*, PhD thesis, Christian-Albrechts-Universität, 1999.
- [128] Y. YANG, M. RATHINAM, AND J. SHEN, *Integral tau methods for stiff stochastic chemical systems*, Journal of Chemical Physics, 134 (2011), p. 044129.

Curriculum Vitae

Personal data

Name	Giacomo Rosilho de Souza
Date of birth	29 December 1989
Nationality	Swiss and Brazilian

Education

2015 – 2020	PhD in Mathematics École Polytechnique Fédérale de Lausanne, Switzerland. Thesis advisor: Prof. A. Abdulle.
2012 – 2014	Master of Science in Mathematics École Polytechnique Fédérale de Lausanne, Switzerland. Thesis advisor: Prof. A. Abdulle.
2009 – 2012	Bachelor of Science in Mathematics École Polytechnique Fédérale de Lausanne, Switzerland.

PhD publications

- [1] A. ABDULLE AND G. ROSILHO DE SOUZA, *A local discontinuous Galerkin gradient discretization method for linear and quasilinear elliptic equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 53 (2019), pp. 1269–1303.
- [2] ——, *A posteriori error analysis of a local adaptive discontinuous Galerkin method for convection-difusion-reaction equations*, Preprint submitted for publication, (2020), arXiv:2004.07148 [math.NA].
- [3] A. ABDULLE, M. J. GROTE, AND G. ROSILHO DE SOUZA, *Stabilized explicit multirate methods for stiff stochastic differential equations*, Preprint submitted for publication, (2020), arXiv:2006.00744 [math.NA].
- [4] A. ABDULLE AND G. ROSILHO DE SOUZA, *Instabilities and order reduction phenomenon of an interpolation based multirate Runge–Kutta–Chebyshev method*, Technical Report, EPFL, (2020), arXiv:2003.03154 [math.NA].

In preparation

- [5] A. ABDULLE AND G. ROSILHO DE SOUZA, *Stabilized explicit multirate methods for stiff stochastic differential equations*, Technical Report, EPFL, (2020).
- [6] A. ABDULLE, L. GANDER, AND G. ROSILHO DE SOUZA, *Optimal stabilized explicit integrators for stiff discrete noise stochastic differential equations*, Technical Report, EPFL, (2020).

Awards

- JOHN BUTCHER PRIZE IN NUMERICAL ANALYSIS (2019).

For the talk: *Multirate explicit stabilized integrators for stiff differential equations*, given at the SciCADE conference in Innsbruck, Austria, 26th July 2019.

The award is for the best student talk at the biennial International Conference on Scientific Computation and Differential Equations (SciCADE). In particular, considering both the academic merit of the content and the presentation itself. The John Butcher Prize in Numerical Analysis and the associated funds are administered by the New Zealand branch of ANZIAM.

Presentations

- MATHICSE RETREAT (Leysin, Switzerland, 27–29 June 2016);
Talk: *Two local time stepping techniques for parabolic equations*.
- MATHICSE RETREAT (Leysin, Switzerland, 14–16 June 2017);
Talk: *A predictor corrector local time stepping scheme for parabolic equations*.
- SWISS NUMERICS DAY (Zürich, Switzerland, 20 April 2018);
Talk: *A priori and a posteriori analysis of a local scheme for elliptic equations*.
- MATHICSE RETREAT (Sainte-Croix, Switzerland, 19–21 June 2018);
Talk: *A local discontinuous Galerkin FEM for linear and quasilinear elliptic equations*.
- MATHICSE RETREAT (Champéry, Switzerland, 11–13 June 2019);
Talk: *Multirate explicit stabilized integrators for stiff differential equations*.
- SCICADE , INTERNATIONAL CONFERENCE ON SCIENTIFIC COMPUTATION AND DIFFERENTIAL EQUATIONS (Innsbruck, Austria, 22–26 July 2019);
Talk: *Multirate explicit stabilized integrators for stiff differential equations*.
- INTERNAL SEMINAR (Basel, Switzerland, 1 November 2019);
Talk: *Multirate explicit stabilized integrators for stiff differential equations*.