Explicit Stabilized Methods for Stiff Stochastic Differential Equations and Stiff Optimal Control Problems

Ibrahim Almuslimani

UNIVERSITÉ DE GENÈVE
Section de mathématiques

FACULTÉ DES SCIENCES
Dr. Gilles Vilmart

Explicit Stabilized Methods for Stiff Stochastic Differential Equations and Stiff Optimal Control Problems

THÈSE

Présentée à la Faculté des Sciences de l'Université de Genève
pour obtenir le grade de Docteur ès Sciences, mention Mathématiques

par

Ibrahim ALMUSLIMANI

de

Halba (Liban)

Thèse N° 5511

GENÈVE
Atelier d'impression ReproMail
2020

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION MATHEMATIQUES

Thèse de Monsieur Ibrahim ALMUSLIMANI

intitulée :

«Explicit Stabilized Methods for Stiff Stochastic Differential Equations and Stiff Optimal Control Problems»

La Faculté des sciences, sur le préavis de Monsieur G. VILMART, docteur et directeur de thèse (Section de mathématiques), Monsieur M. GANDER, professeur ordinaire (Section de mathématiques), Monsieur P. CHARTIER, professeur (INRIA Rennes, Campus de Beaulieu, Rennes, France), Monsieur K. ZYGALAKIS, docteur (School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 3 novembre 2020

Thèse - 5511 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

à mes parents

à Joanna

à Aysha

Ibrahim

Acknowledgment

First of all, I would like to express my sincere gratitude to my Ph.D. supervisor, Dr. Gilles Vilmart, whose insight and knowledge into the subject matter steered me throughout this research. I would like to thank him as well for his kindness, patience, help, motivation, support, as well as his enthusiasm for the project. He always believed in my abilities and encouraged me to achieve more and more. It would not have been possible to complete this work without his guidance.

Besides my advisor, I would like to thank Professors Philippe Chartier, Martin Gander, and Konstantinos Zygalakis for generously offering their time to be the jury of my Ph.D. defense, and for their valuable comments. I thank as well Professor Assyr Abdulle for his collaboration and interesting discussions.

I would also like to extend my sincere thanks to my colleagues and friends at the Department of Mathematics. Many thanks go to my academic brothers Adrien, Guillaume, and Nicolas. Special thanks go to my office mates, Eiichi and Pratik for the interesting discussions, suggestions, and the nice atmosphere in the office and during coffee breaks. Many thanks also go to all the other members of the numerical analysis group: Martin, Bart, Fayçal, Marco, Gabriele, Tommaso, Ding, Thibaut, Vladimir, Conor, Michal, Pablo, Julian, Parisa, Sandie, Giancarlo, and Bo. I thank my friends from other groups at the department: Jhih-Huang, Aitor, Yaroslav, Renaud, Louis-Hadrien, Pascaline, Fathi, Giovanni, and Raphaël. Apart from university members, special thanks also go to the best friends I have made in Geneva, especially Imad, Bilal, Onur, and Shady. Thank you all for the valuable discussions and the good times. I could not forget to thank the secretariat of the Mathematics department at the University of Geneva, Joselle Besson, for helping me in managing all the administrative complications, before and after my arrival to Switzerland.

My deepest gratitude goes to my parents Adnan Almuslimani and Sara Hallak for their unconditional and endless love, care, and support. I would not have reached this stage without their sacrifices. I also thank all the members of my wonderful family, my brothers, sisters, nephews, and nieces, as well as all my friends in Lebanon.

Very big and special thanks go to my favorite mathematician, my wife Joanna Ajaj. She was always there to provide me with love, care, help, moral and mathematical support. Accomplishing this thesis would have been much more difficult without her motivation and encouragement. I would like to thank as well her beautiful family especially my father-in-law for his care and support.

It is impossible to forget my little princess Aysha, whose recent birth gave me a strong push forward to achieve more.

I am grateful to the Swiss National Science Foundation for funding my Ph.D. work and my postdoctoral research project.

الحَمْدُ للهِ الَّذِي بِنِعْمَتِهِ تَتِمُّ الصَّالِحَاتُ

Praise is to Allah by Whose grace good deeds are completed

Abstract

Explicit stabilized methods are an efficient and powerful alternative to implicit schemes for the time integration of stiff systems of differential equations in large dimensions. In the present thesis, we derive new explicit stabilized methods for different types of problems and we analyze their stability and convergence properties. We rigorously prove their efficiency, and we provide numerical experiments that illustrate their performance.

We provide in Chapter 1 an introduction to our work as well as a short summary of the main results presented in this thesis.

Chapter 2 is dedicated to necessary preliminaries, where we recall first the notion of stability of Runge-Kutta methods and we give a crash course on explicit stabilized integrators for deterministic ordinary differential equations (ODEs). Then, we explain briefly some useful notions about numerical integration of stochastic differential equations (SDEs).

In Chapter 3, we introduce a new explicit stabilized scheme of weak order 1 for stiff and ergodic stochastic differential equations (SDEs). In the absence of noise, the new method coincides with the classical deterministic stabilized scheme (or Chebyshev method) for diffusion dominated advection-diffusion problems. For mean-square stable stiff stochastic problems, the scheme has an optimally large extended mean-square stability domain that grows at the same quadratic rate as the deterministic stability domain size in contrast to known existing methods for stiff SDEs [A. Abdulle and T. Li. Commun. Math. Sci., 6(4), 2008, A. Abdulle, G. Vilmart, and K. C. Zygalakis, SIAM J. Sci. Comput., 35(4), 2013]. Combined with postprocessing techniques, the new methods achieve a convergence rate of order two for sampling the invariant measure of a class of ergodic SDEs, achieving a stabilized version of the non-Markovian scheme introduced in [B. Leimkuhler, C. Matthews, and M. V. Tretyakov, Proc. R. Soc. A, 470, 2014]. All the results are illustrated by numerical experiments on different types of problems. In the last section, an extension to advection-diffusion PDEs is discussed.

In Chapter 4, we derive, for the first time, explicit stabilized integrators of orders 1 and 2 for the optimal control of stiff systems. We analyze their favorable stability properties based on the continuous optimality conditions. Furthermore, we study their order of convergence taking advantage of the symplecticity of the corresponding partitioned Runge-Kutta method involved for the adjoint equations. The implementations of the new methods are done completely using two-term recurrence relations for both state (forward) and costate (backward) which reduces the effect of round-off errors that appear in standard Runge-Kutta implementations. The recurrence relations are derived carefully to avoid order reduction phenomenon and make the methods symplectic. Numerical experiments including the optimal control of a nonlinear advection-diffusion PDE illustrate the efficiency of the new approach.

Finally, we give in Chapter 5 an outlook and some ideas for potential future work. We also draw some conclusions.

Résumé

Les méthodes explicites stabilisées sont une alternative efficace et puissante aux schémas implicites pour l'intégration temporelle de systèmes raides d'équations différentielles en grande dimension. Dans la présente thèse, nous développons de nouvelles méthodes explicites stabilisées pour différents types de problèmes et nous analysons leurs propriétés de stabilité et de convergence. Nous prouvons rigoureusement leur efficacité, et nous présentons des expériences numériques qui illustrent leurs performances.

Le chapitre 1 est une brève introduction ainsi qu'un résumé des résultats principaux présentés dans cette thèse.

Dans le chapitre 2, nous rappelons d'abord la notion de stabilité des méthodes Runge-Kutta et nous donnons un cours accéléré sur les intégrateurs explicites stabilisés pour les équations différentielles ordinaires (EDO) déterministes. Ensuite, nous expliquons brièvement quelques notions utiles sur l'intégration numérique des équations différentielles stochastiques (EDS).

Dans le chapitre 3, nous introduisons un nouveau schéma explicite stabilisé d'ordre faible 1 pour les EDS raides et ergodiques. En l'absence de bruit, la nouvelle méthode coïncide avec le schéma stabilisé déterministe classique (ou méthode de Tchebyshev) pour les problèmes d'advection-diffusion dominés par la diffusion. Pour les problèmes stochastiques raides stables en moyenne quadratique, le schéma a un domaine de stabilité en moyenne quadratique étendu optimal qui croît à la même vitesse quadratique que la taille du domaine de stabilité déterministe, contrairement aux méthodes existantes connues pour les EDS raides [A. Abdulle et T. Li. Commun. Math. Sci. 6(4), 2008, A. Abdulle, G. Vilmart, et K. C. Zygalakis, SIAM J. Sci. Comput. 4(4), 2013]. Combinées aux techniques de postprocessing, les nouvelles méthodes atteignent un taux de convergence d'ordre 2 pour l'échantillonnage de la mesure invariante d'une classe de EDS ergodiques, réalisant une version stabilisée de la méthodes non-markovienne introduite dans [B. Leimkuhler, C. Matthews, et M. V. Tretyakov, Proc. R. Soc. A, 470, 2014]. Tous les résultats sont illustrés par des expériences numériques sur différents types de problèmes. Dans la dernière section, une extension aux EDP de type advection-diffusion est discutée.

Dans le chapitre 4, nous dérivons, pour la première fois, des intégrateurs explicites stabilisés des ordres 1 et 2 pour le contrôle optimal des systèmes raides. Nous analysons leurs propriétés de stabilité favorables en basant sur les conditions d'optimalité continue. En outre, nous étudions leur ordre de convergence en tirant avantage de la symplecticité de la méthode Runge-Kutta partitionnée correspondante impliquée pour les équations adjointes. L'implémentation des nouvelles méthodes est effectuée entièrement en utilisant des relations de récurrence à deux termes pour l'état et l'adjoint, ce qui réduit l'effet des erreurs d'arrondi qui apparaissent dans l'implémentation standard des méthodes de Runge-Kutta. Les relations de récurrence sont obtenues avec soin pour éviter le phénomène de réduction d'ordre et rendre les méthodes symplectiques. Des expériences numériques incluant le contrôle optimal d'une EDP d'advection-diffusion non linéaire illustrent l'efficacité de la nouvelle approche.

Nous présentons dans le chapitre 5 quelques perspectives et idées pour de futurs travaux potentiels. Nous tirons également quelques conclusions.

Contents

Chapter 1

Introduction and main results

This chapter is dedicated to introduce the topics studied in the thesis and present its main results briefly. More details and references can be found in the corresponding chapters.

Stiff differential equations are differential equations for which standard explicit numerical integrators are numerically unstable, unless the step size is taken to be extremely small. Intuitively, such equations usually include one or more terms that cause fast variations in the solution. One important example of such equations is diffusion equations. The usage of implicit methods is a good alternative in small dimensions, but for high dimensional problems, in addition to large round-off errors, the cost of implicit methods increases dramatically especially if the problem is severely nonlinear. Explicit stabilized methods serve as an alternative to implicit integrators in high dimensional problems. Indeed, these Runge-Kutta type methods enjoy extended stability domains over the negative real axis which, on the one hand, efficiently reduces the restriction on the step size faced by standard explicit methods, and on the other hand, allow us to avoid solving large dimensional systems of equations. Explicit stabilized methods will be introduced in details in Chapter 2.

In what follows, we present a summary of our main results which will be detailed in chapters 3 and 4 with more details and references.

1.1 Optimal explicit stabilized integrators for stiff and ergodic stochastic differential equations

This contribution is published in [5] in collaboration with Assyr Abdulle and Gilles Vilmart, and detailed in Chapter 3.

We consider Itô systems of stochastic differential equations of the form

$$dX(t) = f(X(t))dt + \sum_{r=1}^{m} g^r(X(t))dW_r(t), \qquad X(0) = X_0 \tag{1.1}$$

where $X(t)$ is a stochastic process with values in $\mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}^d$ is the drift term, $g^r : \mathbb{R}^d \to \mathbb{R}^d$, $r = 1, \dots, m$ are the diffusion terms, and $W_r(t)$, $r = 1, \dots, m$, are

independent one-dimensional Wiener processes fulfilling the usual assumptions. We assume that the drift and diffusion functions are smooth enough and Lipschitz continuous to ensure the existence and uniqueness of a solution of (1.1) on a given time interval $(0, T)$. The simplest numerical method to solve this problem is the Euler-Maruyama method which, analogously to the Euler method for deterministic ODEs, faces severe time step restriction when applied to stiff SDEs like diffusion problems. One can use implicit methods with favorable stability properties, but for very large dimension they are often very costly as they require to solve large dimensional nonlinear problems at every time step. Here we focus on explicit methods which are very useful for large dimensions.

The standard S-ROCK method In 2008, a stochastic explicit stabilized method called S-ROCK (for stochastic orthogonal Runge-Kutta-Chebyshev) was designed in [8], it is a Runge-Kutta method defined as follows, using an explicit two term recurrence relation with s is the number of drift function evaluations

$$\begin{aligned} K_0 &= X_0 \\ K_1 &= X_0 + \mu_1 h f(X_0) \\ K_i &= \mu_i h f(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad i = 2, \dots, s, \\ X_1 &= K_s + \sum_{r=1}^{m} g^r(K_s) \Delta W_r, \end{aligned} \tag{1.2}$$

where,

$$\omega_0 = 1 + \frac{\eta}{s^2}, \quad \omega_1 = \frac{T_s(\omega_0)}{T_s'(\omega_0)}, \quad \mu_1 = \frac{w_1}{w_0}, \tag{1.3}$$

and η is the damping parameter (see Figure 1.2). The coefficients μ_i, ν_i, and κ_i are chosen in an appropriate way as functions of ω_0 and ω_1. T_s is the first kind Chebyshev polynomial defined by $T_s(\cos\theta) = \cos(s\theta)$.

Advantages of S-ROCK

- It is very easy to implement (fully explicit) similarly to the Euler-Maruyama method (recovered for s=1).
- It is consistent, i.e. it has weak order 1 and strong order 1/2.
- Its stability domain grows quadratically with the number of function evaluations which reduces the cost compared to Euler-Maruyama.

Drawbacks of S-ROCK

- The large damping parameter $\eta = \eta_s$ needed to stabilize the stiff noise term is an increasing function of s which reduces the stability domain size down to $\approx 0.33s^2$ much lower than the optimal one which is $2s^2$.
- It has only order 1 of accuracy when approximating the invariant measure of the overdamped Langevin equation.

1.1.1 New second kind Chebyshev methods

New SK-ROCK The new S-ROCK method, denoted SK-ROCK (for stochastic second kind orthogonal Runge-Kutta-Chebyshev method) is defined with a recurrence relation similar to (1.2) except that the noise is introduced in the first internal stage:

$$\begin{aligned} K_0 &= X_0 \\ K_1 &= X_0 + \mu_1 h f(X_0 + \nu_1 Q) + \kappa_1 Q, \\ K_i &= \mu_i h f(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad i = 2, \dots, s. \\ X_1 &= K_s, \end{aligned} \tag{1.4}$$

where $Q = \sum_{r=1}^{m} g^r(X_0)\Delta W_j$.

Advantages of SK-ROCK

- The same advantages as S-ROCK.
- The damping parameter η is fixed to a small value $\eta = 0.05$ which allows a nearly optimal stability domain size (about $(2 - 4/3\eta)s^2$ and $2s^2$ for $\eta = 0$), which makes SK-ROCK much less expensive than S-ROCK (see Section 1.1.2).
- Combined with postprocessing techniques this method achieves order two for the invariant measure for a class of ergodic SDEs (see Section 1.1.3).

1.1.2 Mean-square Stability analysis

Let $s \geq 1$ and $\eta \geq 0$. Applied to the linear test equation $dX = \lambda X dt + \mu X dW$ (widely used in the literature [49]), the new scheme SK-ROCK yields

$$X_{n+1} = R_{SK-ROCK}(\lambda h, \mu\sqrt{h}, \xi_n) X_n$$

where $p = \lambda h, q = \mu\sqrt{h}$, $\xi_n \sim \mathcal{N}(0,1)$ is a Gaussian variable and the stability function given by

$$\begin{aligned} R_{SK-ROCK}(p, q, \xi) &= \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} + \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)}(1 + \frac{\omega_1}{2}p)q\xi \\ &= A(p) + B(p)q\xi, \end{aligned} \tag{1.5}$$

where U_s are the second kind Chebyshev polynomials of degree s, we have that $T_s'(x) = sU_{s-1}(x)$ and so U_s contains sine function. On the other hand the stability function of classical S-ROCK is

$$R_{S-ROCK}(p, q, \xi) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} + \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} q\xi, \tag{1.6}$$

The new idea is to use second kind Chebyshev polynomials to stabilize the noise. These polynomials are defined using the sine function and so they are the optimal choice to put together in a sum of two squares (second moment) with the first kind Chebyshev polynomials which are defined using cosine, and thanks to the relation $T_s^2(x) + (1 - x^2)U_{s-1}^2(x) = 1$ we get the optimal stability domain under the corresponding order conditions (see Figure 1.1).

$$\mathbb{E}(|R_{SK-ROCK}(p,q,\xi)|^2) = A(p)^2 + B(p)^2 q^2$$

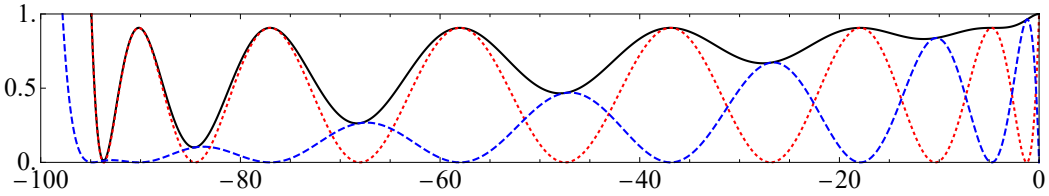


Figure 1.1: Different parts of the second moment of the stability function of SK-ROCK with $q^2 = -2p$, and $s = 7$ stages.

Theorem 1.1.1. *There exists $\eta_0 > 0$ and s_0 such that for all $\eta \in [0, \eta_0]$ and all $s \geq s_0$, for all $p \in [-2\omega_1^{-1}, 0]$ and $p + \frac{1}{2}|q|^2 \leq 0$, we have $\mathbb{E}(|R_{sk-ROCK}(p,q,\xi)|^2) \leq 1$.*

Theorem 1.1.1 is the first result of this kind in the context of stochastic explicit stabilized methods that proves rigorously the large size of the mean square stability domain. In contrast, the stability regions of all the methods previously proposed in the literature were checked numerically and not rigorously. Theorem states that for all η small enough and all s large enough, the length of stability domain is $2w_1^{-1} \approx (2 - 4/3\eta)s^2$ which is arbitrarily close to $2s^2$ when $\eta \to 0$. We have checked numerically that the statement remains valid for any $\eta \geq 0$ and $s \in \mathbb{N}^*$.

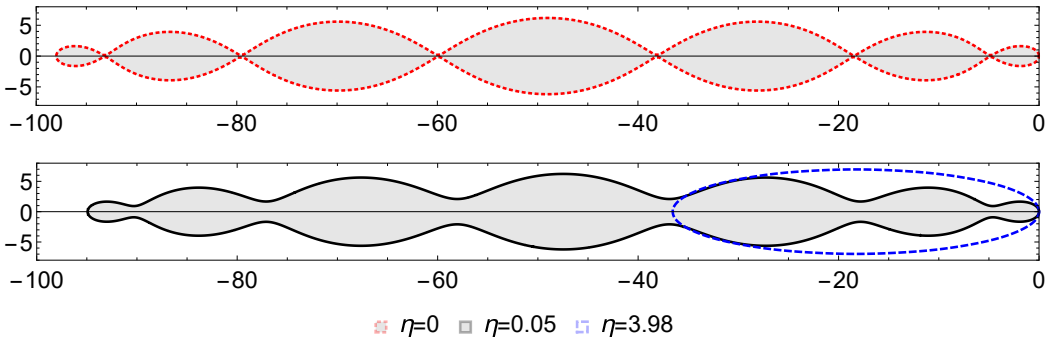


Figure 1.2: Deterministic complex stability domain $\{p \in \mathbb{C};\ |R_{SK-ROCK}(p,0,0)| \leq 1\}$ for different damping parameters η and s=7 stages.

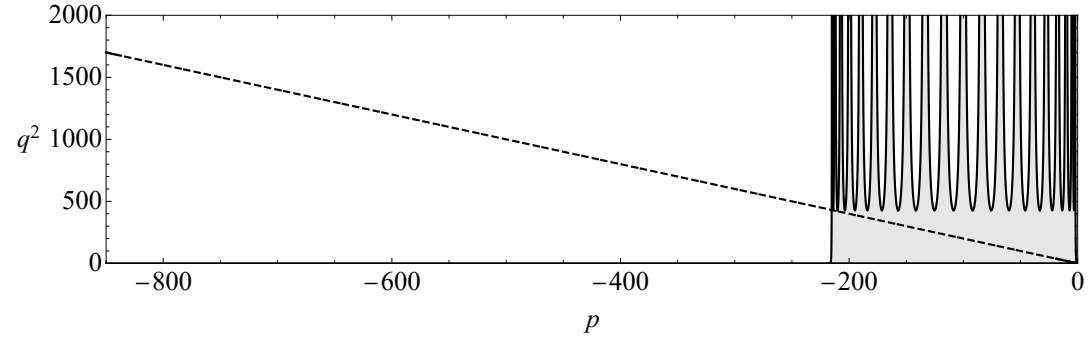
1.1.3 PSK-ROCK: Postprocessed integrator for overdamped Langevin equation

We consider the overdamped Langevin equation in $\mathbb{R}^d$ $(1 \ll d)$,

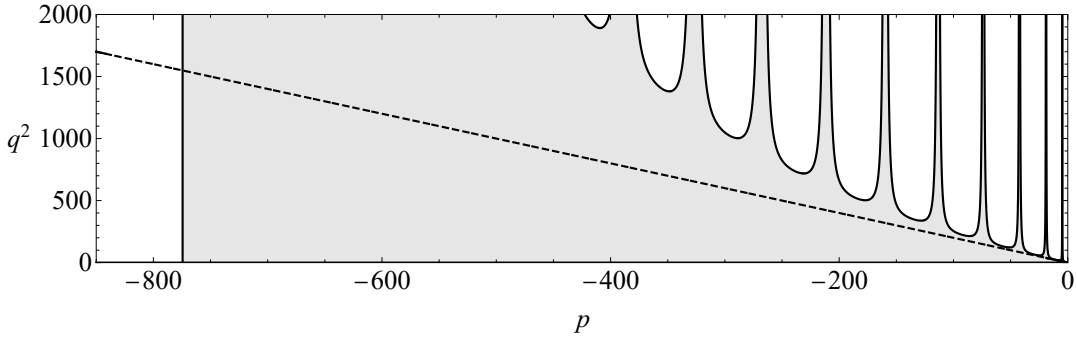
$$dX(t) = -\nabla V(X(t))dt + \sigma dW(t),$$

Under some natural assumptions, the above equation is ergodic with exponential convergence to a unique invariant measure with Gibbs density $\rho_\infty = Z \exp(-2\sigma^{-2}V(x))$, and we have

$$|\mathbb{E}(\phi(X(t)) - \int_{\mathbb{R}^d} \phi(x)\rho_\infty(x)dx| \leq Ce^{-\lambda t},$$



(a) Standard S-ROCK method ($s = 20$, $\eta = 6.95$).



(b) New SK-ROCK method ($s = 20$, $\eta = 0.05$).

Figure 1.3: Stochastic mean-square stability domains. The dashed lines corresponds to the upper boundary $q^2 = -2p$ of the real mean-square stability domain $\mathcal{S}$ of the exact solution.

for test function ϕ and all initial condition X_0, where C, λ are independent of t.

A modification to reach high order for the invariant measure The new method SK-ROCK can be modified to compute efficiently ergodic integrals $\int_{\mathbb{R}^d} \phi(x)\rho_\infty(x)dx$ in high dimension d with order 2 of accuracy. We propose to modify the internal stage K_1 of the method as follows:

$$\begin{aligned} K_1 &= X_0 + \mu_1 h f(X_0 + \nu_1 Q) + \kappa_1 Q + \alpha h\big(f(X_0 + \nu_1 Q) \\ &- 2f(X_0) + f(X_0 - \nu_1 Q)\big). \end{aligned} \tag{1.7}$$

Theorem 1.1.2. *Under the above assumptions, consider the scheme SK-ROCK with modified internal stage K_1 (1.7). Consider in addition a postprocessor defined as*

$$\overline{X}_n = X_n + c\sigma\sqrt{h}\xi.$$

where α and c are chosen appropriately.

Then, $\overline{X}_n$ yields order two for the invariant measure,

$$|\mathbb{E}(\phi(\overline{X}_n) - \int_{\mathbb{R}^d} \phi(x)\rho_\infty(x)dx| \leq C_1 e^{-\lambda t_n} + C_2 h^2,$$

for all $t_n = nh$ with h small enough. C_1 and C_2 are independent of h and n.

Theorem 1.1.2 means that using a posprocessing technique coupled with SK-ROCK, we can approximate the invariant measure of the overdamped Langevin equation with order two of accuracy with negligeable overcost (only 2 additional f evaluations per time step) and we keep the same optimally large stability domain.

Many numerical experiments for stiff linear and nonlinear problems are presented in Chapter 3. They illustrate the advantage of SK-ROCK and PSK-ROCK over other explicit stabilized schemes from the literature (S-ROCK, SROCK2...) with respect to the cost and the order, as well as the error constants.

1.2 Explicit stabilized integrators for stiff optimal control problems

This contribution is published in [14] in collaboration with Gilles Vilmart, and detailed in Chapter 4.

We aim to introduce and analyze symplectic explicit stabilized Runge-Kutta methods of order 2 for the optimal control of systems of ordinary differential equations (ODEs) of the form

$$\min_u \ \Psi(y(T)); \qquad \dot{y}(t) := \frac{dy}{dt}(t) = f(u(t), y(t)), \quad t \in [0,T]; \quad y(0) = y^0, \tag{1.8}$$

where for a fixed final time $T > 0$ and a given initial condition $y^0 \in \mathbb{R}^n$, the function $y : [0,T] \to \mathbb{R}^n$ is the unknown state function, $u : [0,T] \to \mathbb{R}^m$ is the unknown control function. Here, $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ is the given vector field and $\Psi : \mathbb{R}^n \to \mathbb{R}$ is the given cost function, which are assumed to be C^∞ mappings.

We recall that any Runge-Kutta method applied to the linear ODE $\dot{y} = \lambda y$, $y(0) = y^0$ yields an induction $y_n = R(z)^n y^0$ where $z = h\lambda$. $R(z)$ is usually a rational function called the stability function and it reduces to a polynomial in the case of explicit methods. The stability domain is then defined as $\mathcal{S} = \{z \in \mathbb{C};\ |R(z)| \leq 1\}$.

Explicit stabilized methods are Runge-Kutta methods with extended stability domain over the negative real axis. They have been applied to various types of stiff (diffusion) problems. Usually, the stability polynomials of these methods are constructed using Chebyshev polynomials. For more information about Explicit stabilized Runge-Kutta methods we refer to the review [4]. In Figure 1.4 we plot the stability domains of the second-order explicit stabilized method called RKC (Runge-Kutta-Chebyshev) for $s = 13$ stages, and the well known second-order Heun method.

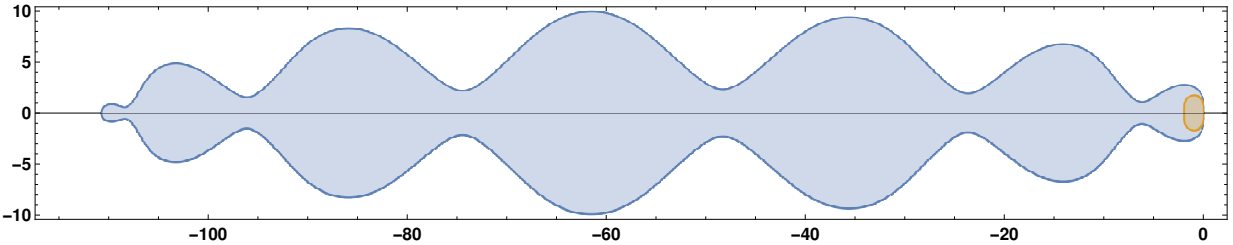


Figure 1.4: Stability domains of the RKC method for $s = 7$ stages (in blue) and the Heun method (in brown).

If we discretize (1.8) using a Runge-Kutta discretization we naturally get the following discrete optimization problem,

$$\begin{aligned} &\min \ \ \Psi(y_N); \quad \text{subject to:} \\ &y_{k+1} = y_k + h\sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \qquad y_{ki} = y_k + h\sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}). \end{aligned} \tag{1.9}$$

Let us denote by $H(u, y, p) := p^T f(u, y)$ the pseudo-Hamiltonian of the system where p is the Lagrange multiplier (or the costate) associated to the state y. Applying Pontryagin's maximum (or minimum) principle, the first order optimality conditions of (1.8) are

$$\begin{aligned} \dot{y}(t) &= f(u(t), y(t)) = \nabla_p H(u(t), y(t), p(t)), \\ \dot{p}(t) &= -\nabla y f(u(t), y(t))p = -\nabla_y H(u(t), y(t), p(t)), \\ 0 &= \nabla_u H(u(t), y(t), p(t)). \ \ t \in [0, T], \quad y(0) = y^0, \quad p(T) = \nabla\Psi(y(T)). \end{aligned} \tag{1.10}$$

By applying the Lagrange theorem to the finite dimensional optimization problem (1.9), and supposing that $b_i \neq 0$, a calculation yields the discrete optimality conditions,

$$\begin{aligned} &y_{k+1} = y_k + h\sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \quad y_{ki} = y_k + h\sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}), \\ &p_{k+1} = p_k - h\sum_{i=1}^{s} \hat{b}_i \nabla_y H(u_{ki}, y_{ki}, p_{ki}), \ p_{ki} = p_k - h\sum_{j=1}^{s} \hat{a}_{ij} \nabla_y H(u_{kj}, y_{kj}, p_{kj}), \\ &0 = \nabla_u H(u_{ki}, y_{ki}, p_{ki}), \quad y_0 = y^0, \quad p_N = \nabla\Psi(y_N), \end{aligned} \tag{1.11}$$

where $k = 0, \ldots, N-1, \ i = 1, \ldots, s$, and the coefficients $\hat{b}_i$ and $\hat{a}_{ij}$ are defined by the following relations which correspond to the *symplecticity* conditions of partitioned Runge-Kutta methods for ODEs,

$$\hat{b}_i := b_i, \qquad \hat{a}_{ij} := b_j - \frac{b_j}{b_i} a_{ji}, \qquad i, j = 1, \ldots s.$$

For symplectic Runge-Kutta methods, the following diagram commutes [27, 16]:

$$\begin{array}{ccc} (1.8) & \xrightarrow{\text{discretization}} & (1.9) \\ \text{optimality conditions}\downarrow & & \downarrow\text{optimality conditions} \\ (1.10) & \xrightarrow[\text{discretization}]{} & (1.11) \end{array}$$

In [27] Hager showed as well that if the state method (a_{ij}, b_i) is of order 2 of accuracy, then the obtained symplectic partitioned scheme $(a_{ij}, b_i) - (\hat{a}_{ij}, \hat{b}_i)$ is automatically of order 2 (no additional coupling order conditions).

A calculation on (1.11) yields the Runge-Kutta method $(\frac{b_j}{b_i} a_{ji}, b_i)$ for the costate (time reversed), which is in fact the time adjoint of $(\hat{a}_{ij}, \hat{b}_i)$ (which is called in the literature the adjoint of (a_{ij}, b_i)). We will call the method $(\frac{b_j}{b_i} a_{ji}, b_i)$ the **double adjoint** of (a_{ij}, b_i), and

we rewrite the costate equations in (1.11) as

$$\begin{aligned} p_k &= p_{k+1} + h\sum_{i=1}^{s} b_i \nabla_y H(u_{ki}, y_{ki}, p_{ki}), \quad k = N-1, \dots, 0 \\ p_{ki} &= p_{k+1} + h\sum_{j=1}^{s} \frac{b_j}{b_i} a_{ji} \nabla_y H(u_{kj}, y_{kj}, p_{kj}), \quad k = N-1, \dots, 0 \quad i = s, \dots, 1 \\ 0 &= \nabla_u H(u_{kj}, y_{kj}, p_{kj}), \qquad k = 0, \dots, N-1 \quad i = 1, \dots, s, \quad p_N = \nabla\Psi(y_N). \end{aligned} \tag{1.12}$$

The following proposition is crucial in our analysis. It implies that when using an explicit method to discretize (1.8), the resulting partitioned method is fully explicit.

Proposition 1.2.1. *If a Runge-Kutta method (a_{ij}, b_i) is explicit, then its double adjoint $(\frac{b_j}{b_i}a_{ji}, b_i)$ is explicit as well.*

1.2.1 New Runge-Kutta-Chebyshev (RKC) method for optimal control

We consider the following implementation of the RKC method applied to (1.8)

$$\begin{aligned} &\min \Psi(y_N), \ \textit{such that} \\ &y_{k0} = y_k, \quad y_{k1} = y_{k0} + \mu_1 h f(u_{k0}, y_{k0}), \\ &y_{ki} = \mu_i h f(u_{k,i-1}, y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \dots, s \\ &y_{k+1} = a_s y_{k0} + b_s T_s(\omega_0) y_{ks}, \end{aligned} \tag{1.13}$$

where,

$$a_s = 1 - b_s T_s(\omega_0), \ b_s = \frac{T_s''(\omega_0)}{(T_s'(\omega_0)^2)}, \ \omega_0 = 1 + \frac{\eta}{s^2}, \ \omega_1 = \frac{T_s'(\omega_0)}{T_s''(\omega_0)}, \ \eta = 0.15,$$

and

$$\mu_i = \frac{2\omega_1 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \ \nu_i = \frac{2\omega_0 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad i = 2, \dots, s.$$

The real positive number η is called the damping parameter and its non zero value helps to include a strip around the negative real axis in the stability domain to make the scheme robust with respect to small perturbations such as a small advection term. The stability function of the s-stage RKC method is $R_s(z) = a_s + b_s T_s(\omega_0 + \omega_1 z)$, where $T_s(\cos\theta) = \cos(s\theta)$ are the Chebyshev polynomials. Besides the extended stability domain, this method has many advantages. The two-term recurrence relations require low memory (only two stages should be stored) and do not introduce round-off errors for large number of stages, in contrast to applying the method using the corresponding Runge-Kutta coefficients (a_{ij}, b_i). In addition, and as mentioned before, applying RKC in both directions (for y and p) leads to an order reduction since the coupling order conditions are not satisfied. Hence, the aim is to search for recurrence formulas of the RKC double adjoint. Before that, we introduce our first theorem of this section which will be very helpful for stability analysis.

Theorem 1.2.2 (Stability function of the double adjoint)**.** *A Runge-Kutta method (a_{ij}, b_i) and its double adjoint $(\frac{b_j}{b_i}a_{ji}, b_i)$ share the same stability function $R(z)$.*

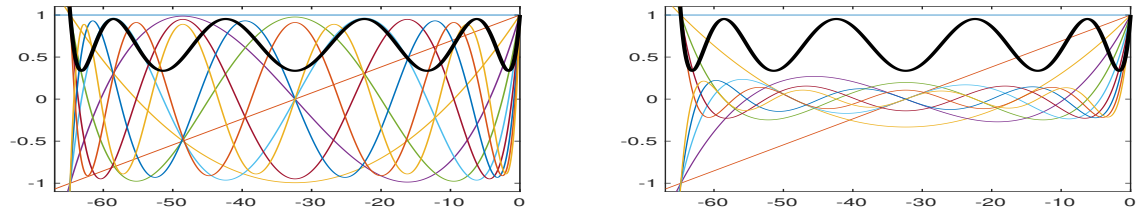


Figure 1.5: Internal stages (thin curves) and stability polynomial (bold curve) of the RKC method (left) and its double adjoint (right) of order two for $s = 10$ internal stages.

Theorem 1.2.2 means that RKC and its double adjoint share the same stability region. A full stability analysis for the internal stages of the new scheme is available in Chapter 4, and is summarized in Theorem 1.2.4 below.

We may now state our main result in the context of stiff optimal control problems:

Theorem 1.2.3 (Recurrence formulas of the RKC double adjoint)**.** *The double adjoint of the scheme* (1.13) *is given by the recurrence*

$$\begin{aligned}
p_N &= \nabla\Psi(y_N), \quad p_{ks} = p_{k+1},\\
p_{k,s-1} &= p_{ks} + \frac{\mu_s}{\nu_s} h\nabla_y H(u_{k,s-1}, y_{k,s-1}, p_{ks}),\\
p_{k,s-j} &= \frac{\mu_{s-j+1}\alpha_{s-j+1}}{\alpha_s - j} h\nabla_y H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1})\\
&\quad + \frac{\nu_{s-j+1}\alpha_{s-j+1}}{\alpha_{s-j}} p_{k,s-j+1},\\
&\quad + \frac{(1-\nu_{s-j+2})\alpha_{s-j+2}}{\alpha_{s-j}} p_{k,s-j+2}, \quad j = 2,\dots,s-1,\\
p_{k0} &= \mu_1\alpha_1 h\nabla_y H(u_{k0}, y_{k0}, p_{k1}) + \alpha_1 p_{k1} + (1-\nu_2)\alpha_2 p_{k2} + a_s p_{k+1},\\
p_k &= p_{k0},\\
\nabla_u H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) &= 0, \quad j = 1,\dots,s,
\end{aligned} \tag{1.14}$$

where the coefficients α_j are defined using the induction

$$\begin{aligned}
\alpha_s &= b_s T_s(\omega_0), \quad \alpha_{s-1} = \nu_s\alpha_s,\\
\alpha_{s-j} &= \nu_{s-j+1}\alpha_{s-j+1} + (1-\nu_{s-j+2})\alpha_{s-j+2}, \quad j = 2\dots s-1.
\end{aligned} \tag{1.15}$$

In Figure 1.5, we plot the internal stages, as well as the stability functions of the RKC method (1.13) and its double adjoint (1.14) for $s = 13$ stages. The following two theorems show the good stability properties, as well as the order 2 of convergence of the new scheme.

Theorem 1.2.4 (Stability)**.** *For $\eta = 0$, the stability functions of the internal stages $R_{s,i}(z)$ of the RKC double adjoint* (1.14)*, are bounded by* 1 *for all $z \in [-\frac{2}{3}s^2 + \frac{2}{3}, 0]$ and all $s \in \mathbb{N}$.*

Remark 1.2.5. *As a corollary, using the continuity of the coefficients with respect to the damping parameter η, the above theorem is true for all small enough damping $\eta > 0$.*

Theorem 1.2.6 (Convergence)**.** *The method* (1.13)-(1.14) *has order 2 for the optimal control problem* (1.8)*.*

Numerical experiments of different problems including the optimal control of a nonlinear advection-diffusion partial differential equation are presented in Chapter 4.

Chapter 2

Preliminaries

In this chapter, we provide the essential tools needed to read and understand the main chapters of the thesis. We recall some useful definitions and results on Runge-Kutta methods, especially explicit stabilized methods, and on numerical integration of stochastic differential equations. All necessary preliminaries on optimal control are presented in the corresponding Chapter 4.

2.1 Introduction to explicit stabilized Runge-Kutta methods

In this section we introduce explicit stabilized methods which are one of the main ingredients of the results presented in this thesis. First, a crash course on Runge-Kutta methods and their stability is presented. Then, a short revision on Chebyshev polynomials and their property is done. Finally, we present explicit stabilized methods in details and we give examples of three different integrators of this type. We include these explanations for the sake of completeness since explicit stabilized methods serve as a main ingredient of our work.

2.1.1 Stability of Runge-Kutta methods

Let us first recall the definition of a Runge-Kutta method for ordinary differential equations (ODEs),

$$\dot{y}(t) := \frac{dy}{dt} = f(t, y(t)), \qquad y(t_0) = y^0, \tag{2.1}$$

where $y : [t_0, T] \to \mathbb{R}^n$ is the unknown solution, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth vertor field, and $y^0 \in \mathbb{R}^n$ is a given initial condition. For simplicity, we consider a uniform discretization of the interval $[t_0, T]$ with $N + 1$ points for $N \in \mathbb{N}$, and denote by $h = (T - t_0)/N$ the stepsize.

Definition 2.1.1. *For a given integer s and real coefficients b_i, a_{ij} $(i, j = 1, \dots, s)$, an s-stage Runge-Kutta method, $y_k \approx y(t_k)$, $t_k = t_0 + kh$, to approximate the solution of* (2.1),

is defined, for all $k = 0, \ldots, N-1$, *by*

$$\begin{aligned} y_{ki} &= y_k + h \sum_{j=1}^{s} a_{ij} f(t_k + c_j h, y_{kj}), \quad i = 1, \ldots, s, \\ y_{k+1} &= y_k + h \sum_{i=1}^{s} b_i f(t_k + c_i h, y_{ki}), \end{aligned} \tag{2.2}$$

where $y_0 = y^0$ *and* $\sum_{j=1}^{s} a_{ij} = c_i$.

The coefficients are usually displayed in a Butcher tableau as follows

$$\begin{array}{c|c} c_i & a_{ij} \\ \hline & b_i \end{array} := \begin{array}{c|c} c & A \\ \hline & b^T \end{array}, \tag{2.3}$$

and we will sometimes use the notation (a_{ij}, b_i). If the matrix A is strictly lower triangular, i.e. $a_{ij} = 0$ for all $i \leq j$ (or strictly upper triangular by reverting the indices), then the method is explicit, otherwise it is implicit. For $s = 1$, there exists one and only one consistent explicit Runge-Kutta method of one stage, which is simply the famous explicit Euler method $y_{k+1} = y_k + hf(t_k, y_k)$, with Butcher tableau

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}.$$

Definition 2.1.2. *A Runge-Kutta method* (2.2) *with stepsize* h, *applied to a sufficiently smooth problem of the form* (2.1), *is said to be of order* p, *if its local error (error after one step) satisfies*

$$\|y(t_0 + h) - y_1\| = \mathcal{O}(h^{p+1}).$$

The following tables represent the implicit Euler method of order 1, and the second order schemes: Heun method (or explicit trapezoidal rule), implicit and explicit midpoint rules, and (implicit) trapezoidal rule respectively.

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

Figure 2.1 illustrates graphically how to get the approximation y_1 of $y(t_1)$ using Heun method

$$y_1 = y_0 + \frac{h}{2}[f(y_0) + f(y_0 + hf(y_0))]$$

with the following Runge-Kutta formulation

$$\begin{aligned} Y_1 &= f(y_0) \\ Y_2 &= f(y_0 + hf(Y_1)) \\ y_1 &= y_0 + \frac{h}{2}(f(Y_1) + f(Y_2)). \end{aligned}$$

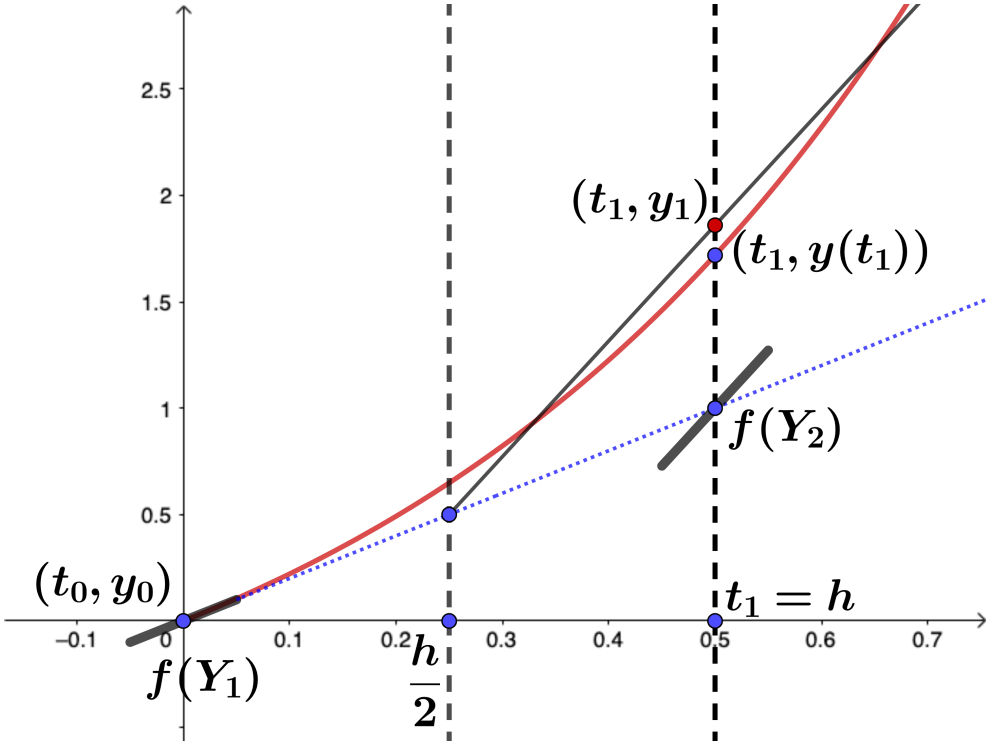


Figure 2.1: Graphical illustration of the approximation by Heun method after one step.

A Runge-Kutta method (a_{ij}, b_i) is of order one if and only if its coefficients satisfy

$$\sum_{i=1}^{s} b_i = 1.$$

For order two, one more condition is required which is

$$\sum_{i,j=1}^{s} b_i a_{ij} = \sum_{i=1}^{s} b_i c_i = \frac{1}{2}.$$

For more details about the order conditions of Runge-Kutta methods in the context of initial value ODEs, we refer for example to the book [28, Chap. III].

Throughout the thesis, and for simplicity of the presentation, we will consider only autonomous problems where the vector field f does not depend explicitly on time t

$$\dot{y} = f(y), \qquad y(t_0) = y^0. \tag{2.4}$$

However, we highlight that our results also apply straightforwardly to non-autonomous problems. A standard approach is to consider the augmented system with $z(t) = t$, i.e. $\frac{dz}{dt} = 1$, $z(0) = 0$ and define $\tilde{y}(t) = (y(t), z(t))^T$, see e.g. [28, Chap. III] for details.

We denote by $y_{k+1} = \Phi_h(y_k)$ the numerical flow of (2.2), while the time adjoint method Φ_h^* of Φ_h is the inverse map of the original method with reversed time step $-h$, i.e., $\Phi_h^* := \Phi_{-h}^{-1}$ [28, Sect. II.3]. We recall that the time adjoint of an s-stage Runge-Kutta method (a_{ij}, b_i) (2.2) is again an s-stage Runge-Kutta method with the same order of accuracy and its coefficients (a_{ij}^*, b_i^*) are given by

$$a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j} \text{ and } b_i^* = b_{s+1-i}, \text{ where } i, j = 1, \dots, s.$$

Stability is a crucial property of numerical integrators for solving stiff problems and we refer to the book [29] for a detailed study of stiff ODEs and stability of numerical integrators. We say that the solution $y(t)$ of the ODE (2.4) (with $f(0) = 0$) is stable if $\lim_{t\to\infty} y(t) = 0$. A Runge-Kutta method is said to be stable if the numerical solution stays bounded along the integration process. Applying a Runge-Kutta method (2.2) to the linear test problem (with fixed parameter $\lambda \in \mathbb{C}$ and whose exact solution is stable when $\Re(\lambda) < 0$),

$$\dot{y} = \lambda y, \quad y(0) = y_0, \tag{2.5}$$

with stepsize h yields a recurrence of the form $y_{k+1} = R(h\lambda)y_k$ and by induction we get $y_k = R(h\lambda)^k y_0$. The function $R(z)$ is called the stability function of the method and the stability domain is defined as

$$\mathcal{S} := \{z \in \mathbb{C}; |R(z)| \leq 1\}, \tag{2.6}$$

and y_k remains bounded if and only if $h\lambda \in \mathcal{S}$. The same result also applies to the internal stages of the Runge-Kutta method, for all $i = 1, \ldots, s$, where s is the number of internal stages, $y_{k_i} = R_i(h\lambda)y_k$, for some rational functions R_i. Remark that $R(z)$ is a rational function for implicit methods, but in the case of explicit methods the stability function $R(z)$ reduces to a polynomial which explains that the stability domain is necessarily bounded in this case. The simplest Runge-Kutta type method to integrate ODEs (2.1) is the order one explicit Euler method $y_{k+1} = y_k + hf(y_k)$ with stability polynomial $R(z) = 1 + z$. Indeed, applied to the linear test problem (2.5), the method reads

$$y_{k+1} = y_k + h\lambda y_k = (1 + h\lambda)y_k.$$

However, its stability domain $\mathcal{S}$ is very small. Indeed, it reduces to the disc of center -1 and radius 1 in the complex plane as show in Figure 2.2a, which yields a severe timestep restriction and makes it very expensive for stiff problems. For instance in the case of ODEs arising from the space discretization of diffusion partial differential equations (PDEs) with mesh size Δx, using the explicit Euler method yields the famous very restrictive CFL condition $h \leq C\Delta x^2$. On the other hand, one can easily prove that the order one implicit Euler method $y_{k+1} = y_k + hf(y_{k+1})$ has as stability function $R(z) = \frac{1}{1-z}$ and hence its stability domain is the complementary set of the disk of center 1 and radius 1 (Figure 2.2b), which makes the method unconditionally stable for diffusion problems. Nevertheless, this comes at the price of solving large systems of size proportional to $(1/\Delta x)^d$ (d is the spatial dimension) at each time step.

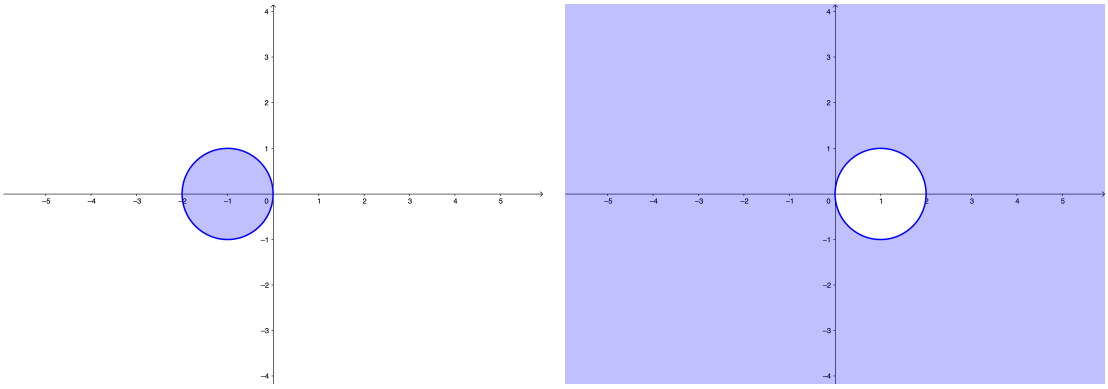
This is the motivation for seeking a new type of methods which combines, as much as possible, the advantages of explicitness, i.e. ease of implementation and avoiding to solve large systems, and good stability properties.

Finally we will recall a few useful definitions:

Definition 2.1.3. *A Runge-Kutta method is called A-stable if its stability domain contains the left complex half place $\mathbb{C}^-$.*

Definition 2.1.4. *A Runge-Kutta method is called L-stable if it is A-stable and its stability function $R(z)$ goes to zero as z tends to $\pm\infty$.*

Example 2.1.5. *The implicit Euler method is L-stable.*



(a) Explicit Euler. (b) Implicit Euler.

Figure 2.2: Stability domains of the explicit and the implicit Euler methods.

2.1.2 A quick revision of Chebyshev polynomials

Let us first recall the definitions and some properties of the first and second kind Chebyshev polynomials, which will be very useful throughout this thesis.

Definition 2.1.6. *The first kind Chebyshev polynomial of degree s is the unique polynomial satisfying $T_s(\cos\theta) = \cos(s\theta)$.*

One can easily see that for $x \in [-1,1]$, $T_s(x) = \cos(s \arccos x)$ and hence $|T_s(x)| \leq 1$. These polynomials are orthogonal on $[-1,1]$ with respect to the inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\frac{dx}{\sqrt{1-x^2}},$$

which means that the family of polynomials $(T_i)_{i\leq s}$ forms a basis of the space of polynomials of degree $\leq s$.

The first kind Chebyshev polynomials can be defined for every complex number z by the recurrence

$$\begin{aligned} T_0(z) &= 1, \qquad T_1(z) = z, \\ T_s(z) &= 2zT_{s-1}(z) - T_{s-2}(z), \quad s \geq 2. \end{aligned} \tag{2.7}$$

Definition 2.1.7. *The second kind Chebyshev polynomial of degree s is the unique polynomial satisfying*

$$U_s(\cos\theta) = \frac{\sin((s+1)\theta)}{\sin\theta}.$$

The second kind Chebyshev polynomials are orthogonal with respect to the inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\sqrt{1-x^2}dx,$$

and for every $z \in \mathbb{C}$, we have

$$\begin{aligned} U_0(z) &= 1, \qquad U_1(z) = 2z, \\ U_s(z) &= 2zU_{s-1}(z) - U_{s-2}(z), \quad s \geq 2. \end{aligned} \tag{2.8}$$

The fact that both first and second kind Chebyshev polynomials have the same recurrence relation will be very useful in our result of Chapter 3. Note that second kind Chebyshev polynomials were not used before in the literature for explicit stabilized methods.

Here are also some useful properties: for all $s \in \mathbb{N}$, and all $z \in \mathbb{C}$,

- $T_s(1) = 1$,
- $T'_s(1) = s$,
- $U_s(z) = \frac{1}{s} T'_{s+1}(z)$,
- $T_s(z)^2 - (z^2 - 1) U_{s-1}(z)^2 = 1$.

2.1.3 Explicit stabilized methods

The idea of explicit stabilized methods (as introduced in [54], see the survey [4]) is to construct explicit Runge-Kutta integrators with extended stability domain that grows quadratically with the number of stages s of the method along the negative real axis, and hence allows to use large time steps typically for problems arising from diffusion partial differential equations (or diffusion dominant advection-diffusion-reaction PDEs) for which the eigenvalues of the Jacobian matrix of the obtained vector field are on (or very close to) the negative real axis and are very large in modulus (stiff ODEs). In order to construct an explicit stabilized integrator of order p, the first step is to find a polynomial $R_s(z)$ of degree s and order p, i.e.

$$R_s(z) = 1 + z + \cdots + \frac{z^p}{p!} + \mathcal{O}(z^{p+1}), \tag{2.9}$$

that solves the following problem

$$\begin{aligned} &\text{Find} \quad R_s(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^p}{p!} + \alpha_{p+1} z^{p+1} + \cdots + \alpha_s z^s, \\ &|R_s(z)| \leq 1 \ \text{ for } \ z \in [-l_s^p, 0], \quad \text{with } l_s^p \text{ as large as possible.} \end{aligned} \tag{2.10}$$

Note that for $p > 2$, stability functions of the form (2.9) guarantee the order p only for linear problems and some additional order conditions have to be satisfied by the method in order to have order p for nonlinear problems.

In this section we will present some explicit stabilized schemes of orders one and two from the literature. The presented methods will be the main ingredients to our new results.

2.1.3.1 Optimal first order Chebyshev methods

It is well known in the literature of explicit stabilized methods that the solution of Problem (2.10) for $p = 1$ is the shifted Chebyshev polynomials

$$R_s(z) = T_s\left(1 + \frac{z}{s^2}\right), \tag{2.11}$$

where $T_s(.)$ is the first kind Chebyshev polynomial of degree s. For a given integer s and real number x, this polynomial stays bounded between -1 and 1 for $-1 \leq 1 + x/s^2 \leq 1$

i.e. $-2s^2 \leq x \leq 0$, which means that the real negative interval $[-2s^2, 0]$ in included in the stability domain (see Figure 2.3), and $l_s^1 = 2s^2$ is the optimal length. For higher orders we have

$$l_s^2 \approx 0.82s^2, \quad l_s^3 \approx 0.49s^2 \quad l_s^4 \approx 0.34s^2.$$

Approximations of l_s^p up to order $p = 11$ can be found in [2]. The fact that the length of the stability domain on the negative real axis enjoys a quadratic growth with respect to the number of stages s is crucial to the success of explicit stabilized Runge-Kutta methods.

There are many approaches to construct, for a given integer s, a Runge-Kutta method having (2.11) as stability function, the most reasonable approach with respect to memory and round-off errors is the one considered by Van der Houwen and Soomeijer [54] which uses the two term recurrence relation (2.7) of the Chebyshev polynomials to construct the numerical method given by

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_k + \frac{h}{s^2} f(y_{k0}), \\ y_{ki} &= \frac{2h}{s^2} f(y_{k,i-1}) + 2y_{k,i-1} - y_{k,i-2}, \quad i = 2, \ldots, s \\ y_{k+1} &= y_{ks}, \end{aligned} \tag{2.12}$$

where $k = 0, \ldots, N-1$. It can be easily verified, by induction and using (2.7), that applied to the test problem (2.5), the above method leads, for the internal stages, to

$$y_{ki} = T_i(1 + h\lambda/s^2)y_k, \quad i = 1 \ldots, s,$$

and produces after one step

$$y_{k+1} = y_{ks} = T_s(1 + h\lambda/s^2)y_k.$$

The method has low memory requirements (only two stages have to be stored) and reasonable propagation of round-off errors even for large values of s needed in practice [54].

As can be seen in the first plot of Figure 2.3, the width of the stability domain reduces to zero at some points which are the local extrema of (2.11), i.e. the points $x_i \in \mathbb{R}^-$ where $R_s(x_i) = T_s(1 + x_i/s^2) = \pm 1$ which can cause instability in the case of small perturbations (small advection term for example). Here comes the importance of what we call *damping* to make the scheme robust with respect to small perturbations of the eigenvalues. If one sets

$$R_s^\eta(z)y_k = \frac{T_s(\omega_0 + \omega_1 z)}{T_s(\omega_0)} y_k, \quad \omega_0 := 1 + \frac{\eta}{s^2}, \quad \omega_1 := \frac{T_s(\omega_0)}{T_s'(\omega_0)}, \tag{2.13}$$

then the polynomials (2.13) have the correct order and oscillate between $-1+\eta$ and $1-\eta$, where $\eta > 0$ is called the damping parameter. This makes the stability domain a bit wider and ensures that it includes a strip around the negative real axis (a neighborhood of the negative real interval $[-l_s, 0]$), but it comes with the cost of loosing a bit of the length of the stability domain (the length reduces from $2s^2$ to $\approx 2 - \frac{4}{3}\eta$). By increasing the value of η, a larger strip around the negative real axis can be included in the stability domain which becomes shorter. A typical value for the damping parameter is $\eta = 0.05$, in this case the loss in the length of the stability domain is negligible, where l_s becomes $\approx 1.94s^2$ (see Figure 2.3).

The order one Chebyshev method for solving a stiff ODE (2.1) is defined as an explicit s-stage Runge-Kutta method by the recurrence

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_k + \mu_1 h f(y_{k0}), \\ y_{ki} &= \mu_i h f(y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \ldots, s \\ y_{k+1} &= y_{ks}, \end{aligned} \tag{2.14}$$

where $k = 0, \ldots, N-1$, and

$$\mu_1 := \frac{\omega_1}{\omega_0}, \quad \mu_i := \frac{2\omega_1 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad \nu_i := \frac{2\omega_0 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad i = 2, \ldots, s. \tag{2.15}$$

For $\eta = 0$ (without damping) we get $\omega_0 = 1,\ \omega_1 = 1/s^2,\ \mu_1 = 1/s^2,\ \mu_i = 2/s^2,\ \nu_i = 2$, for all $i = 2, \ldots, s$, and te method reduces to (2.12). One can easily check that the (family) of methods (2.14) has the same first order of accuracy as the explicit Euler method (recovered for $s = 1$). Note that instead of the standard Runge-Kutta method formulation (2.2) with coefficients (a_{ij}, b_i), the one step method $y_{k+1} = \Phi_h(y_k)$ in (2.14) should be implemented using a recurrence relation (indexed by i) inspired from the relation (2.7) on Chebyshev polynomials. This implementation (2.14) yields a good stability [54] of the scheme with respect to round-off errors. The most interesting feature of this scheme is its stability behavior. Indeed, the method (2.14) applied to (2.5) yields, with $z = \lambda h$,

$$y_{k+1} = R_s^\eta(z) y_k = \frac{T_s(\omega_0 + \omega_1 z)}{T_s(\omega_0)} y_k. \tag{2.16}$$

A large real negative interval $(-C_\eta s^2, 0)$ is included in the stability domain of the method $\mathcal{S} := \{z \in \mathbb{C}; |R_s^\eta(z)| \leq 1\}$. For the internal stages, we have analogously

$$y_{ki} = R_{s,i}^\eta(z) y_k = \frac{T_i(\omega_0 + \omega_1 z)}{T_i(\omega_0)} y_k.$$

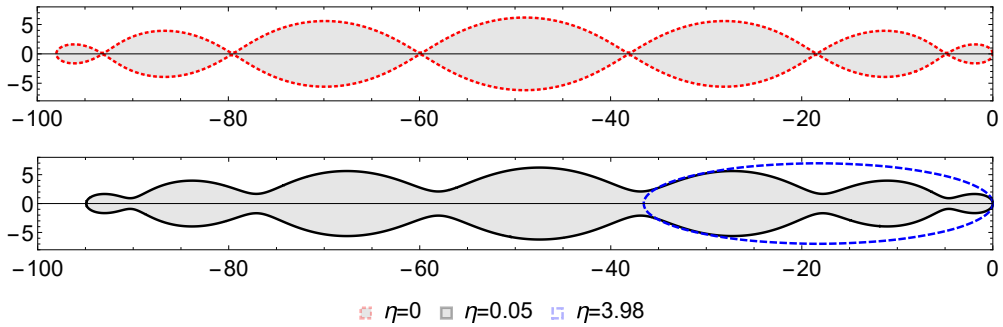
The constant $C_\eta = 2 - 4/3\,\eta + \mathcal{O}(\eta^2)$ depends on the damping parameter η and for $\eta = 0$, it reaches the maximal value $C_0 = 2$ (also optimal with respect to all possible stability polynomials for explicit schemes of order 1). Hence, given the stepsize h, for dissipative vector fields with a Jacobian having large real negative eigenvalues (such as diffusion problems) with spectral radius $\lambda_{\max}$ at y_n, the parameter s for the next step y_{n+1} can be chosen adaptively as[1]

$$s := \left[\sqrt{\frac{h\lambda_{\max} + 1.5}{2 - 4/3\,\eta}} + 0.5 \right], \tag{2.17}$$

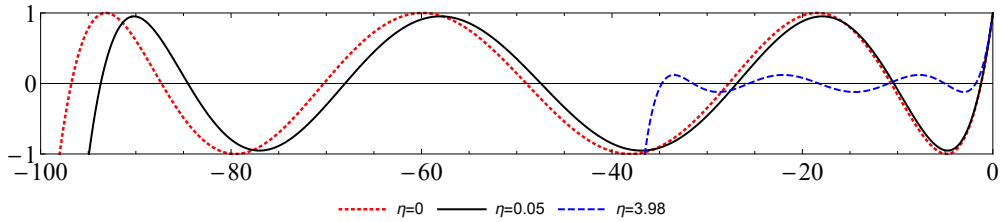
see [3] in the context of stabilized schemes of order two with adaptive stepsizes. The method (2.14) is much more efficient as its stability domain increases *quadratically* with the number s of function evaluations while a composition of s explicit Euler steps (same cost per time step) has a stability domain that only increases *linearly* with s.

In Figure 2.4 we plot the internal stages for $s = 10$ and different values $\eta = 0$ and $\eta = 0.05$, respectively. We observe that in the absence of damping ($\eta = 0$), the stability function (here a polynomial) is bounded by 1 in the large real interval $[-2s^2, 0]$ of width $2 \cdot 10^2 = 200$.

[1]The notation $[x]$ stands for the integer rounding of real numbers.

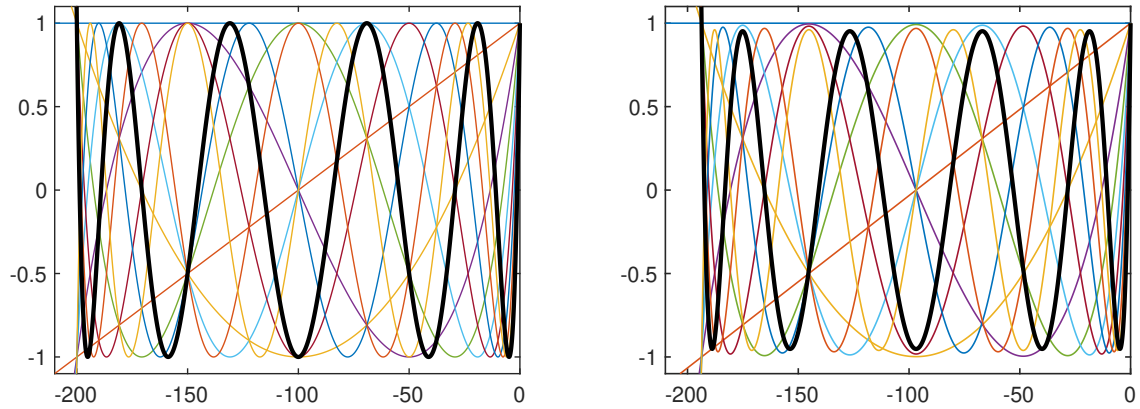


(a) Complex stability domain for different damping parameters η.



(b) Corresponding stability functions $R_7^\eta(z)$.

Figure 2.3: Stability domains and stability functions of the Chebyshev method for $s = 7$ and different damping values $\eta = 0, 0.05, 3.98$.



(a) $\eta = 0$ (b) $\eta = 0.05$

Figure 2.4: Internal stages (thin curves) and stability polynomials (bold curves) of the Chebyshev method (2.14) for $s = 10$ with and without damping.

2.1.3.2 Second order RKC methods

To design a second order method, we need the stability polynomial to satisfy[2]

$$R(z) = 1 + z + \frac{z^2}{2} + \mathcal{O}(z^3). \tag{2.18}$$

[2]Indeed, up to order two, the order conditions for nonlinear problems are the same as the order conditions for linear problems [28, Chap. III].

In [15], Bakker introduced a correction to the first order shifted Chebyshev polynomials to get the following second order polynomial

$$R_s^\eta(z) = a_s + b_s T_s(\omega_0 + \omega_2 z), \tag{2.19}$$

where,

$$a_s = 1 - b_s T_s(\omega_0), \quad b_s = \frac{T_s''(\omega_0)}{(T_s'(\omega_0)^2)}, \quad \omega_0 = 1 + \frac{\eta}{s^2}, \quad \omega_2 = \frac{T_s'(\omega_0)}{T_s''(\omega_0)}, \quad \eta = 0.15. \tag{2.20}$$

For each s, $|R_s^\eta(z)|$ remains bounded by $a_s + b_s = 1 - \eta/3 + \mathcal{O}(\eta^2)$ for z in the stability interval (except for a small interval near the origin). The stability interval along the negative real axis is approximately $[-0.65s^2, 0]$, and covers about 80% of the optimal stability interval for second order stability polynomials, and the formula now for calculating s for a given time step h is

$$s := \left\lceil \sqrt{\frac{h\lambda_{\max} + 1.5}{0.65}} + 0.5 \right\rceil. \tag{2.21}$$

Using the recurrence relation of the Chebyshev polynomials, the RKC method as introduced in [54] is defined by

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_{k0} + h b_1 \omega_2 f(y_{k0}), \\ y_{ki} &= y_{k0} + \mu_i' h (f(y_{k,i-1}) - a_{i-1} f(y_{k0})) + \nu_i'(y_{k,i-1} - y_{k0}) + \kappa_i'(y_{k,i-2} - y_{k0}), \\ y_{k+1} &= y_{ks}, \end{aligned} \tag{2.22}$$

where $k = 0, \ldots, N-1$, and

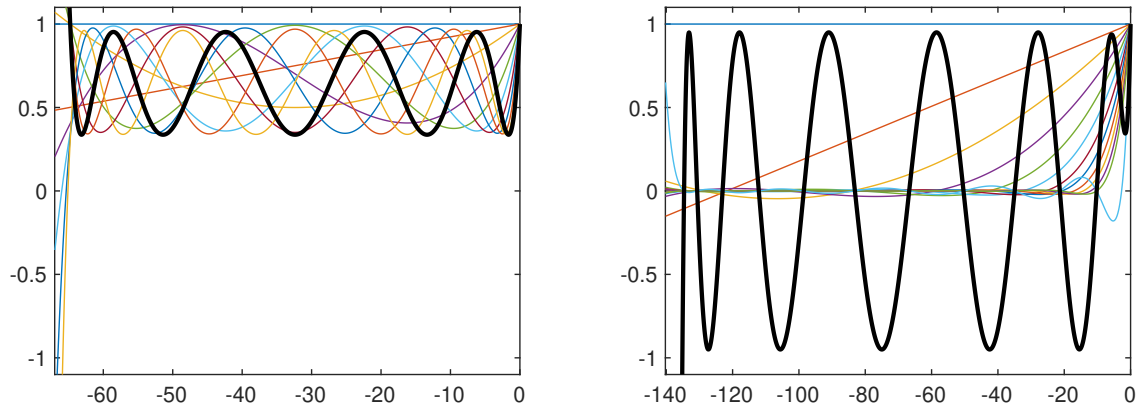
$$\mu_i' = \frac{2b_i\omega_2}{b_{i-1}}, \quad \nu_i' = \frac{2b_i\omega_0}{b_{i-1}}, \quad \kappa_i' = -\frac{b_i}{b_{i-2}}, \quad b_i = \frac{T_i''(\omega_0)}{(T_i'(\omega_0)^2)}, \quad a_i = 1 - b_i T_i(\omega_0), \tag{2.23}$$

for $i = 2, \ldots, s$. As in (2.19), the stability functions of the internal stages are given by $R_i^\eta(z) = a_i + b_i T_i(\omega_0 + \omega_2 z)$, where $i = 0, \ldots, s-1$, and the parameters a_i and b_i are chosen such that the above stages are consistent $R_i^\eta(z) = 1 + \mathcal{O}(z)$. The parameters b_0 and b_1 are free ($R_0^\eta(z)$ is constant and only order 1 is possible for R_1^η(z)) and the values $b_0 = b_1 = b_2$ are suggested in [52]. Figure 2.5a illustrates the stability polynomials of the internal stages of the RKC method (2.22) for $s = 10$ stages.

Figure 2.6 illustrates the advantage of RKC over the known Heun method of order 2 by showing the big difference between the stability domains of both methods, where we set $s = 7$ stages for RKC.

2.1.3.3 Nearly optimal second order family: ROCK2 methods

For explicit stabilized methods of higher order with optimal stability domains, no analytical expressions are known for the stability polynomials. The ROCK methods (for orthogonal Runge-Kutta-Chebyshev) considered in [2, 3, 9] are explicit stabilized methods of nearly optimal stability polynomials (cover 98% of the optimal domain), built on a recurrence relation, and have been obtained for orders 2 and 4. Using the fact that the optimal



(a) RKC (2.22). (b) ROCK2 (2.25).

Figure 2.5: Internal stages (thin curves) and stability polynomials (bold curve) of the RKC method for $s = 10$ and the ROCK2 method for $s = 13$ stages.

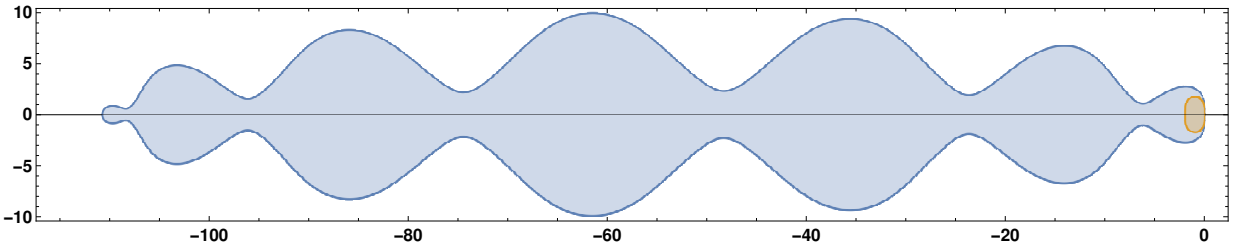


Figure 2.6: Stability domains of the RKC method for $s = 7$ stages (in blue) and the Heun method (in brown).

stability polynomials of even order p possess exactly p complex zeros [1], the key idea is then to seek, for a given p, an approximate solution of the problem (2.10) of the form

$$R_s(z) = w_p(z)P_{s-p}(z), \tag{2.24}$$

where P_{s-p} is a member of the family $\{P_j\}_{j\geq 0}$, with P_j is a polynomial of degree j. This family is orthogonal with respect to the weight function

$$w_p(z)^2/\sqrt{1-z^2},$$

where $w_p(z)$ is a positive polynomial of degree p, and its zeros are the p complex zeros of $R_s(z)$ (very close to the complex zeros of (2.10). In particular, for $p = 1$, we recover the optimal order 1 shifted Chebyshev polynomials (2.11) with $w_1(z) = 1$. For a given integer $s \geq 2$, the second order ROCK2 method reads

$$\begin{aligned} y_{k0} &= y_k, \\ y_{k1} &= y_k + \mu_1'' h f(y_{k0}), \\ y_{ki} &= \mu_i'' h f(y_{k,i-1}) + \nu_i'' y_{k,i-1} + (1-\nu_i'')y_{k,i-2}, \quad i = 2,\ldots,s-2, \end{aligned} \tag{2.25}$$

and then the quadratic factor $w_2(z) = 1 + 2\sigma z + \tau z^2$ is represented by a two-stage *finishing procedure*

$$\begin{aligned} y_{k,s-1} &= y_{k,s-2} + h\sigma f(y_{k,s-2}), \\ y^\star_{ks} &= y_{k,s-1} + h\sigma f(y_{k,s-1}), \\ y_{ks} &= y^\star_{ks} - h\sigma(1 - \frac{\tau}{\sigma^2})(f(y_{k,s-1}) - f(y_{k,s-2})). \\ y_{k+1} &= y_{ks}. \end{aligned} \tag{2.26}$$

The coefficients μ''_i and ν''_i are precomputed numerically (independently of f). Applying the method to the linear test problem $y' = \lambda y$, we obtain for all $i = 0, \ldots, s-2$, $y_{ki} = P_i(z)y_k$ and $y_{ks} = w_2(z)y_{k,s-2}$, hence $y_{k+1} = R_s(z)y_k$, where $z = \lambda h$.

Figure 2.5b, illustrates the internal stages and the stability function of the ROCK2 method for $s = 13$ stages. We see that $l_s \approx 136 \approx 0.81s^2$, and the optimal l_s for second order polynomials is approximately $0.82s^2$.

A fourth order explicit stabilized method called ROCK4 is introduced in [3], and is constructed in a similar procedure as ROCK2. The additional difficulty for order $p > 2$ is that the order conditions are no more the same as for linear problems [28, Chap. III].

2.2 Introduction to numerical integration of stochastic differential equations

A stochastic differential equation (SDE) is a differential equation in which one or more terms are random, i.e., depends on a *white noise* which is the formal derivative of a *Brownian motion*. We will clarify these terms throughout this section. Stochastic differential equations have important applications in various domains. Indeed, they are used to model stock prices, molecular dynamics, physical systems subject to thermal fluctuations, and other phenomena in physics, chemistry, economy and other domains. In this section, we will recall some preliminaries about stochastic differential equations (SDEs). A few definitions and main properties of Itô stochastic integrals are introduced. We will recall as well the definition of mean square stability for the exact and the numerical solution of an SDE.

The reader is supposed to have a basic background knowledge of stochastic processes and numerical integration of ordinary differential equations (ODEs). We refer to the review article [32] for a practical and more detailed introduction to numerical methods for stochastic differential equations including useful basic Matlab programs.

2.2.1 Brownian motion and Itô stochastic integral

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 2.2.1. *A Brownian motion (or Wiener process) $(W(t))_{t\in[0,T]}$ over a closed interval $[0, T]$ is a stochastic process that satisfies the following:*

i. W is real valued and $W(0) = 0$.

ii. The trajectories of W are continuous almost surely on $[0, T]$.

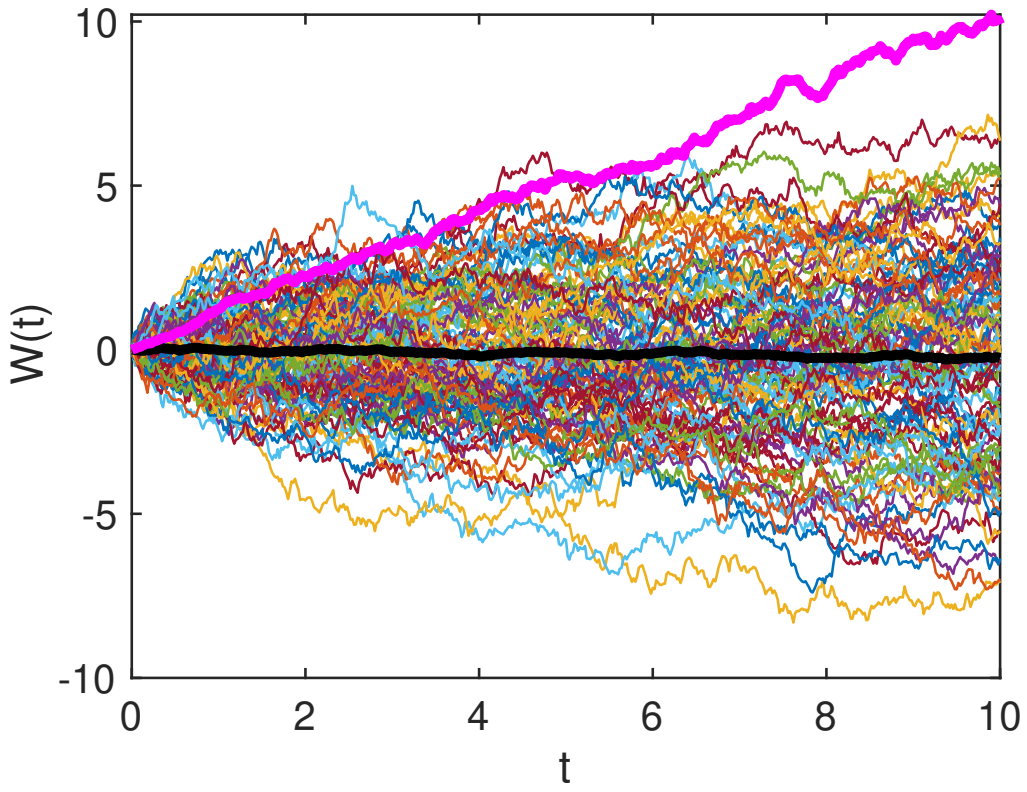


Figure 2.7: A plot of 100 Brownian paths, their average (thick black curve), and their variance (thick magenta curve).

iii. The increments of W are independent, i.e., for $0 \leq s < t \leq u < v \leq T$, the increments $W(t) - W(s)$ and $W(v) - W(u)$ are independent.

iv. For $0 \leq s < t \leq T$, the random variable given by the increment $W(t) - W(s)$ is normally distributed with mean 0 and variance $t-s$, i.e., $W(t) - W(s) \sim \mathcal{N}(0, t-s)$.

It can be proven that the trajectories of a Brownian motion are differentiable nowhere with probability 1. In Figure 2.7, we plot 100 trajectories of a Brownian motion together with their average and variance with respect to time.

2.2.1.1 Stochastic integrals

For a given (integrable) function $g : [0, T] \to \mathbb{R}$, the integral $\int_0^T g(t)dt$ can be approximated using the Riemann sum

$$\sum_{j=0}^{N-1} g(t_j)(t_{j+1} - t_j), \tag{2.27}$$

where $t_j = jh$, and $h = T/N$ for some integer N, then the integral may be defined by taking the limit of (2.27) when $h \to 0$ (i.e. $N \to \infty$). Using the same analogy, we may

define the stochastic integral $\int_0^T g(t)dW(t)$ as the limit when $h \to 0$ of

$$\sum_{j=0}^{N-1} g(t_j)(W(t_{j+1}) - W(t_j)). \tag{2.28}$$

The "left-hand" sum (2.28) gives rise to what is know as the *Itô* integral. Evaluating g in the sum at the midpoint $(t_j + t_{j+1})/2$ gives the *Stratonovich* integral. Throughout this thesis we will use the Itô version, but we recall that a simple transformation converts from Itô to Stratonovich and vice versa.

Example 2.2.2. *We have, using Itô integration,*

$$\int_0^T W(t)dW(t) = \frac{W(T)^2}{2} - \frac{T}{2}. \tag{2.29}$$

Indeed,

$$\begin{aligned} W(T)^2 &= \left(\sum_{j=0}^{N-1}(W(t_{j+1}) - W(t_j))\right)^2 \\ &= \sum_{j=0}^{N-1}(W(t_{j+1}) - W(t_j))^2 + 2\sum_{i=1}^{N-1}\sum_{j=0}^{i-1}(W(t_{j+1}) - W(t_j))(W(t_{i+1}) - W(t_i)) \\ &= \sum_{j=0}^{N-1}(W(t_{j+1}) - W(t_j))^2 + 2\sum_{i=0}^{N-1} W(t_i)(W(t_{i+1}) - W(t_i)) \quad (W(t_0) = 0), \end{aligned}$$

the term $\sum_{j=0}^{N-1}(W(t_{j+1}) - W(t_j))^2$ can be shown to have expected value T and variance $\mathcal{O}(h)$, which means that as $h \to 0$, it converges to T. On the other hand, the term $2\sum_{i=0}^{N-1} W(t_i)(W(t_{i+1}) - W(t_i))$ clearly converges to $2\int_0^T W(t)dW(t)$ as $h \to 0$, and hence we have the desired equality. Note that using Stratonovich integration we get

$$\int_0^T W(t) \circ dW(t) = W(T)^2/2,$$

without the Itô term $-T/2$.

2.2.2 Itô formula and stochastic differential equations

An autonomous stochastic differential equation can be written in integral form as

$$X(t) = X_0 + \int_0^t f(X(s))ds + \int_0^t g(X(s))dW(s), \quad 0 \le t \le T, \tag{2.30}$$

where the initial condition $X_0 \in \mathbb{R}^d$ is a vector of random variables (it can also be deterministic), and the vector fields f (called the drift) and g (called the diffusion) are functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ fulfilling some usual technical assumptions.

Usually, the equation (2.30) is written in the form of differential equation as

$$dX(t) = f(X(t))dt + g(X(t))dW(t), \quad X(0) = X_0, \quad 0 \le t \le T. \tag{2.31}$$

Here $dW(t)$ is called the *white noise* and it can be seen formally as the time derivative of the Wiener process $W(t)$. Now let $X(t)$ be a stochastic process verifying (2.31), and let $u : \mathbb{R}^d \to \mathbb{R}^d$ be a function of class C^2. For simplicity of the notation we set $d = 1$, then the Itô differentiation formula (or stochastic chain rule) reads

$$\begin{aligned} d(u(X(t))) &= u'(X(t))dX(t) + \frac{1}{2}u''(X(t))g(X(t))^2dt \\ &= u'(X(t))f(X(t))dt + u'(X(t))g(X(t))dW(t) + \frac{1}{2}u''(X(t))g(t)^2dt. \end{aligned} \tag{2.32}$$

The term $\frac{1}{2}u''(X(t))g(t)^2dt$ is called the "Itô term".

Example 2.2.3. *1. Consider the linear SDE*

$$dX = \lambda X dt + \mu X dW, \quad X(0) = X_0, \quad 0 \leq t \leq T. \tag{2.33}$$

Suppose that X stays different than zero and let $Y = ln(X)$, and let us apply the Itô formula

$$\begin{aligned} dY &= Y'(X)dX + \frac{1}{2}\mu^2X^2Y''(X)dt \\ &= \frac{1}{X}(\lambda X dt + \mu X dW) - \frac{1}{2}\mu^2X^2\frac{1}{X^2}dt \\ &= \lambda dt + \mu dW - \frac{1}{2}\mu^2 dt \\ &= (\lambda - \frac{1}{2}\mu^2)dt + \mu dW, \end{aligned}$$

integrating both sides, we get $Y(t) = Y(0) + (\lambda - \frac{1}{2}\mu^2)t + \mu W(t)$ and hence

$$X(t) = X_0 \exp((\lambda - \frac{1}{2}\mu^2)t + \mu W(t)). \tag{2.34}$$

It can be easily verified that $X(t)$ defined in (2.34) solves (2.33).

2. Another way to prove (2.29) is to consider the SDE $dX = dW(t)$, $X(0) = 0$ (i.e $X(t) = W(t)$), and $u(X) = X^2$, we have then $u'(X) = 2X$ and $u''(X) = 2$. The Itô formula yields

$$du(X(t)) = d(W(t)^2) = 1dt + 2W(t)dW(t).$$

Using $W(0) = 0$ we get

$$W(T)^2 = T + 2\int_0^T W(t)dW(t),$$

and hence

$$\int_0^T W(t)dW(t) = \frac{W(T)^2}{2} - \frac{T}{2}.$$

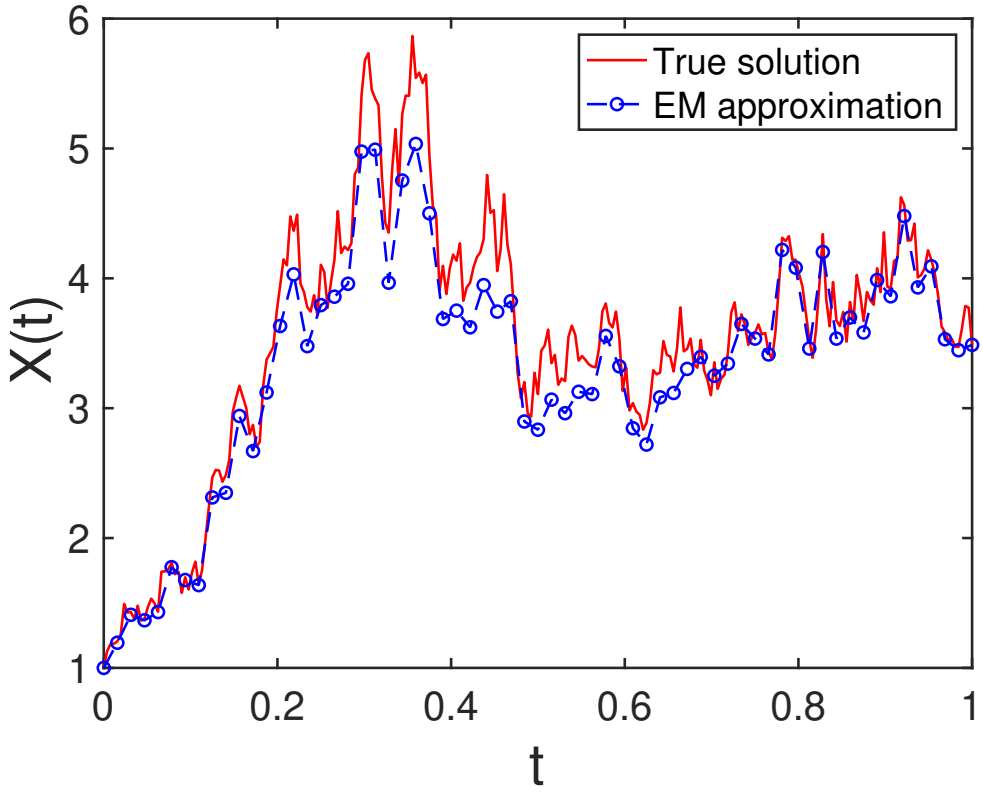


Figure 2.8: Exact solution of (2.33) and its approximation using Euler-Maruyama method (2.36).

2.2.3 Numerical Integration and mean square stability

In this section we will introduce standard explicit numerical methods to integrate an SDE of the form (2.31). We will analyze their convergence and stability properties after defining those notions.

We consider a uniform discretization of the interval $[0, T]$, $t_0 = 0 < t_1 < \cdots < t_N = T$, where $t_n = nh$, $h = T/N$ and $n = 0, \ldots, N$. A one step numerical integrator for the approximation of (2.31) at time $t = nh$ is a discrete dynamical system of the form

$$X_{n+1} = \Psi(X_n, h, \xi_n) \tag{2.35}$$

where X_n is an approximation of $X(t_n)$ and ξ_n are independent random vectors. The simplest one step method to integrate (2.31) is the Euler-Maruyama method defined as

$$X_{n+1} = X_n + hf(X_n) + g(X_n)\Delta W_n, \quad X(0) = X_0, \tag{2.36}$$

where $\Delta W_n = W(t_{n+1}) - W(t_n) \sim \mathcal{N}(0, hI_d)$ are the discrete Brownian increments. This is just a generalization of the deterministic Euler method recovered in the absence of noise. In Figure 2.8, we plot the one realization of the exact solution (2.34) of the linear SDE (2.33) as well as its approximation using Euler-Maruyama scheme for $T = 1$ and $N = 64$.

Truncating the Itô-Taylor expansion of the exact solution at an appropriate point, we get the Milstein method

$$X_{n+1} = X_n + hf(X_n) + g(X_n)\Delta W_n + \frac{1}{2}g'(X_n)g(X_n)((\Delta W_n)^2 - h), \quad X(0) = X_0, \quad (2.37)$$

which turns out to be more accurate than the Euler-Maruyama scheme with respect to the strong order of convergence which will be defined in the next section. Another common method is the generalization of the well known $\theta-$method defined by

$$X_{n+1} = X_n + h(1-\theta)f(X_n) + h\theta f(X_{n+1}) + g(X_n)\Delta W_n, \quad X(0) = X_0, \quad (2.38)$$

which is clearly implicit for $\theta \neq 0$ and coincides with the Euler-Maruyama method (2.36) for $\theta = 0$.

2.2.3.1 Strong, Weak, and invariant measure convergence of stochastic numerical integrators

Consider an SDE of the form (2.31), and let us denote by the sequence $\{X_n\}_{n\geq 0}$ a numerical approximation of its solution. We denote by $\|.\|_p$ the usual $p-$norm on $\mathbb{R}^d$, and in practice we often use $p = 1, 2$. Finally, denote by $\mathbb{E}(Y)$ the expected value of a random variable Y.

Definition 2.2.4. *A numerical method $\{X_n\}_{n\geq 0}$ to approximate (2.31) is said to be of strong order $r > 0$, if for all h small enough and $t_n = nh \leq T$, we have*

$$\mathbb{E}(\|X_n - X(t_n)\|_p) \leq Ch^r, \quad (2.39)$$

where C is independent of h and n.

The strong order of convergence measures the decay rate of the mean of the error between each individual trajectory and its numerical approximation as $h \to 0$. An alternative which turns out to be very useful in practice, is to measure the rate at which the error of the trajectories' means decay, i.e., measuring the convergence in law of the numerically computed stochastic process to the exact solution. This introduces the concept of weak convergence.

Definition 2.2.5. *A numerical method $\{X_n\}_{n\geq 0}$ to approximate (2.31) is said to be of weak order $q > 0$, if for all test function ϕ satisfying some smoothness and polynomial growth conditions, and all h small enough and $t_n = nh \leq T$, we have*

$$|\mathbb{E}(\phi(X_n)) - \mathbb{E}(\phi(X(t_n)))| \leq Ch^q, \quad (2.40)$$

where C is independent of h and n.

For example, it can be shown that the Euler-Maruyama method (2.36) is of strong $1/2$ while the Milstein method (2.37) is of strong order 1. However, they share the same weak order 1 of convergence. A very useful result is that for Lipschitz test functions ϕ, if a method has strong order r this implies that it is of weak order r as well.

For ergodic SDEs, i.e., when (2.31) has a unique invariant measure μ satisfying for each test function ϕ and for any deterministic initial condition $X_0 = x$,

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(X(s))ds = \int_{\mathbb{R}^d} \phi(y)d\mu(y), \qquad \text{almost surely,} \tag{2.41}$$

one is interested in approximating numerically the long-time dynamics and to design numerical schemes with a unique invariant measure such that

$$\left| \lim_{N\to\infty} \frac{1}{N+1}\sum_{n=0}^{N} \phi(X_n) - \int_{\mathbb{R}^d} \phi(y)d\mu(y) \right| \leq Ch^p, \tag{2.42}$$

where C is independent of h small enough and X_0. In such a situation, we say that the numerical scheme has order p with respect to the invariant measure. This type of convergence measures the rate at which the numerical method approximates the law of the solution (or the steady state) over longtime. For instance, the Euler-Maruyama method has order 1 with respect to the invariant measure. In fact, if a method converges weakly with order q, then it is of order q with respect to the invariant measure for ergodic SDEs, i.e., with the notations of the above definitions, $p \geq q \geq r$. In Chapter 3, we will use the idea of postprocessors to construct improved integrators in order to approximate the invariant measure of ergodic SDEs with higher order of convergence with respect to the invariant measure and with negligible overcost. To make the reader familiar with postprocessing techniques for SDEs, we refer to the paper [56].

2.2.3.2 Mean square stability

This section is dedicated to introduce the notions of stochastic stability. We consider again the SDE (2.31) where zero is an equilibrium, i.e. $f(0) = g(0) = 0$. The idea of stability in the context of ODE does not generalize straightforwardly to the SDE case, and we should say more precisely what we mean by "$\lim_{t\to\infty} X(t)$" . The following definitions introduce the two main kinds of stochastic stability.

Definition 2.2.6. *The solution $X(t)$ of the equation (2.31) is said to be:*

- *stochastically asymptotically stable (AS) if there exists $\delta > 0$ such that, for all initial conditions X_0 satisfying $\|X_0\| \leq \delta$, we have*

$$\lim_{t\to\infty} \|X(t)\| = 0 \quad \textit{almust surely.} \tag{2.43}$$

- *mean square stable (MS) if there exists $\delta > 0$ such that, for all initial conditions X_0 satisfying $\|X_0\| \leq \delta$, we have*

$$\lim_{t\to\infty} \mathbb{E}(X(t)^2) = 0. \tag{2.44}$$

Following the idea of stability of deterministic ODEs, we consider again the scalar linear test equation

$$dX = \lambda X dt + \mu X dW, \quad X(0) = X_0, \quad 0 \leq t \leq T, \tag{2.45}$$

and we recall that its true solution is

$$X(t) = X_0 \exp((\lambda - \frac{1}{2}\mu^2)t + \mu W(t)). \tag{2.46}$$

It can be proved that the solution (2.46) is *asymptotically stable* if and only if

$$\Re\{\lambda - \frac{1}{2}\mu^2\} < 0, \tag{2.47}$$

and it is *mean square stable* if and only if

$$\Re\{\lambda\} + \frac{1}{2}|\mu|^2 < 0. \tag{2.48}$$

One can easily see that if (2.45) is mean square stable then it is automatically asymptotically stable. Thus, throughout our study we will focus on the mean square stability which turns out to be more required in practice. Now it is time to study the mean square stability of numerical methods.

Definition 2.2.7. *We say that a numerical method* $\{X_n\}_{n\geq 0}$ *applied to the linear test problem* (2.45) *is mean square stable if and only if* $\lim_{n\to\infty} \mathbb{E}(|X_n|^2) = 0$.

Applying a one-step method to the linear SDE (2.45) with $\mathbb{E}(|X_0|^2) < \delta$ for some $\delta > 0$, yields to a recurrence of the form

$$X_{n+1} = R(p, q, \xi_n)X_n, \tag{2.49}$$

where $p = \lambda h$, $q = \mu\sqrt{h}$, and $\xi_n \sim \mathcal{N}(0, 1)$. The rational function $R(p, q, \xi)$ is called the *stability function* of the method. Note that $\sqrt{h}\xi_n$ represents $\Delta W_n \sim \mathcal{N}(0, h)$. For example, the Euler-Matuyama method (2.36), which can be written as

$$X_{n+1} = X_n + h\lambda X_n + \mu\sqrt{h}\xi_n X_n,$$

admits $R(p, q, \xi) = 1 + p + q\xi$ as stability function, whereas that of Milstein method is $R(p, q, \xi) = 1 + p + q\xi + \frac{1}{2}q^2(\xi^2 - 1)$. Using the independence of ξ_n and X_n, and taking the expected value of (2.49) on both sides, we can write

$$\mathbb{E}(|X_{n+1}|^2) = \mathbb{E}(|R(p, q, \xi_n)|^2)\mathbb{E}(|X_n|^2),$$

and by induction we get

$$\mathbb{E}(|X_n|^2) = \mathbb{E}(|R(p, q, \xi_n)|^2)^n\mathbb{E}(|X_0|^2).$$

Naturally we define the numerical mean square *stability domain* of a one-step method as

$$\mathcal{S}^{MS}_{num} = \{(p, q) \in \mathbb{C}^2 \ ; \ |R(p, q, \xi)| < 1\}. \tag{2.50}$$

Definition 2.2.8. *A numerical method* $\{X_n\}_{n\geq 0}$ *is said to be mean square A-stable if*

$$\mathcal{S}^{MS}_{exact} \subset \mathcal{S}^{MS}_{num},$$

where $\mathcal{S}^{MS}_{exact} = \{(p, q) \in \mathbb{C}^2 \ ; \ \Re\{p\} + \frac{1}{2}|q|^2 < 0\}$.

Theorem 2.2.9 (Higham 00')**.** *The* $\theta-$*method* (2.38) *is mean square A-stable if and only if* $\theta \geq \frac{1}{2}$.

2.2.3.3 Monte Carlo method

Consider and SDE with unknown process $X : [0,T] \to \mathbb{R}^d$, which we would like to approximate using a numerical integrator with time step $h = T/N$, where T is the final time and N is the number of steps. In order to approximate expected values of the form $\mathbb{E}(\phi(X(T)))$, we calculate a large number M of independent trajectories $\{X_N^k\}_{n=0,\dots,N}^{k=1,\dots,M}$, then we approximate the average of $\phi(X_N)$, using the law of large numbers

$$\mathbb{E}(\phi(X(T))) \approx \mathbb{E}(\phi(X_N)) \approx \frac{1}{M}\sum_{k=1}^{M}\phi(X_N^k). \tag{2.51}$$

This method for approximating the expected value of a random variable is called Monte Carlo method. The central limit theorem implies that the error of this approximation is $\mathcal{O}(1/\sqrt{M})$. For instance, if the numerical integrator used to calculate the trajectories is of weak order q the weak error will look like

$$\begin{aligned}\left|\mathbb{E}(\phi(X(t_n))) - \frac{1}{M}\sum_{k=1}^{M}\phi(X_n^k)\right| &\le |\mathbb{E}(\phi(X(t_n))) - \mathbb{E}(\phi(X_N))| \\ &\quad + \left|\mathbb{E}(\phi(X_n)) - \frac{1}{M}\sum_{k=1}^{M}\phi(X_N^k)\right| \\ &= \mathcal{O}(h^q + 1/\sqrt{M}).\end{aligned}$$

To reach an accuracy ε, supposing that we use the Euler-Maruyama method (weak order 1) as our numerical integrator, we set $h = 1/\sqrt{M} = \varepsilon$, the total cost of the above approximation is

$$\frac{MT}{h} = \mathcal{O}\left(\frac{M}{h}\right) = \mathcal{O}(\varepsilon^{-3}).$$

In order to reduce this cost, we use variance reduction techniques. Giles introduced in [26] a variance reduction technique called Multilevel Monte Carlo method (MLMC) to reduce the computational complexity of estimating an expected value arising from a stochastic differential equation using Monte Carlo path simulations. In [6], Abdulle and Blumenthal consider a stabilized MLMC method by coupling standard MLMC with explicit stabilized integrators. The total cost using MLMC reduces to $\mathcal{O}(\varepsilon^2(\log\varepsilon)^2)$.

Chapter 3

Optimal explicit stabilized integrators for stiff and ergodic SDEs

Note: Sections 3.1 to 3.6 are quoted identically from the paper [5] in collaboration with Assyr Abdulle and Gilles Vilmart, while Section 3.7 discusses possible extensions and outlook.

3.1 Introduction

We consider Itô systems of stochastic differential equations of the form

$$dX(t) = f(X(t))dt + \sum_{r=1}^{m} g^r(X(t))dW_r(t), \qquad X(0) = X_0 \tag{3.1}$$

where $X(t)$ is a stochastic process with values in $\mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}^d$ is the drift term, $g^r : \mathbb{R}^d \to \mathbb{R}^d$, $r = 1, \ldots, m$ are the diffusion terms, and $W_r(t)$, $r = 1, \ldots, m$, are independent one-dimensional Weiner processes fulfilling the usual assumptions. We assume that the drift and diffusion functions are smooth enough and Lipschitz continuous to ensure the existence and uniqueness of a solution of (3.1) on a given time interval $(0, T)$. We consider autonomous problems to simplify the presentation, but we emphasise that the scheme can also be extended to non-autonomous SDEs. A one step numerical integrator for the approximation of (3.1) at time $t = nh$ is a discrete dynamical system of the form

$$X_{n+1} = \Psi(X_n, h, \xi_n) \tag{3.2}$$

where h denotes the stepsize and ξ_n are independent random vectors. Analogously to the deterministic case, standard explicit numerical schemes for stiff stochastic problems, such as the simplest Euler-Maruyama method defined as

$$X_{n+1} = X_n + hf(X_n) + \sum_{r=1}^{m} g^r(X_n)\Delta W_{n,r}, \qquad X(0) = X_0, \tag{3.3}$$

where $\Delta W_{n,r} = W_r(t_{n+1}) - W_r(t_n)$ are the Brownian increments, face a severe timestep restriction [33, 29, 36], and one can use an implicit or semi-implicit scheme with favorable

stability properties. In particular, it is shown in [33] that the implicit θ-method of weak order one is mean-square A-stable if and only if $\theta \geq 1/2$, while weak order two mean-square A-stable are constructed in [11]. An alternative approach is to consider explicit stabilized schemes with extended stability domains, as proposed in [7, 8]. In [8] the deterministic Chebyshev method is extended to the context of mean-square stiff stochastic differential equations with Itô noise, while the Stratonovitch noise case is treated in [7]. In place of a standard small damping, the main idea in [7, 8] is to use a large damping parameter η optimized for each number s of stages to stabilize the noise term. This yields a family of Runge-Kutta type schemes with extended stability domain with size $L_s \simeq 0.33s^2$. This stability domain size was improved to $L_s \simeq 0.42s^2$ in [12] where a family of weak second order stabilized schemes (and strong order one under suitable assumptions) is constructed based on the deterministic ROCK2 method [9].

For ergodic SDEs, i.e., when (3.1) has a unique invariant measure μ satisfying for each test function ϕ and for any deterministic initial condition $X_0 = x$,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \phi(X(s))ds = \int_{\mathbb{R}^d} \phi(y)d\mu(y), \qquad \text{almost surely,} \tag{3.4}$$

one is interested in approximating numerically the long-time dynamics and to design numerical scheme with a unique invariant measure such that

$$\left| \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \phi(X_n) - \int_{\mathbb{R}^d} \phi(y)d\mu(y) \right| \leq Ch^r, \tag{3.5}$$

where C is independent of h small enough and X_0. In such a situation, we say that the numerical scheme has order r with respect to the invariant measure. For instance, the Euler-Maruyama method has order 1 with respect to the invariant measure. In [39] the following non-Markovian scheme with the same cost as the Euler-Maruyama method was proposed for Brownian dynamics, i.e where the vector field is a gradient $f(x) = -\nabla V(x)$ and the noise is additive ($g(x) = \sigma$),

$$X_{n+1} = X_n + hf(X_n) + \sigma \frac{\Delta W_{n,j} + \Delta W_{n+1,j}}{2}, \qquad X(0) = X_0, \tag{3.6}$$

and it was shown in [40] that (3.6) has order 2 with respect to the invariant measure for Brownian dynamics. However, the admissible stepsizes for such an explicit method to be stable may face a severe restriction and alternatively to switching to drift-implicit methods, one may ask if a stabilized version of such an attractive non-Markovian scheme exists.

In this chapter we introduce a new family of explicit stabilized schemes with optimal mean-square stability domain of size $L_s = Cs^2$, where $C \geq 2 - \frac{4}{3}\eta$ and $\eta \geq 0$ is a small parameter. We emphasize that in the deterministic case, $L_s = 2s^2$ is the largest, i.e. optimal, stability domain along the negative real axis for an explicit s-stage Runge-Kutta method [29]. We note that the Chebyshev method (3.8) (with $\eta = 0$) realizes such an optimal stability domain. The new schemes have strong order $1/2$ and weak order 1. The main ingredient for the design of the new schemes is to consider second kind Chebyshev polynomials, in addition to the usual first kind Chebyshev polynomials involved in the deterministic Chebychev method and stochastic extensions [8, 7]. For stiff stochastic problems, the stability domain sizes are close to the optimal value $2s^2$ and in the deterministic

setting the method coincide with the optimal first order explicit stabilized method. Thus these methods are more efficient than previously introduced stochastic stabilized methods [8, 12]. For ergodic dynamical systems, in the context of the ergodic Brownian dynamics, the new family of explicit stabilized schemes allows for a postprocessing [56] (see also [17, 38] in the context of Runge-Kutta methods) to achieve order two of accuracy for sampling the invariant measure. In this context, our new methods can be seen as a stabilized version of the non-Markovian scheme (3.6) introduced in [39, 40].

This chapter is organized as follows. In Section 3.2, we introduce the new family of schemes with optimal stability domain and we recall the main tools for the study of stiff integrators in the mean-square sense. We then analyze its mean-square stability properties (Section 3.3), and convergence properties (Section 3.4). In Section 3.5, using a postprocessor we present a modification with negligible overcost that yields order two of accuracy for the invariant measure of a class of ergodic overdamped Langevin equation. Finally, Section 3.6 is dedicated to the numerical experiments that confirm our theoretical analysis and illustrate the efficiency of the new schemes.

3.2 New second kind Chebyshev methods

In this section we introduce our new stabilized stochastic method. We first briefly recall the concept of stabilized methods. In the context of ordinary differential equations (ODEs),

$$\frac{dX(t)}{dt} = f(X(t)), \qquad X(0) = X_0, \tag{3.7}$$

and the Euler method $X_1 = X_0 + hf(X(0))$, a stabilization procedure based on recurrence formula has been introduced in [54]. Its construction relies on Chebyshev polynomials (hence the alternative name "Chebyshev methods"), $T_s(\cos x) = \cos(sx)$ and it is based on the explicit s-stage Runge-Kutta method

$$\begin{aligned} K_0 &= X_0, \quad K_1 = X_0 + h\mu_1 f(K_0), \\ K_i &= \mu_i h f(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad j = 2, \dots, s, \\ X_1 &= K_s, \end{aligned} \tag{3.8}$$

where

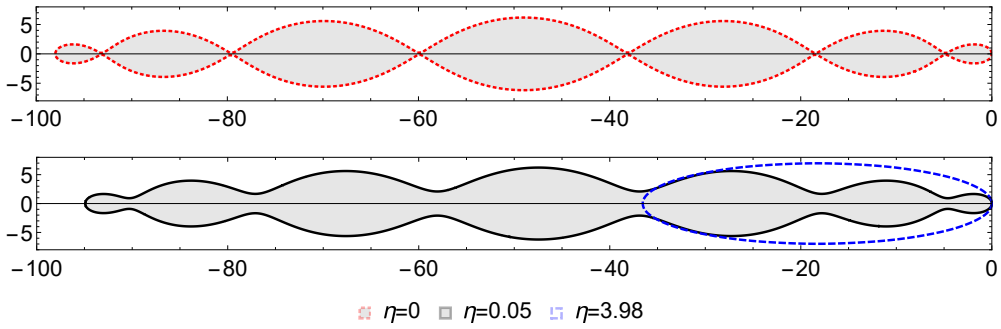
$$\omega_0 = 1 + \frac{\eta}{s^2}, \quad \omega_1 = \frac{T_s(\omega_0)}{T_s'(\omega_0)}, \quad \mu_1 = \frac{\omega_1}{\omega_0}, \tag{3.9}$$

and for all $i = 2, \dots, s$,

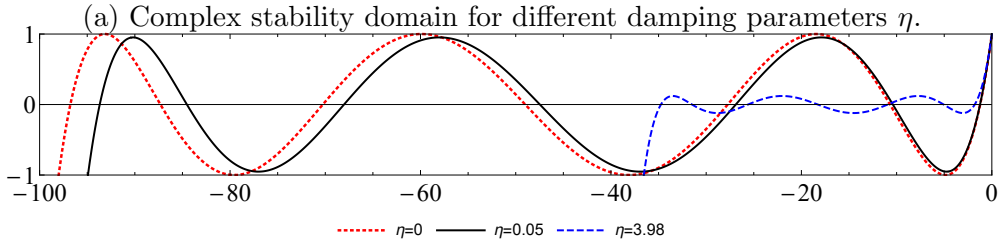
$$\mu_i = \frac{2\omega_1 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad \nu_i = \frac{2\omega_0 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad \kappa_i = -\frac{T_{i-2}(\omega_0)}{T_i(\omega_0)} = 1 - \nu_i. \tag{3.10}$$

One can easily check that the (family) of methods (3.8) has the same first order accuracy as the Euler method (recovered for $s = 1$). In addition, the scheme (3.8) has a low memory requirement (only two stages should be stored when applying the recurrence formula) and it has a good internal stability with respect to round-off errors [54]. The attractive feature of such a scheme comes from its stability behavior. Indeed, the method (3.8) applied to the linear test problem $dX(t)/dt = \lambda X(t)$ yields, using the recurrence relation

$$T_j(p) = 2pT_{j-1}(p) - T_{j-2}(p), \tag{3.11}$$



(a) Complex stability domain for different damping parameters η.



(b) Corresponding stability functions $R_{7,\eta}(p)$.

Figure 3.1: Stability domains and stability functions of the deterministic Chebyshev method for $s = 7$ and different damping values $\eta = 0, 0.05, 3.98$.

where $T_0(p) = 1, T_1(p) = p$, with $p = \lambda h$,

$$X_1 = R_{s,\eta}(p)X_0 = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)}X_0, \tag{3.12}$$

where the dependence of the stability function $R_{s,\eta}$ on the parameters s and η is emphasized with a corresponding subscript. The real negative interval $(-C_s(\eta) \cdot s^2, 0)$ is included in the stability domain of the method

$$\mathcal{S} := \{p \in \mathbb{C}; |R_{s,\eta}(p)| \leq 1\}. \tag{3.13}$$

The constant $C_s(\eta) \simeq 2 - 4/3\,\eta$ depends on the so-called damping parameter η and for $\eta = 0$, it reaches the maximal value $C_s(0) = 2$. Hence, given the stepsize h, for systems with a Jacobian having large real negative eigenvalues (such as diffusion problems) with spectral radius $\lambda_{\max}$ at X_n, the parameter s for the next step X_{n+1} can be chosen adaptively as[1]

$$s = \left[\sqrt{\frac{h\lambda_{\max} + 1.5}{2 - 4/3\,\eta}} + 0.5\right], \tag{3.14}$$

see [3] in the context of deterministically stabilized schemes of order two with adaptative stepsizes. The method (3.8) is much more efficient as its stability domain increases *quadratically* with the number s of function evaluations while a composition of s explicit Euler steps (same cost) has a stability domain that only increases *linearly* with s. In Figure 3.1(a) we plot the complex stability domain $\{p \in \mathbb{C}\,;\, |R_{s,\eta}(p)| \leq 1\}$ for $s = 7$ stages and different values $\eta = 0$, $\eta = 0.05$ and $\eta = 3.98$, respectively. We also plot in Figure 3.1(b) the corresponding stability function $R_{s,\eta}(p)$ as a function of p real, to illustrate

[1]The notation $[x]$ stands for the integer rounding of real numbers.

that the stability domain along the negative real axis corresponds to the values for which $|R_{s,\eta}(p)| \leq 1$. We observe that in the absence of damping ($\eta = 0$), the stability domain includes the large real interval $[-2 \cdot s^2, 0]$ of width $2 \cdot 7^2 = 98$. However for all p that are a local extrema of the stability function, where $|R_{s,\eta}(p)| = 1$, the stability domain is very thin and does not include a neighbourhood close to the negative real axis. To make the scheme robust with respect to small perturbations of the eigenvalues, it is therefore needed to add some damping and a typical value is $\eta = 0.05$, see for instance the reviews [55, 4]. The advantage is that the stability domain now includes a neighbourhood of the negative real axis portion. The price of this improvement is a slight reduction of the stability domain size $C_\eta s^2$, where $C_\eta \simeq 2 - \frac{4}{3}\eta$. Chebyshev methods have been first generalized for Itô SDEs in [8] (see [7] for Stratonovitch SDEs) with the following scheme denoted S-ROCK,[2]

$$\begin{aligned} K_0 &= X_0 \\ K_1 &= X_0 + \mu_1 h f(X_0) \\ K_i &= \mu_i h f(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad i = 2, \ldots, s, \\ X_1 &= K_s + \sum_{r=1}^{m} g^r(K_s) \Delta W_r, \end{aligned} \tag{3.15}$$

where the coefficients μ_i, ν_i, κ_i are defined in (3.9),(3.10). In contrast to the deterministic method (3.8), where η is chosen small and fixed (typically $\eta = 0.05$), in stochastic case for the classical S-ROCK method [8], the damping $\eta = \eta_s$ is not small and chosen as an increasing function of s that plays a crucial role in stabilizing the noise and in obtaining an increasing portion of the true stability domain (3.19) as s increases.

In the context of stiff SDEs, a relevant stability concept is that of mean-square stability. A test problem widely used in the literature is [49, 33, 19, 53] ,

$$dX(t) = \lambda X(t) dt + \mu X(t) dW(t), \qquad X(0) = 1, \tag{3.16}$$

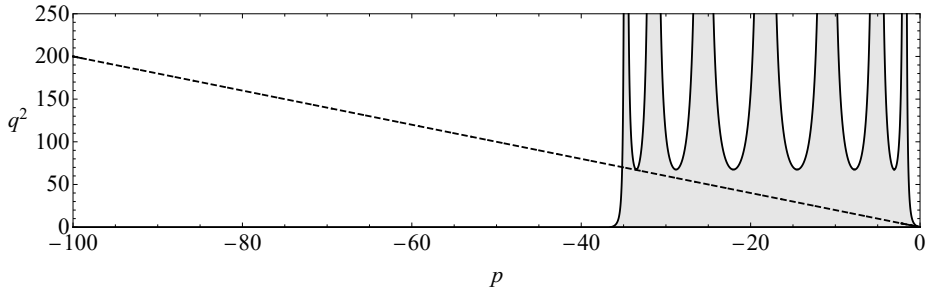
in dimensions $d = m = 1$ with fixed complex parameters λ, μ. Note that other stability test problems in multiple dimensions are also considered in [18] and references therein. The exact solution of (3.16) is called mean-square stable if $\lim_{t\to\infty} \mathbb{E}\big(|X(t)|^2\big) = 0$ and this holds if and only if $(\lambda, \mu) \in \mathcal{S}^{MS}$, where

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

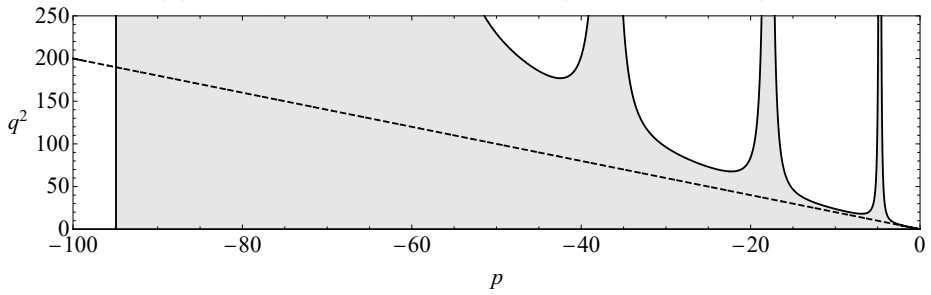
Indeed, the exact solution of (3.16) is given by $X(t) = \exp((\lambda - \frac{1}{2}\mu^2)t + \mu W(t))$, and an application of the the Itô formula yields $\mathbb{E}(|X(t)|^2) = \exp((\Re(\lambda) + \frac{1}{2}\mu^2)t)$ which tends to zero at infinity if and only if $\Re(\lambda) + \frac{1}{2}\mu^2 < 0$. We say that a numerical scheme $\{X_n\}$ for the test problem (3.16) is mean-square stable if and only if $\lim_{n\to\infty} \mathbb{E}(|X_n|^2) = 0$. For a one-step integrator applied to the test SDE (3.16), we obtain in general a induction of the form

$$X_{n+1} = R(p, q, \xi_n) X_n, \tag{3.17}$$

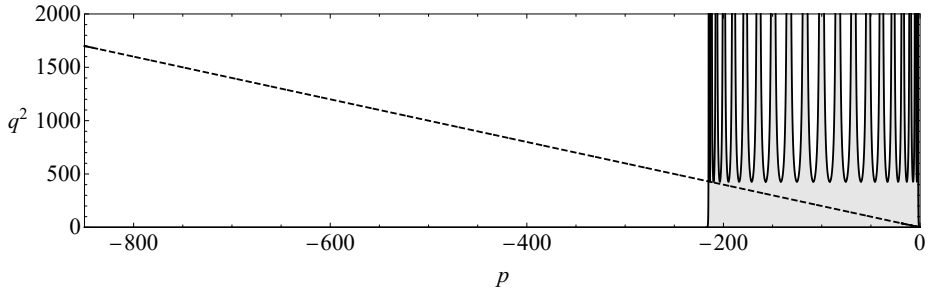
[2] A variant with analogous stability properties is proposed in [8] with $g^r(K_s)$ replaced by $g^r(K_{s-1})$ in (3.15).



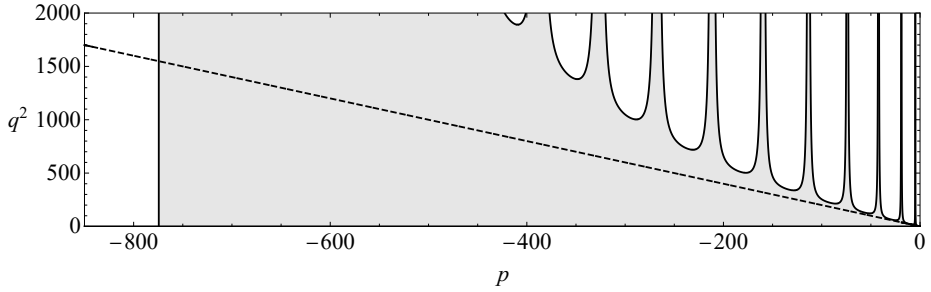
(a) Standard S-ROCK method ($s = 7$, $\eta = 3.98$).



(b) New SK-ROCK method ($s = 7$, $\eta = 0.05$).



(c) Standard S-ROCK method ($s = 20$, $\eta = 6.95$).



(d) New SK-ROCK method ($s = 20$, $\eta = 0.05$).

Figure 3.2: Mean-square stability domains of the standard and new stochastic Chebyshev methods in the p–q^2 plane for $s = 7, 20$ stages, respectively. The dashed lines corresponds to the upper boundary $q^2 = -2p$ of the real mean-square stability domain $\mathcal{S} \cap \mathbb{R}^2$ of the exact solution.

where $p = \lambda h, q = \mu\sqrt{h}$, and ξ_n is a random variable (e.g. a Gaussian $\xi_n \sim \mathcal{N}(0,1)$ or a discrete random variable). Using $\mathbb{E}(|X_{n+1}|^2) = \mathbb{E}(|R(p,q,\xi_n)|^2)\mathbb{E}(|X_n|^2)$, we obtain the mean-square stability condition [49, 33]

$$\lim_{n\to\infty} \mathbb{E}(|X_n|^2) = 0 \iff (p,q) \in \mathcal{S}_{num}, \tag{3.18}$$

where we define $\mathcal{S}_{num} = \{(p,q) \in \mathbb{C}^2 \, ; \mathbb{E}|R(p,q,\xi)|^2 < 1\}$. The function $R(p,q,\xi_n)$ is called the stability function of the one-step integrator. For instance, the stability function of the Euler-Maruyama method (3.3) reads $R(p,q,\xi) = 1 + p + q\xi$ and we have $\mathbb{E}(|R(p,q,\xi)|^2) = (1+p)^2 + q^2$.

We say that a numerical integrator is mean-square A-stable if $\mathcal{S}^{MS} \subset \mathcal{S}_{num}$. This means that the numerical scheme applied to (3.16) is mean-square stable for all all $h > 0$ and all $(\lambda, \mu) \in \mathcal{S}^{MS}$ for which the exact solution of (3.16) is mean-square stable. An explicit Runge-Kutta type scheme cannot however be mean-square stable because its stability domain $\mathcal{S}_{num}$ is necessary bounded along the p-axis. Following [7, 8], we consider the following portion of the true mean-square stability domain

$$\mathcal{S}_a = \{(p,q) \in (-a, 0) \times \mathbb{R} \; ; \; p + \frac{1}{2}|q|^2 < 0\}, \tag{3.19}$$

and define for a given method

$$L = \sup\{a > 0 \; ; \; \mathcal{S}_a \subset \mathcal{S}_{num}\}. \tag{3.20}$$

We search for explicit schemes for which the length L of the stability domain is large. For example, for the classical S-ROCK method [8], the value $\eta = 3.98$ is the optimal damping maximising L for $s = 7$ stages and we can see in Figure 3.1 that this damping reduces significantly the stability domain compared to the optimal deterministic domain.

The new S-ROCK method, denoted SK-ROCK (for stochastic second kind orthogonal Runge-Kutta-Chebyshev method) introduced in this chapter is defined as

$$\begin{aligned} K_0 &= X_0 \\ K_1 &= X_0 + \mu_1 h f(X_0 + \nu_1 Q) + \kappa_1 Q \\ K_i &= \mu_i h f(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad i = 2, \ldots, s. \\ X_1 &= K_s, \end{aligned} \tag{3.21}$$

where $Q = \sum_{r=1}^m g^r(X_0)\Delta W_r$, and $\mu_1 = \omega_1/\omega_0, \nu_1 = s\omega_1/2, \kappa_1 = s\omega_1/\omega_0$ and $\mu_i, \nu_i, \kappa_i, \; i = 2, \ldots, s$ are given by (3.10), with a fixed small damping parameter η. In the absence of noise ($g^r = 0, r = 1, \ldots, m$, deterministic case), this method coincides with the standard deterministic order 1 Chebychev method, see the review [4]. We observe that the new class of methods (3.21) is closely related to the standard S-ROCK method (3.15). Comparing the two schemes (3.21) and (3.15), the two differences are on the one hand that the noise term is computed at the first internal stage K_1 for (3.21), whereas it is computed at the final stage in (3.15), and on the other hand, for the new method (3.21) the damping parameter η involved in (3.9) is fixed and small independently of s (typically $\eta = 0.05$), whereas for the standard method (3.15), the damping η is an increasing function of s, optimized numerically for each number of stages s.

If we apply the above scheme (3.21) to the linear test equation (3.16), we obtain

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

where

$$\mathbb{E}(|R(p,q,\xi)|^2) = A(p)^2 + B(p)^2 q^2, \tag{3.22}$$

and

$$A(p) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} \qquad B(p) = \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)}(1 + \frac{\omega_1}{2}p)$$

correspond to the drift and diffusion contributions, respectively. The above stability function (see Lemma 3.3.1 in Section 3.3) is obtained by using the recurrence relation for the first kind Chebyshev polynomials (3.11) and the similar recurrence relation for the second kind Chebyshev polynomials

$$U_j(p) = 2pU_{j-1}(p) - U_{j-2}(p), \tag{3.23}$$

where $U_0(p) = 1, U_1(p) = 2p$. Notice that the relation $T_s'(p) = sU_{s-1}(p)$ between first and second kind Chebyshev polynomials will be repeatedly used in our analysis.

In Figure 3.2(b)(d), we plot the mean-square stability domain of the SK-ROCK method for $s = 7$ and $s = 20$ stages, respectively and the same small damping $\eta = 0.05$ as for the deterministic Chebyshev method. We observe that the stability domain has length $L_s \simeq (2 - \frac{4}{3}\eta)s^2$. For comparison, we also include in Figure 3.2(a)(c) the mean-square stability domain of the standard S-ROCK method with smaller stability domain size $L_s \simeq 0.33 \cdot s^2$.

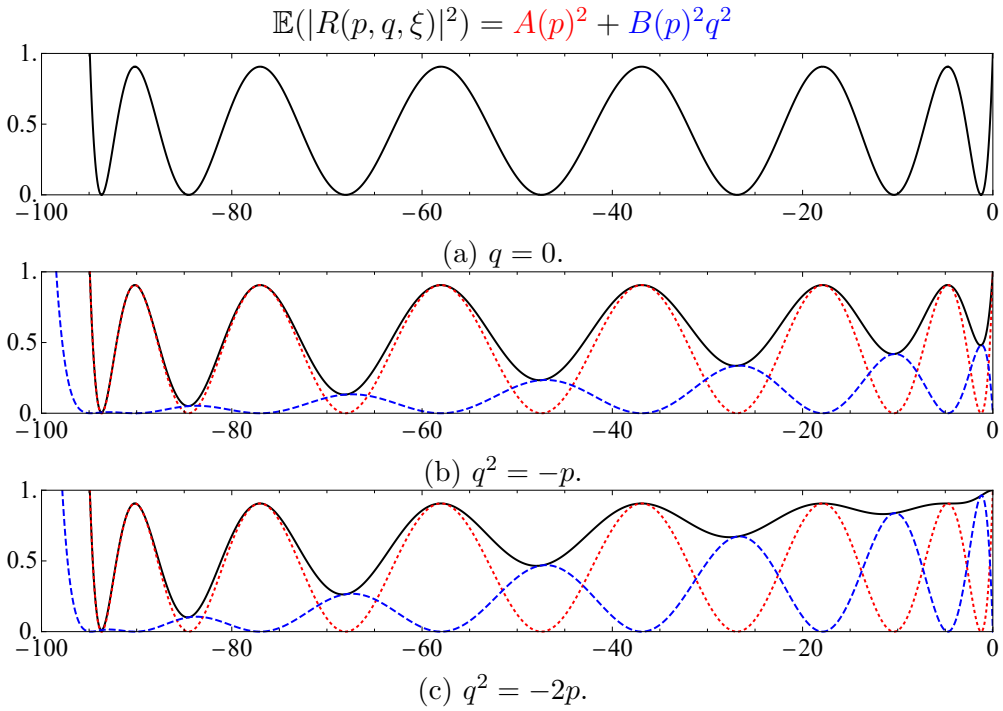
In Figure 3.3, we plot the stability function $\mathbb{E}(|R(p, q, \xi)|^2)$ in (3.22) as a function of p for various scaling of the noise for $s = 7$ stages and damping $\eta = 0.05$. We see that it is bounded by 1 for $p \in (-2(1 - \frac{2}{3}\eta)s^2, 0)$ which is proved asymptotically in Theorem 3.3.2. The case $q = 0$ corresponds to the deterministic case, and we see in Figure 3.3(a), the polynomial $\mathbb{E}(|R(p, 0, \xi)|^2) = A(p)^2$. Noticing that $\mathbb{E}(|R(p, q, \xi)|^2)$ is an increasing function of q, the case $q^2 = -2p$ represented in Figure 3.3(c) corresponds to the upper border of the stability domain $\mathcal{S}_L$ defined in (3.19) (note that this is the stability function value along the dashed boundary in Figure 3.2), while the scaling $q^2 = -p$ in Figure 3.3(c) is an intermediate regime. In Figures 3.3(b)(c), we also include the drift function $A(p)^2$ (red dotted lines) and diffusion function $B(p)^2q^2$ (blue dashed lines), and it can be observed that their oscillations alternate, which means that any local maxima of one function is close to a zero of the other function. This is not surprising because $A(p)$ and $B(p)$ are related to the first kind and second kind Chebyshev polynomials, respectively, corresponding to the cosine and sine functions. This also explains how a large mean-square stability domain can be achieved by the new SK-ROCK method (3.21) with a small damping parameter η, in contrast to the standard S-ROCK method (3.15) from [8] that uses a large and s-dependent damping parameter η with smaller stability domain size $L_s \simeq 0.33 \cdot s^2$(see Figures 3.3(a)(c)).

3.3 Mean-square stability analysis

In this section, we prove asymptotically that the new SK-ROCK methods have an extended mean-square stability domain with size Cs^2 growing quadratically as a function of the number of internal stages s, where the constant $C \geq 2 - \frac{4}{3}\eta$ is the same as the optimal constant of the standard Chebyshev method in the deterministic case, using a fixed and small damping parameter η.

Lemma 3.3.1. *Let $s \geq 1$ and $\eta \geq 0$. Applied to the linear test equation*

$$dX = \lambda X dt + \mu X dW,$$



(a) $q = 0$.

(b) $q^2 = -p$.

(c) $q^2 = -2p$.

Figure 3.3: Stability function (3.22) of the new SK-ROCK method as a function of p in (solid black lines), for $s = 7, \eta = 0.05$ and various noise scalings $q^2 = 0, -p, -2p$, respectively. We also include the drift contribution $A(p)^2$ (red dotted lines) and diffusion contribution $B(p)^2q^2$ (blue dashed lines).

the scheme (3.21) *yields*

$$X_{n+1} = R(\lambda h, \mu\sqrt{h}, \xi_n)X_n$$

where $p = \lambda h, q = \mu\sqrt{h}$, $\xi_n \sim \mathcal{N}(0,1)$ *is a Gaussian variable and the stability function given by*

$$R(p,q,\xi) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} + \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)}(1 + \frac{\omega_1}{2}p)q\xi. \tag{3.24}$$

Proof. Indeed, we take advantage that T_j and U_j have the same recurrence relations (3.11),(3.23), and only the initialization changes with $T_1(x) = x$ and $U_1(x) = 2x$, we deduce $Q = X_0\mu\sqrt{h}\xi$, and we obtain by induction on $i \geq 1$,

$$K_i = \frac{T_i(\omega_0 + \omega_1 p)}{T_i(\omega_0)}X_0 + \frac{U_{i-1}(\omega_0 + \omega_1 p)}{T_i(\omega_0)}(1 + \frac{\omega_1 p}{2})s\omega_1 Q$$

and we use $T_s'(x) = xU_{s-1}(x)$ and $s\omega_1/T_s(\omega_0) = 1/U_{s-1}(\omega_0)$, which yields the result for $X_1 = K_s$. □

For a positive damping η, we prove the following main result of this section, showing that a quadratic growth $L \geq (2 - 4/3\,\eta)s^2$ of the mean-square stability domain defined in (3.20) is achieved for all η small enough and all stage number s large enough.

Theorem 3.3.2. *There exists* $\eta_0 > 0$ *and* s_0 *such that for all* $\eta \in [0, \eta_0]$ *and all* $s \geq s_0$, *for all* $p \in [-2\omega_1^{-1}, 0]$ *and* $p + \frac{1}{2}|q|^2 \leq 0$, *we have* $\mathbb{E}(|R(p,q,\xi)|^2) \leq 1$.

Remark 3.3.3. *We deduce from Theorem 3.3.2, that the mean-square stability domain size* (3.20) *of SK-ROCK grows as* $(2 - 4/3\,\eta)s^2$ *which is arbitrarily close to the optimal stability domain size* $2s^2$ *for* $\eta \to 0$*. Indeed, for* $s \to \infty$ *and all* $\eta \le \eta_0$*, we have*

$$2\omega_1^{-1}s^{-2} \to 2\frac{\tanh(\sqrt{2\eta})}{\sqrt{2\eta}} = 2 - 4/3\,\eta + \mathcal{O}(\eta^2)$$

and for all s, η*, we have* $2\omega_1^{-1} \ge (2 - 4/3\,\eta)s^2$*. In addition, in the special case of a zero damping* ($\eta = 0$)*, the stability function* (3.24) *reduces to*

$$R(p, q, \xi) = T_s(1 + \frac{p}{s^2}) + s^{-1}U_{s-1}(1 + \frac{p}{s^2})(1 + \frac{p}{2s^2})q\xi,$$

and it holds

$$\mathbb{E}(|R(p, q, \xi)|^2) \le 1,$$

for all $s \ge 1$*, for all* $p \in [-2s^2, 0]$ *and all* $q \in \mathbb{C}$ *such that* $p + |q|^2/2 \le 0$*. Indeed, for* $p \in [-2s^2, 0]$*, we denote* $\cos\theta = 1 + \frac{p}{s^2} \in [-1, 1]$ *and using*

$$T_s(\cos(\theta)) = \cos(s\theta), \qquad \sin(\theta)U_{s-1}(\cos(\theta)) = \sin(s\theta),$$

we obtain

$$\mathbb{E}(|R(p, q, \xi)|^2) \le \mathbb{E}(|R(p, \sqrt{-2p}) = \cos(s\theta)^2 + \sin(s\theta)^2\frac{1 + \cos\theta}{2} \le 1,$$

where we used $-2p = 2s^2(1 - \cos\theta)$*,* $1 + \frac{p}{2s^2} = \frac{1+\cos\theta}{2}$ *and* $\sin^2\theta = (1 + \cos\theta)(1 - \cos\theta)$*.*

Before we prove Theorem 3.3.2, we have the following lemma, see [54] for analogous results.

Lemma 3.3.4. *We have the following convergences as* $s \to \infty$ *to analytic functions*[3] *uniformly for* z *in any bounded set of the complex plan,*

$$T_s(1 + z/s^2) \to \cosh\sqrt{2z},$$
$$s^{-1}U_{s-1}(1 + z/s^2) \to \alpha(z) := \frac{\sinh\sqrt{2z}}{\sqrt{2z}},$$
$$\omega_1 s^2 \to \Omega(\eta)^{-1}, \quad \Omega(\eta) := \frac{\tanh\sqrt{2\eta}}{\sqrt{2\eta}}.$$

Proof. We prove the uniform convergence of the first limit only, since it will be useful in the proof of the next theorem. The others can be proved in a similar way.

First, let us write the two functions of η in Taylor series,

$$s^{-1}U_{s-1}(\omega_0) = s^{-1}\sum_{n=0}^{s-1}\frac{U_{s-1}^{(n)}(1)}{n!}(\frac{\eta}{s^2})^n = \sum_{n=0}^{s-1}\left(\frac{1}{n!}\prod_{k=0}^{n}(1 - \frac{k^2}{s^2})\prod_{k=0}^{n}\frac{1}{2k+1}\right)\eta^n,$$
$$\alpha(\eta) = \sum_{n=0}^{\infty}\frac{2^n\eta^n}{(2n+1)!} = \sum_{n=0}^{\infty}\left(\frac{1}{n!}\prod_{k=0}^{n}\frac{1}{2k+1}\right)\eta^n,$$

[3]Note that for $z < 0$, we can use $\sqrt{2z} = i\sqrt{-2z}$ and obtain $T_s(1 + z/s^2) \to \cos(\sqrt{-2z})$ for $s \to \infty$ and similarly $\alpha(z) = \mathrm{sinc}(\sqrt{-2z})$.

where we used the formula $sU_{s-1}^{(n-1)}(1) = T_s^{(n)}(1) = \prod_{k=0}^{n-1} \frac{s^2-k^2}{2k+1}$. Subtracting the above two identities, we deduce

$$\begin{aligned}\sup_{\eta\in[-\eta_0,\eta_0]} \left|s^{-1}U_{s-1}(\omega_0) - \alpha(\eta)\right| &\leq \sum_{n=0}^{s-1} \frac{\eta_0^n}{n!}\left(1 - \prod_{k=0}^{n}(1-\frac{k^2}{s^2})\right)\prod_{k=0}^{n}\frac{1}{2k+1} + \sum_{n=s}^{\infty}\frac{\eta_0^n}{n!} \\ &\leq \sum_{n=0}^{s-1} \frac{\eta_0^n}{n!}\left(1 - \prod_{k=0}^{n}(1-\frac{k^2}{s^2})\right)\frac{1}{2s-1} + \sum_{n=s}^{\infty}\frac{\eta_0^n}{n!} \quad (3.25)\end{aligned}$$

Noticing that $\frac{\eta_0^n}{n!}\left(1-\prod_{k=0}^{n}(1-\frac{k^2}{s^2})\right)\frac{1}{2s-1}$ converges to zero as $s \to \infty$ and is bounded by $\frac{\eta_0^n}{n!}$ for all integers s, n, which is the general term of the convergent series of $\exp(\eta_0) = \sum_{n=0}^{\infty}\frac{\eta_0^n}{n!}$, the Lebesgue dominated convergence theorem implies that (3.25) converges to zero as $s \to \infty$, which concludes the proof. □

Lemma 3.3.5. *For all η small enough and all s large enough, we have the following estimate:*

$$\frac{s^2\omega_1}{T_s(w_0)^2}\frac{1-(1-\omega_1)^2}{1-(\omega_0-\omega_1)^2} \leq 1 \quad (3.26)$$

Proof of Lemma 3.3.5. Using the Lemma 3.3.4 we have for $s \to \infty$, uniformly for all $\eta \in [0, \eta_0]$,

$$\frac{s^2\omega_1}{T_s(\omega_0)^2} \to \frac{2\sqrt{2\eta}}{\sinh(2\sqrt{2\eta})} \text{ and } \frac{1-(1-\omega_1)^2}{1-(\omega_0-\omega_1)^2} \to \frac{1}{1-\Omega(\eta)\eta}.$$

Now if we expand both functions in Taylor series we get:

$$\frac{2\sqrt{2\eta}}{\sinh(2\sqrt{2\eta})} = 1 - \frac{4}{3}\eta + \mathcal{O}(\eta^2), \qquad \frac{1}{1-\Omega(\eta)\eta} = 1 + \eta + \mathcal{O}(\eta^2), \quad (3.27)$$

and this implies that for all s large enough and all $\eta \leq \eta_0$,

$$\frac{s^2\omega_1}{T_s(w_0)^2}\frac{1-(1-\omega_1)^2}{1-(\omega_0-\omega_1)^2} \leq (1 - \frac{4}{3}\eta_0 + \mathcal{O}(\eta_0^2))(1 + \eta_0 + \mathcal{O}(\eta_0^2)) = 1 - \frac{1}{3}\eta_0 + \mathcal{O}(\eta_0^2), \quad (3.28)$$

which is less than 1 for η_0 small enough. □

Remark 3.3.6. *Numerical evidence suggests that the result of Theorem 3.3.2 holds for all $s \geq 1$ and all $\eta \geq 0$. Indeed, it can be checked numerically that* (3.26) *holds for all $\eta \in (0, 1)$ and all $s \geq 1$.*

Proof of Theorem 3.3.2. Setting $x = w_0 + w_1 p$, a calculation yields

$$\begin{aligned}\mathbb{E}(|R(p,q,\xi)|^2) &\leq \mathbb{E}(|R(p,\sqrt{-2p},\xi)|^2) \\ &= \frac{T_s(x)^2}{T_s(w_0)^2} + \frac{U_{s-1}(x)^2}{U_{s-1}(w_0)^2}(1+\frac{w_1}{2}p)^2(-2p)\end{aligned}$$

The proof is conducted in two steps where we treat separately the cases $p \in [-2\omega_1^{-1}, -1]$ and $p \in [-1, 0]$. For the first case $p \in [-2\omega_1^{-1}, -1]$, which corresponds to $x \in [-1+\eta/s^2, \omega_0-\omega_1]$,

we have

$$\begin{aligned}\mathbb{E}(|R(p,q,\xi)|^2) &= \frac{T_s(x)^2}{T_s(w_0)^2} + \frac{U_{s-1}(x)^2}{U_{s-1}(w_0)^2}\left(1-\frac{w_0-x}{2}\right)^2 2\frac{w_0-x}{w_1} \\ &= \frac{T_s(x)^2}{T_s(w_0)^2} + U_{s-1}(x)^2(1-x^2)Q_s(x)\end{aligned}$$

where we denote

$$Q_s(x) = \frac{s^2\omega_1}{T_s(w_0)^2}\left(\frac{1+x-\frac{\eta}{s^2}}{2}\right)\frac{1-(x-\frac{\eta}{s^2})^2}{1-x^2}$$

First, we note that $\frac{1+x-\frac{\eta}{s^2}}{2} \in [0, 1-\frac{\omega_1}{2}]$. Next, using $\frac{\eta}{s^2} \leq 2$, we deduce

$$\frac{d}{dx}\left(\frac{1-(x-\frac{\eta}{s^2})^2}{1-x^2}\right) = \frac{2\eta}{s^2}\frac{1+x^2-\eta/s^2 x}{(1-x^2)^2} \geq \frac{2\eta}{s^2}\frac{(1-x)^2}{(1-x^2)^2} \geq 0.$$

Thus, $\frac{1-(x-\frac{\eta}{2})^2}{1-x^2}$ is an increasing function of x, smaller than its value at $x = \omega_0 - \omega_1$,

$$\frac{1-(x-\frac{\eta}{s^2})^2}{1-x^2} \leq \frac{1-(1-\omega_1)^2}{1-(\omega_0-\omega_1)^2}$$

Using Lemma 3.3.5 we obtain $|Q_s(x)| \leq 1$. This yields $\mathbb{E}(|R(p,q,\xi)|^2) \leq T_s(x)^2 + U_{s-1}(x)^2(1-x^2) = 1$.

For the second case $p \in [-1, 0]$ which corresponds to $x \in [\omega_0 - \omega_1, \omega_0]$, we deduce from $T_s(x)^2 + U_{s-1}(x)^2(1-x^2) = 1$ that

$$\mathbb{E}(|R(p,q,\xi)|^2) \leq \frac{1}{T_s(w_0)^2} + \frac{U_{s-1}(x)^2}{U_{s-1}(w_0)^2}\left((1+\frac{w_1}{2}p)^2(-2p) - \frac{(1-x^2)U_{s-1}(\omega_0)^2}{T_s(w_0)^2}\right)$$

Using Lemma 3.3.4, we get

$$\begin{aligned}\mathbb{E}(|R(p,q,\xi)|^2) &\leq \frac{1}{T_s(w_0)^2} + \frac{U_{s-1}(x)^2}{U_{s-1}(w_0)^2}\left((1+\frac{w_1}{2}p)^2(-2p) - \frac{(1-x^2)U_{s-1}(\omega_0)^2}{T_s(w_0)^2}\right) \\ &\to l(\eta,p) := \frac{1}{\cosh^2\sqrt{2\eta}} + \frac{\alpha(\eta+p/\Omega(\eta))^2}{\alpha(\eta)^2}(-2p(\Omega(\eta)-1) + 2\Omega(\eta)^2\eta),\end{aligned}$$

for $s \to \infty$, where the above convergence is uniform for $p \in [0,1]$,$\eta \leq \eta_0$. Using the fact that $\Omega(\eta) = 1 - \frac{2}{3}\eta + O(\eta^2)$, we deduce

$$\frac{\partial l}{\partial \eta}|_{\eta=0} = -2 + \alpha(p)^2(-\frac{4}{3}p + 2).$$

By Taylor series in the neighbourhood of zero we have $\alpha(p)^2 = 1 + \frac{2}{3}p + \frac{8}{45}p^2 + O(p^3)$, and for $p \in [-1, 0]$, $\alpha(p)^2 \leq 1 + \frac{2}{3}p + \frac{8}{45}p^2$, thus for all $p \in [-1, 0]$,

$$\frac{\partial l}{\partial \eta}|_{\eta=0} \leq -2 + (1 + \frac{2}{3}p + \frac{8}{45}p^2)(-\frac{4}{3}p + 2) = -\frac{8}{135}p^2(4p+9) \leq 0.$$

Therefore, there exists η_0 small enough such that for all $p \in [-1, 0], \eta \leq \eta_0$, $l(\eta, p) \leq l(0, p) = 1$. This concludes the proof of Theorem 3.3.2. □

3.4 Convergence analysis

We show in this section that the proposed scheme (3.21) has strong order 1/2 and weak order 1 for general systems of SDEs of the form (3.1) with Lipschitz and smooth vector fields, analogously to the simplest Euler-Maruyama method.

We denote by $C_P^4(\mathbb{R}^d, \mathbb{R}^d)$ the set of functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ that are 4 times continuously differentiable with all derivatives with at most polynomial growth. The following theorem shows that the proposed SK-ROCK has strong order 1/2 and weak order 1 for general SDEs.

Theorem 3.4.1. *Consider the system of SDEs* (3.1) *on a time interval of length $T > 0$, with $f, g \in C_P^4(\mathbb{R}^d, \mathbb{R}^d)$, Lipschitz continuous. Then the scheme* (3.21) *has strong order* 1/2 *and weak order* 1*,*

$$\mathbb{E}|(\|X(t_n) - X_n\|) \leq Ch^{1/2}, \qquad t_n = nh \leq T, \tag{3.29}$$

$$|\mathbb{E}(\phi(X(t_n))) - \mathbb{E}(\phi(X_n))| \leq Ch, \qquad t_n = nh \leq T, \tag{3.30}$$

for all $\phi \in C_P^4(\mathbb{R}^d, \mathbb{R})$, where C is independent of n, h.

For the proof the Theorem 3.4.1, the following lemma will be useful. It relies on the linear stability analysis of Lemma 3.3.1.

Lemma 3.4.2. *The scheme* (3.21) *has the following Taylor expansion after one timestep,*

$$X_1 = X_0 + hf(X_0) + \sum_{r=1}^{m} g^r(X_0)\Delta W_r + h\left(\frac{T_s''(\omega_0)\omega_1^2}{T_s(\omega_0)} + \frac{\omega_1}{2}\right) f'(X_0) \sum_{r=1}^{m} g^r(X_0)\Delta W_r + h^2 R_h(X_0),$$

where all the moments of $R_h(X_0)$ are bounded uniformly with respect to h assumed small enough, with a polynomial growth with respect to X_0.

Proof. Using the definition (3.21) of the scheme and the recurrence relations (3.11),(3.23), we obtain by induction on $i = 1, \ldots, s$,

$$\begin{aligned} K_i = X_0 &+ h\frac{T_i'(\omega_0)\omega_1}{T_i(\omega_0)} f(X_0) + \frac{sT_i'(\omega_0)\omega_1}{iT_i(\omega_0)} \sum_{r=1}^{m} g^r(X_0)\Delta W_r \\ &+ h\left(\frac{sT_i''(\omega_0)\omega_1^2}{iT_i(\omega_0)} + \frac{sT_i'(\omega_0)\omega_1^2}{2iT_i(\omega_0)}\right) f'(X_0) \sum_{r=1}^{m} g^r(X_0)\Delta W_r + h^2 R_{i,h}(X_0), \end{aligned} \tag{3.31}$$

and $R_{i,h}(X_0)$ has the properties claimed on $R_h(X_0)$. Using $\omega_1 = T_s(\omega_0)/T_s'(\omega_0)$, this yields the result for $X_1 = K_s$. □

Proof of Theorem 3.4.1. A well-known theorem of Milstein [45] (see [46, Chap. 2.2]) allows to infer the global orders of convergence from the error after one step. We first show that for all $r \in \mathbb{N}$ the moments $\mathbb{E}(|X_n|^{2r})$ are bounded for all n, h with $0 \leq nh \leq T$ uniformly with respect to all h sufficiently small. Then, it is sufficient to show the local error estimate

$$|\mathbb{E}(\phi(X(t_1))) - \mathbb{E}(\phi(X_1))| \leq Ch^2,$$

for all initial value $X(0) = X_0$ and where C has at most polynomial growth with respect to X_0, to deduce the weak convergence estimate (3.30). For the strong convergence (3.30), using the classical result from [44], it is sufficient to show in addition the local error estimate

$$\mathbb{E}(\|X(t_1) - X_1\|) \leq Ch$$

for all initial value $X(0) = X_0$ and where C has at most polynomial growth with respect to X_0. These later two local estimates are an immediate consequence of Lemma 3.4.2.

To conclude the proof of the global error estimates, it remains to check that for all $r \in \mathbb{N}$ the moments $\mathbb{E}(|X_n|^{2r})$ are bounded uniformly with respect to all h small enough for all $0 \leq nh \leq T$. We use here the approach of [46, Lemma 2.2, p. 102] which states that it is sufficient to show

$$|\mathbb{E}(X_{n+1} - X_n|X_n)| \leq C(1 + |X_n|)h, \qquad |X_{n+1} - X_n| \leq M_n(1 + |X_n|)\sqrt{h}, \tag{3.32}$$

where C is independent of h and M_n is a random variable with moments of all orders bounded uniformly with respect to all h small enough. These estimates are a straightforward consequence of the definition (3.21) of the scheme and the linear growth of f, g (a consequence of their Lipschitzness). This concludes the proof of Theorem 3.4.1. □

Remark 3.4.3. *In the case of additive noise, i.e. $g^r, r = 1, \ldots, m$ are constant functions, one can show that the order of strong convergence* (3.29) *become* 1*, analogously to the case of the Euler-Maruyama method. For a general multiplicative noise, a scheme of strong order one can also be constructed with $\mathbb{E}(|R(p, q, \xi)|^2) \leq 1$ for all $p \in [-2\omega_1^{-1}, 0]$ and all q with $p + \frac{|q|^2}{2} \leq 0$, as it can be check numerically. The idea is to modify the first stages of the scheme such that the stability function* (3.24) *becomes*

$$R(p, q, \xi) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} + \frac{U_{s-1}(\omega_0 + \omega_1 p)^2}{U_{s-1}(\omega_0)^2}(1 + \frac{w_1}{2}p - \frac{\omega_1^4}{2}p^2)\left(q\xi + q^2\frac{\xi^2 - 1}{2}\right).$$

We refer to [12, Remark 3.2] for details.

3.5 Long term accuracy for Brownian dynamics

In this section we discuss the long-time accuracy of the SK-ROCK for Brownian dynamics (also called overdamped Langevin dynamics). We will see that using postprocessing techniques we can derive an SK-ROCK method that captures the invariant measure of Brownian dynamics with second order accuracy. In doing so, we do not need our stabilized method to be of weak order 2 on bounded time intervals and we obtain a method that is cheaper than the second weak order S-ROCK2 method proposed in [12], as S-ROCK2 uses many more function evaluations per time-step and a smaller stability domain.

3.5.1 An exact SK-ROCK method for the Orstein-Uhlenbeck process

We consider the 1-dimensional Orstein-Uhlenbeck problem with 1-dimensional noise with constants $\delta, \sigma > 0$,

$$dX(t) = -\delta X(t)dt + \sigma dW(t), \tag{3.33}$$

that is ergodic and has a Gaussian invariant measure with mean zero and variance given by $\lim_{t\to\infty} \mathbb{E}(X(t)^2) = \sigma^2/(2\delta)$. Applying the SK-ROCK method to the above system we obtain

$$X_{n+1} = A(p)X_n + B(p)\sigma\sqrt{h}\xi_n \tag{3.34}$$

where $p = -\delta h$, $\xi_n \sim \mathcal{N}(0,1)$ is a Gaussian variable and similarly as for (3.24) we have

$$A(p) = \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)}, \quad B(p) = \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)}(1 + \frac{\omega_1}{2}p). \tag{3.35}$$

A simple calculation (using that $|A(p)| < 1$) gives

$$\lim_{n\to\infty} \mathbb{E}(X_n^2) = \frac{\sigma^2}{2\delta}R(p), \qquad R(p) = \frac{2pB(p)^2}{A(p)^2 - 1}.$$

From the above equation, we see that the SK-ROCK method has order r for the invariant measure of (3.33) if and only if $R(p) = 1 + \mathcal{O}(p^r)$ and a short calculation using (3.35) reveals that $R(p) = 1 + \mathcal{O}(p)$, it has order one for the invariant measure (this is of course not surprising because the SK-ROCK has weak order one). We next apply the techniques of postprocessed integrators popular in the deterministic literature [20] and proposed in the stochastic context in [56]. The idea is to consider a postprocessed dynamics $\overline{X}_n = G_n(X_n)$ (of negligible cost) such that the process $\overline{X}_n$ approximates the invariant measure of the dynamical system with higher order. For the process (3.33), we consider the postprocessor

$$\overline{X}_n = X_n + c\sigma\sqrt{h}\xi_n, \tag{3.36}$$

which yields $\lim_{n\to\infty} \mathbb{E}(\overline{X}_n^2) = \frac{\sigma^2}{2\delta}(R(p) - 2c^2 p)$. In the case of the SK-ROCK method with $\eta = 0$ (zero damping), we have $A(p) = T_s(1 + p/s^2), B(p) = U_{s-1}(1 + p/s^2)(1 + p/(2s^2))/s$. Setting $c = 1/(2s)$ and using the identity $(1 - x^2)U_{s-1}^2(x) = 1 - T_s^2(x)$ with $x = 1 + p/s^2$ reveals that $R(p) - 2c^2 p = 1$ and we obtain

$$\lim_{n\to\infty} \mathbb{E}(\overline{X}_n^2) = \frac{\sigma^2}{2\delta}. \tag{3.37}$$

Hence the postprocessed SK-ROCK method (that will be denoted PSK-ROCK) captures exactly the invariant measure of the 1-dimensional Orstein-Uhlenbeck problem (3.33). Such a behavior is known for the drift-implicit θ method with $\theta = 1/2$ (see [22] in the context of the stochastic heat equation) and has recently also been shown for the non-Markovian Euler scheme [39]. In [56] an interpretation of the scheme [39] as an Euler-Maruyama method with postprocessing (3.36) with $c = 1/2$ has been proposed and we observe that this is exactly the same postprocessor as for the PSK-ROCK method (with $s = 1, \eta = 0$). As the SK-ROCK method with zero damping is mean-square stable (see Remark 3.3.3 for $\eta = 0$), it can be seen as a stabilized version of the scheme [39]. However, the PSK-ROCK method with $s > 1$ and zero damping is not robust to use as its stability domain along the drift axis does not allow for any imaginary perturbation at the points where $|T_s(1 + p/s^2)| = 1$ and it is not ergodic (see Remark 3.5.2 below).

Stability analysis for Orstein-Uhlenbeck Let $M \in \mathbb{R}^{d\times d}$ denote a symmetric matrix with eigenvalues $-\lambda_d \leq \ldots \leq -\lambda_1 < 0$, and consider the d-dimensional Orstein-Uhlenbeck problem

$$dX(t) = MX(t)dt + \sigma dW(t) \tag{3.38}$$

where $W(t)$ denotes a d-dimensional standard Wiener process. The following theorem shows that the damping parameter $\eta > 0$ plays an essential role to warranty the convergence to the numerical invariant measure $\rho_\infty^h(x)dx$ at an exponentially fast rate.

Theorem 3.5.1. *Let $\eta > 0$. Consider the scheme (3.21) with postprocessor (3.36) applied to (3.38) with stepsize h and stage parameter s such that $2\omega_1^{-1} \geq h\lambda_d$. Then, for all $h \leq \eta/\lambda_1$,$\phi \in C_P^1(\mathbb{R}^d, \mathbb{R})$,*

$$|\mathbb{E}(\phi(\overline{X}_n)) - \int_{\mathbb{R}^d} \phi(x)\rho_\infty^h(x)dx| \leq C\exp(-\lambda_1(1+\eta)^{-1}t_n)$$

where C is independent of $h, n, s, \lambda_1, \ldots, \lambda_d$.

Proof. It is sufficient to show the estimate

$$|A(-\lambda_j h)| \leq \exp(-\lambda_1(1+\eta)^{-1}h) \tag{3.39}$$

for all $h \leq h_0$, where we denote $A(z) = T_s(\omega_0 + \omega_1 z)/T_s(\omega_0)$. Indeed, considering two initial conditions X_0^1, X_0^2 for (3.21) and the corresponding numerical solutions X_n^1, X_n^2 (obtained for the same realizations of $\{\xi_n\}$) with postprocessors $\overline{X}_n^1, \overline{X}_n^2$, we obtain

$$X_n^1 - X_n^2 = A(hM)(X_{n-1}^1 - X_{n-1}^2)$$

and using the matrix 2-norm $\|A(hM)\| = \max_j |A(-\lambda_j h)|$ and (3.39), we deduce by induction on n,

$$\|\overline{X}_n^1 - \overline{X}_n^2\| = \|X_n^1 - X_n^2\| \leq \exp(-\lambda_1(1+\eta)^{-1}t_n)\|X_0^1 - X_0^2\|,$$

and taking $\overline{X}_0^2$ distributed according to the numerical invariant measure yields the result.

For the proof of (3.39), let $z = -\lambda_j h$. Consider first the case $z \in (-\eta\omega_1^{-1}s^{-2}, 0)$. Using the convexity of $A(z)$ on $[-\eta\omega_1^{-1}s^{-2}, 0]$ (note that $T_s'(x)$ is increasing on $[1,\infty)$), we can bound $A(z)$ by the affine function passing by the points $(x_1, A(x_1)), (x_2, A(x_2))$ with $x_1 = -\eta\omega_1^{-1}s^{-2}$, $x_2 = 0$,

$$A(z) \leq 1 + z(1 - 1/T_s(\omega_0))\eta^{-1}\omega_1 s^2$$

Using $\omega_1 s^2 \geq 1$ and $T_s(\omega_0) \geq 1 + \eta$, we obtain

$$(1 - 1/T_s(\omega_0))\eta^{-1}\omega_1 s^2 \geq (1 - (1+\eta)^{-1})\eta^{-1} = (1+\eta)^{-1}.$$

This yields for all $z \in [-\eta\omega_1^{-1}s^{-2}, 0]$,

$$A(z) \leq 1 + z(1+\eta)^{-1} \leq \exp(z(1+\eta)^{-1})$$

where we used the convexity of $\exp(z(1+\eta)^{-1})$ bounded from below by its tangent at $z = 0$. We obtain

$$A(-\lambda_j h) \leq e^{-\lambda_j h(1+\eta)^{-1}} \leq e^{-\lambda_1 h(1+\eta)^{-1}}.$$

We now consider the case $z \in [-L_s, -\eta\omega_1^{-1}s^{-2}]$. We have $|\omega_0 + \omega_1 z| \leq 1$, thus

$$|T_s(\omega_0 + \omega_1 z)| \leq 1$$

and

$$|A(z)| \leq 1/T_s(\omega_0) \leq \exp(-\lambda_1(1+\eta)^{-1}h)$$

for all $h \leq (1+\eta)\log(T_s(\omega_0))/\lambda_1$, and thus also for $h \leq \eta/\lambda_1$ where we use $T_s(\omega_0) \geq 1+\eta$ and $(1+\eta)\log(1+\eta) \geq \eta$. This concludes the proof. □

Remark 3.5.2. *Note that $\eta > 0$ is a crucial assumption in Theorem 3.5.1. Indeed, the estimate of Theorem 3.5.1 is false for $\eta = 0$ already in dimension $d = 1$ for all $s > 1$: for a stepsize h such that $1 - h\lambda_1/s^2 = \cos(\pi/s)$ we obtain $A(-\lambda_1 h) = -1$ and $B(-\lambda_1 h) = 0$ in (3.35) (corresponding to the local extrema $p = -\lambda_1 h$ of $A(p)$ closest to zero) and $X_n = (-1)^n X_0$ for all n, and the scheme is not ergodic. In addition, notice that Theorem 3.5.1 allows to use an h-dependent value of η such as $\eta = h\widetilde{\lambda}_1$ where $\widetilde{\lambda}_1 \geq \lambda_1$ is an upper bound for λ_1. In this case, the exponential convergence of Theorem 3.5.1 holds for all stepsize $h \leq 1$.*

We end this section by noting that being exact for the invariant measure of Brownian dynamics (3.40) is only true for the PSK-ROCK method (or the method in [39]) in the linear case, i.e. for a quadratic potential V. Second order accuracy for the invariant measure has been shown for the method [39]) in [40] for general nonlinear Brownian dynamics (3.40). This will also be shown for the nonlinear PSK-ROCK method in the next section.

3.5.2 PSK-ROCK: a second order postprocessed SK-ROCK method for nonlinear Brownian dynamics

We consider the overdamped Langevin equation,

$$dX(t) = -\nabla V(X(t))dt + \sigma dW(t), \tag{3.40}$$

where the stochastic process $X(t)$ takes values in $\mathbb{R}^d$ and $W(t)$ is a d-dimensional Wiener process. We assume that the potential $V : \mathbb{R}^d \to \mathbb{R}$ has class C^∞ and satisfies the at least quadratic growth assumption

$$x^T \nabla V(x) \geq C_1 x^T x - C_2 \tag{3.41}$$

for two constants $C_1, C_2 > 0$ independent of $x \in \mathbb{R}^d$. The above assumptions warranty that the system (3.40) is ergodic with exponential convergence to a unique invariant measure with Gibbs density $\rho_\infty = Z\exp(-2\sigma^{-2}V(x))$,

$$|\mathbb{E}(\phi(X(t)) - \int_{\mathbb{R}^d} \phi(x)\rho_\infty(x)dx| \leq Ce^{-\lambda t},$$

for test function ϕ and all initial condition X_0, where C, λ are independent of t.

We propose to modify the internal stage $K_1 = X_0 + \mu_1 h f(X_0 + \nu_1 Q) + \kappa_1 Q$ of the method (3.21) as follows:

$$K_1 = X_0 + \mu_1 h f(X_0 + \nu_1 Q) + \kappa_1 Q + \alpha h\big(f(X_0 + \nu_1 Q) - 2f(X_0) + f(X_0 - \nu_1 Q)\big), \tag{3.42}$$

where α is a parameter whose optimal value depends on s and η is discussed below.

Remark 3.5.3. *Notice that for $\alpha = 0$, we recover the original definition from* (3.21)*. We note that the parameter α does not modify the stability function of Lemma 3.3.1, and yields a perturbation of order $\mathcal{O}(h^2)$ in the definition of X_1. Thus, the results of Theorem 3.3.2 and Theorem 3.4.1 remain valid for any value of α for the scheme* (3.21) *with modified internal stage* (3.42)*.*

Theorem 3.5.4. *Consider the Brownian dynamics* (3.40)*, where we assume that $V : \mathbb{R}^d \to \mathbb{R}$ has class C^∞, with ∇V globally Lipschitz and satisfying* (3.41)*. Consider the scheme* (3.21) *applied to* (3.40) *with modified internal stage K_1 defined in* (3.42) *with α defined in* (3.43)*, and the postprocessor defined as*

$$\overline{X}_n = X_n + c\sigma\sqrt{h}\xi,$$

where

$$c^2 = -\frac{1}{4} + \frac{\omega_1}{2} + \frac{\omega_1 T_s''(\omega_0)}{T_s'(\omega_0)} - \frac{\omega_1^2 T_s''(\omega_0)}{4T_s(\omega_0)}, \qquad \alpha = \frac{2}{s\omega_0\omega_1}(c^2 + \frac{\omega_1^2 T_s''(\omega_0)}{2T_s(\omega_0)} - r_s), \tag{3.43}$$

and r_s is defined by induction as $r_0 = 0$, $r_1 = \frac{s^2\omega_1^3}{4\omega_0} := \Delta_1$ and

$$r_i = \nu_i r_{i-1} + \kappa_i r_{i-2} + \Delta_i, \qquad \Delta_i = \mu_i \frac{sT_{i-1}'(\omega_0)\omega_1}{(i-1)T_{i-1}(\omega_0)}, \qquad i = 2, \dots s.$$

Then, $\overline{X}_n$ yields order two for the invariant measure, i.e. (3.5) *holds with $r = 2$, and in addition*

$$|\mathbb{E}(\phi(\overline{X}_n) - \int_{\mathbb{R}^d} \phi(x)\rho_\infty(x)dx| \leq C_1 e^{-\lambda t_n} + C_2 h^2, \tag{3.44}$$

for all $t_n = nh$, $\phi \in C_P^\infty(\mathbb{R}^d, \mathbb{R})$, where C_1, C_2 are independent of h assumed small enough, and C_2 is independent of the initial condition X_0.

The proof of Theorem 3.5.4 relies on the following postprocessing analysis from [56]. Consider a scheme (3.2) with bounded moments and assumed ergodic when applied to (3.40), where the potential V satisfies the above ergodicity assumptions. Assume that the scheme has a weak Taylor expansion after one time step of the form

$$\mathbb{E}(\phi(X_1)|X_0 = x) = \phi(x) + h\mathcal{L}\phi(x) + h^2\mathcal{A}_1\phi(x) + \mathcal{O}(h^3), \tag{3.45}$$

and consider a postprocessor of the form $\overline{X}_n = G_n(X_n)$ where

$$\mathbb{E}(\phi(\overline{X}_1)|X_1 = x) = \phi(x) + h\overline{\mathcal{A}}_1\phi(x) + \mathcal{O}(h^3), \tag{3.46}$$

where the constants in $\mathcal{O}$ in (3.45),(3.46) have at most a polynomial growth with respect to x. Here $\mathcal{L}\phi = \phi' f + \sigma^2/2\Delta\phi$ denotes generator of the SDE and $\mathcal{A}_1, \overline{\mathcal{A}}_1$ are linear differential operators with smooth coefficients. Note that $\mathcal{A}_1 \neq \mathcal{L}^2/2$ in general (otherwise the scheme has weak order 2). If the condition $(A_1 + [\mathcal{L}, \overline{\mathcal{A}}_1])^*\rho_\infty = 0$ holds, equivalently,

$$\left\langle \mathcal{A}_1\phi + [\mathcal{L}, \overline{\mathcal{A}}_1]\phi \right\rangle = 0 \tag{3.47}$$

for all test function ϕ, where we define $\langle\phi\rangle = \int_{\mathbb{R}^d} \phi\rho_\infty dx$, then it is shown in [56, Theorem 4.1] that $\overline{X}_n$ has order two for the invariant measure, i.e. the convergence estimates (3.5) with $r = 2$ and (3.44) hold. Before we can apply the above result, the following lemma allow to compute the weak Taylor expansion of the modified scheme.

Lemma 3.5.5. *Consider the scheme* (3.21) *with modified stage* (3.42) *and assume the hypotheses of Theorem 3.5.4. Then* (3.45) *holds where the linear differential operator* $\mathcal{A}_1$ *is given by*

$$\begin{aligned}\mathcal{A}_1\phi &= \frac{1}{2}\phi''(f,f)+\frac{\sigma^2}{2}\sum_{i=1}^d\phi'''(e_i,e_i,f)+\frac{\sigma^4}{8}\sum_{i,j=1}^d\phi^{(4)}(e_i,e_i,e_j,e_j)+c_2\phi'f'f\\ &+ c_3\frac{\sigma^2}{2}\phi'\sum_{i=1}^d f''(e_i,e_i)+c_4\sigma^2\sum_{i=1}^d\phi''(f'e_i,e_i),\end{aligned}\tag{3.48}$$

where $f=-\nabla V(x)$ *and*

$$c_2=\frac{\omega_1^2T_s''(\omega_0)}{2T_s(\omega_0)},\qquad c_3=r_s+\frac{\omega_0}{s\omega_1}\alpha,\qquad c_4=\frac{T_s''(\omega_0)\omega_1}{T_s'(\omega_0)}+\frac{\omega_1}{2}.\tag{3.49}$$

Proof. Adapting the proof of Lemma 3.4.2, the internal stage K_i defined in (3.21) (and (3.42) for $i=1$) satisfies (3.31) where $h^2R_{i,h}(X_0)$ can be replaced by

$$\frac{\omega_1^2T_i''(\omega_0)}{2T_i(\omega_0)}h^2f'(X_0)f(X_0)+\tilde{r}_i\frac{\sigma^2}{2}f''(X_0)(\xi_n,\xi_n)+h^{5/2}\widetilde{R}_i+h^3\widetilde{R}_{i,h}(X_0),\tag{3.50}$$

where $\mathbb{E}(\widetilde{R}_i)=0$ and all the moments of $\widetilde{R}_i,\widetilde{R}_{i,h}(X_0)$ are bounded with polynomial growth with respect to X_0. Here, $\tilde{r}_i$ is defined by induction as $\tilde{r}_0=0$, $\tilde{r}_1=\Delta_1+\alpha$, and

$$\tilde{r}_i=\nu_i\tilde{r}_{i-1}+\kappa_i\tilde{r}_{i-2}+\Delta_i,\quad i=2,\dots,s.$$

We have $\mathbb{E}(\widetilde{R}_i)=0$ because $\widetilde{R}_i$ is a linear combination of $f'(X_0)f'(X_0)\xi_n$, $f''(X_0)(f(X_0),\xi_n)$, and $f'''(X_0)(\xi_n,\xi_n,\xi_n)$ with zero mean values (recall that odd moments of ξ_n vanish). Next, observing that the difference $d_i=\tilde{r}_i-r_i$ satisfies $d_0=0,d_1=\alpha$, and $d_i=\nu_id_{i-1}+\kappa_id_{i-2}$, $i=2,\dots,s$, we deduce

$$\tilde{r}_i=r_i+d_i,\qquad d_i=\frac{U_{i-1}(\omega_0)}{T_i(\omega_0)}\omega_0\alpha\ \ \forall\ \ i=0,..,s.$$

In particular, taking $i=s$ in (3.31),(3.50), and expanding (3.45), we deduce that (3.48) holds with c_2,c_3,c_4 defined in (3.49) where we note that $c_3=\widetilde{r}_s=r_s+d_s$. □

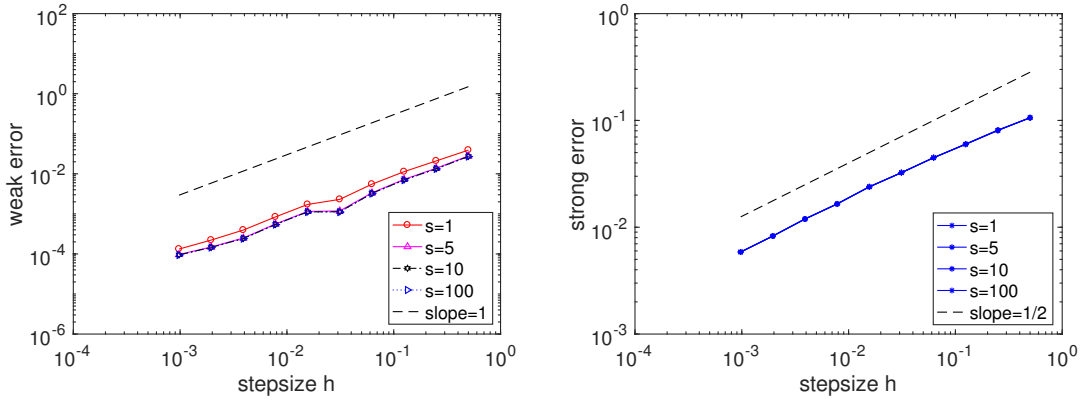
Proof of Theorem 3.5.4. Following the proof of [56, Theorem 4.2] (see also [38, Theorem 5.8]) where we apply repeatedly integration by parts for the integral in (3.47), using Lemma 3.5.5 for the expression of $\mathcal{A}_1$, we deduce that the quantity in (3.47) satisfies

$$\left\langle\mathcal{A}_1\phi+[\mathcal{L},\overline{\mathcal{A}}_1]\phi\right\rangle=\sum_{i=1}^d\left\langle(c_3-c_2-c^2)\frac{\sigma^2}{2}\phi'f''(e_i,e_i)+(c_4-\frac{1}{4}-\frac{c_2}{2}-c^2)\sigma^2\phi''(f'e_i,e_i)\right\rangle,$$

where we use $[\mathcal{L},\overline{\mathcal{A}}_1]=-c^2\sigma^2\big(1/2\,\phi'\sum_{i=1}^d f''(e_i,e_i)+\sum_{i=1}^d\phi''(f'e_i,e_i)\big)$ for $\overline{\mathcal{A}}_1\phi=c^2\sigma^2/2\,\Delta\phi$. We see that the above quantity (3.47) vanishes if $c_3-c_2-c^2=c_4-\frac{1}{4}-\frac{c_2}{2}-c^2=0$, equivalently,

$$c_3-c_2=c^2=c_4-\frac{1}{4}-\frac{c_2}{2}.\tag{3.51}$$

For the values of α,c defined in (3.43), we obtain that (3.51) indeed holds and we deduce that the order two condition (3.47) for the invariant measure is satisfied. This concludes the proof. □



(a) Weak error $|\mathbb{E}(\text{arcsinh}(X(T)) - \mathbb{E}(\text{arcsinh}(X_N))|$.

(b) Strong error $\mathbb{E}(|X(T) - X_N|)$.

Figure 3.4: Nonlinear problem (3.52). Strong and weak convergence plots using SK-ROCK with final time $T = 1$, stepsizes $h = 2^{-p}, p = 1..10$, 10^4 samples and number of stages $s = 1, 5, 10, 100$.

3.6 Numerical experiments

In this Section, we illustrate numerically our theoretical analysis and we show the performance of the proposed SK-ROCK method and its postprocessed modification PSK-ROCK.

3.6.1 A nonlinear nonstiff problem

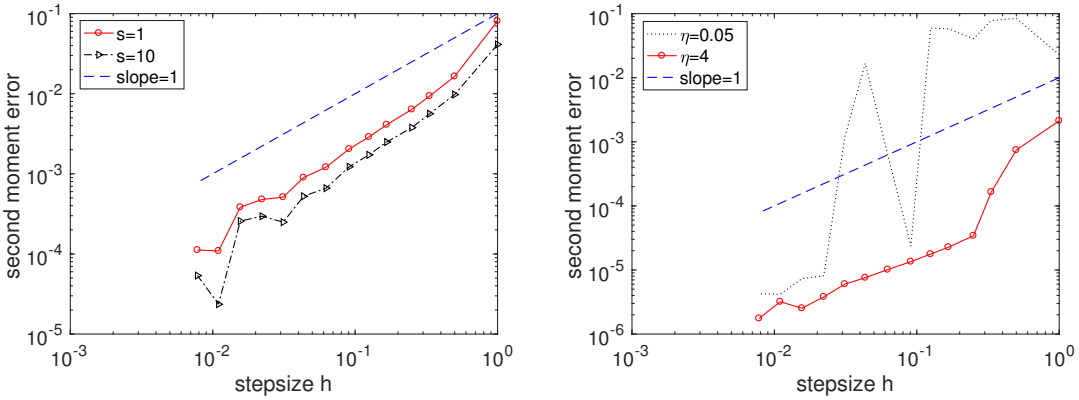
We first consider the following non-stiff nonlinear SDE,

$$dX = \left(\frac{1}{4}X + \frac{1}{2}\sqrt{X^2+1}\right)dt + \sqrt{\frac{1}{2}(X^2+1)}dW, \quad X(0) = 0. \tag{3.52}$$

whose exact solution is $X(t) = \sinh(\frac{t}{2} + \frac{W(t)}{\sqrt{2}})$. In Figure 3.4, we consider the SK-ROCK method (3.21) and plot the strong error $\mathbb{E}(|X(T) - X_N|$ and the weak error $|\mathbb{E}(\text{arcsinh}(X(T)) - \mathbb{E}(\text{arcsinh}(X_N))|$ at the final time $T = Nh = 1$ using 10^4 samples and number of stages $s = 1, 5, 10, 100$. We obtain convergence slopes 1 and 1/2, respectively, which confirms Theorem 3.4.1 stating the strong order 1/2 and weak order 1 of the proposed scheme. Note that $s = 1$ stage is sufficient for the stability of the scheme in the non-stiff case. The results for $s = 5, 10, 100$ yield nearly identical curves which illustrates that the error constants of the method are nearly independent of the stage number of the scheme.

3.6.2 Nonlinear nonglobally Lipschitz stiff problems

Consider the following nonlinear SDE in dimensions $d = 2$ with a one-dimensional noise ($d = 2, m = 1$). This is a modification of a one-dimensional population dynamics model



(a) Non-stiff case $-\lambda_1 = \mu_1 = 1$. (b) Stiff case $-\lambda_1 = \mu_1^2 = 100$.

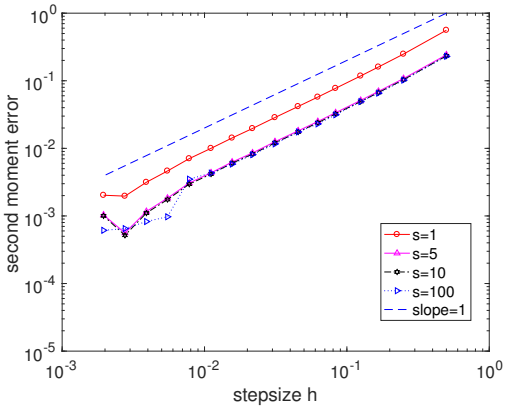
Figure 3.5: Nonlinear problem (3.53) with $\nu = 2, \mu_2 = 0.5, \lambda_2 = -1$. Weak convergence plots using SK-ROCK for $\mathbb{E}(X(T)^2)$ where $T = 1$, $h = T/[2^{i/2}], i = 1, \ldots, 14$, and 10^6 samples. For the stiff case (b), the method uses the following number of stages respectively: $s = 8, 6, 5, 4, 4, 3, 3, 3, 2, 2, 2, 1, 1, 1$ (with damping $\eta = 0.05$) and $s = 13, 9, 8, 7, 6, 5, 4, 4, 3, 3, 3, 3, 2, 2$ (with damping $\eta = 4$).

from [25, Chap. 6.2] considered in [11, 12, 6] for testing stiff integrator performances,

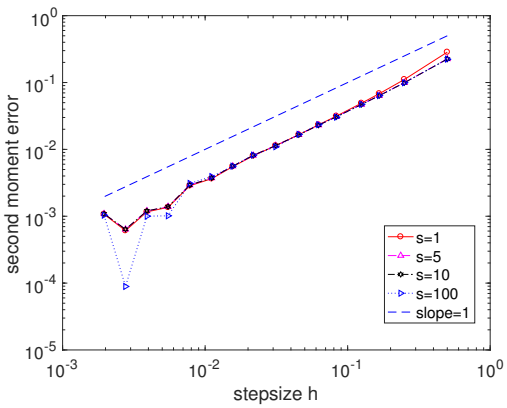
$$\begin{aligned} dX &= (\nu(Y-1) - \lambda_1 X(1-X))dt - \mu_1 X(1-X)dW, \quad X(0) = 0.95, \\ dY &= -\lambda_2 Y(1-Y)dt - \mu_2 Y(1-Y)dW, \quad Y(0) = 0.95. \end{aligned} \tag{3.53}$$

Observe that linearizing (3.53) close to the equilibrium $(X, Y) = (1, 1)$, we recover for $\nu = 0$ the scalar test problem (3.16). In Figure 3.5 we consider the SK-ROCK method applied to (3.53) with parameters that are identical to those used in [6, Sect. 4.2]. We take the initial condition $X(0) = Y(0) = 0.95$ close to this steady state and use the parameters $\nu = 2, \mu_2 = 0.5, \lambda_2 = -1$. In a nonstiff regime ($-\lambda_1 = \mu_1 = 1$ in Figure 3.5(a)), we observe a convergence slope 1 for the second moment $\mathbb{E}(X(T)^2)$ which illustrates the weak order one of the scheme, although our analysis in Theorem 3.4.1 applies only for globally Lipschitz vector fields. The stage number $s = 1$ is sufficient for stability, but we also include for comparison the results for $s = 10$ (note that the results for $s = 50, 100$ not displayed here are nearly identical to the case $s = 10$). The convergence curves are obtained as an average over 10^6 samples. In a stiff regime ($-\lambda_1 = \mu_1^2 = 100$ in Figure 3.5(b)), we observe for the standard small damping $\eta = 0.05$ a stable but not very accurate convergence, due to the severe nonlinear stiffness. However, considering a slightly larger damping $\eta = 4$, in the spirit of the S-ROCK method, yields a stable integration for all considered timesteps and all trajectories and we observe a line with slope one for the SK-ROCK method. Here, given the timesteps h, the numbers of stages s are adjusted as proposed in (3.54) where $\lambda_{\max} = |\lambda_1| = 100$.

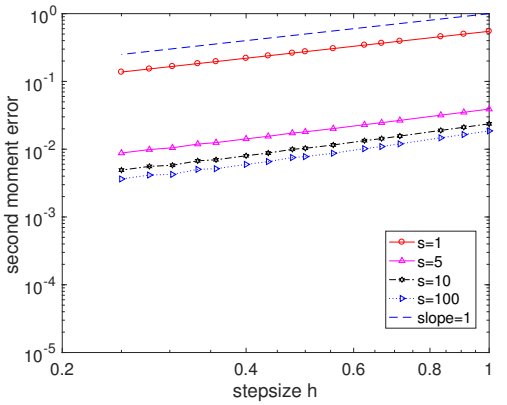
Remark 3.6.1. *For severely stiff problems, alternatively to switching to drift-implicit schemes [33, 11], one can consider in SK-ROCK a slightly larger damping η and the corresponding stage parameter s below, similar to* (3.14) *and chosen such that the mean-square*



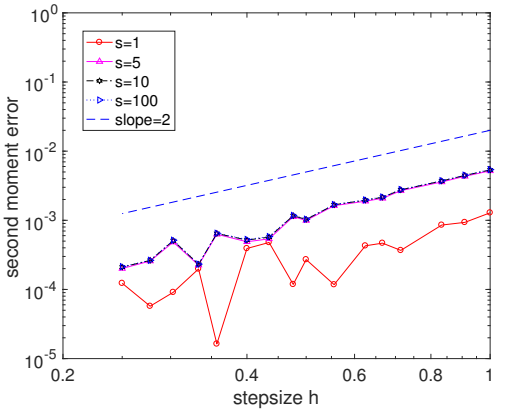
(a) SK-ROCK with $T = 0.5$.



(b) PSK-ROCK with $T = 0.5$.



(c) SK-ROCK with $T = 10$.



(d) PSK-ROCK with $T = 10$.

Figure 3.6: Linear additive problem (3.55). Second moment error $\mathbb{E}(X(T)^2)$ for for short time $T = 0.5$ (top pictures) and long time $T = 10$ (bottom pictures) without (SK-ROCK) or with a postprocessor (PSK-ROCK). where, $h = T/[10\times 2^{i/8}]$ for $T = 10$, and $h = T/[2^{i/2}]$ for $T = 0.5$ with $i = 1, \ldots, 16$, and 10^8 samples.

stability domain length (3.20) *satisfies* $L > h\lambda_{\max}$,

$$s = \left[\sqrt{\frac{h\lambda_{\max} + 1.5}{2\Omega(\eta)}} + 0.5\right], \tag{3.54}$$

where $\Omega(\eta)$ *is given in Lemma 3.3.4.*

3.6.3 Linear case: Orstein-Uhlenbeck process

We now illustrate numerically in details the role of the postprocessor introduced in Theorem 3.5.4 for the linear Orstein-Uhlenbeck process in dimension $d = m = 1$,

$$dX = -\lambda X dt + \sigma dW, \quad X(0) = 2 \tag{3.55}$$

where we choose $\lambda = 1$ and $\sigma = \sqrt{2}$.

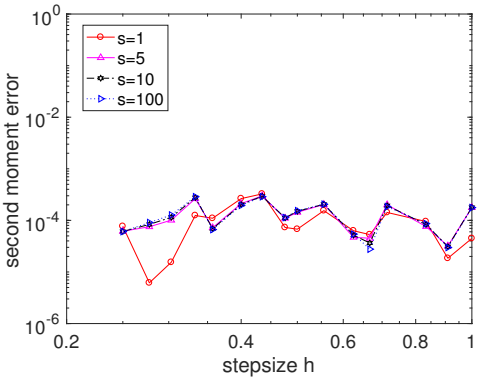


Figure 3.7: PSK-ROCK without damping ($\eta = 0$). Second moment error of problem (3.55), with $T = 10$, $h = T/[2^{i/2}]$, and $s = 1, 5, 10, 100$, using $M = 10^8$ samples (the Monte-Carlo error has size $M^{-1/2} = 10^{-4}$).

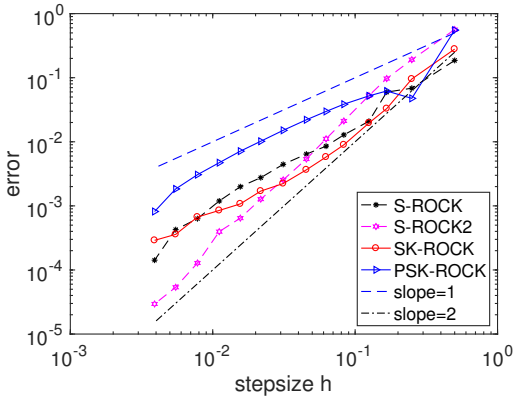
In Figure 3.6, we consider the SK-ROCK and PSK-ROCK methods with $s = 1, 5, 10, 100$ stages, respectively. For a short time $T = 0.5$ (Fig. 3.6(a)(b)), we observe weak convergence slopes one for both SK-ROCK and PSK-ROCK (second moment $\mathbb{E}(X(T)^2)$) as predicted by Theorem 3.4.1, and the postprocessor has nearly no effect of the errors. For a long time $T = 10$ where the solution of this ergodic SDE is close to equilibrium, we observe that the weak order one of SK-ROCK (Fig. 3.6(c)) is improved to order two using the postprocessor in PSK-ROCK (Fig. 3.6(d)), which confirms the statement of Theorem 3.5.4 that the postprocessed scheme has order two of accuracy for the invariant measure. For comparison, in Figure 3.7, we also include the results of PSK-ROCK without damping ($\eta = 0$) using $M = 10^8$ samples. We recall that for the scalar linear Orstein-Uhlenbeck process, the PSK-ROCK method with zero damping is exact for the invariant measure (see Section 3.5.1). We observe only Monte-Carlo errors with size $\simeq M^{-1/2} = 10^{-4}$, which confirms that the PSK-ROCK method has no bias at equilibrium for the invariant measure in the absence of damping, as shown in (3.37). We emphasise however that this exactness results holds only for linear problems, and a positive damping parameter η should be used for nonlinear SDEs for stabilization, as shown in Sections 3.3 and 3.5.1.

3.6.4 Nonglobally Lipschitz Brownian dynamics

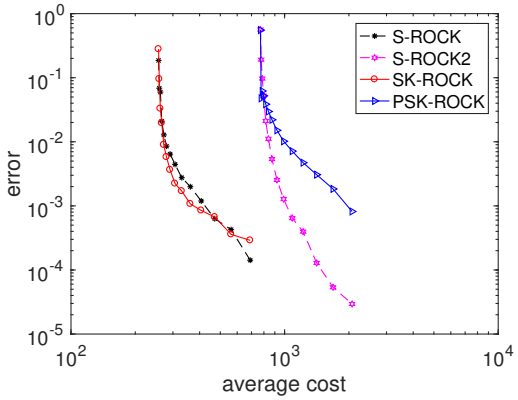
To illustrate the advantage of the PSK-ROCK method applied to nonglobally Lipschitz ergodic Brownian dynamics, we next consider the following double well potential $V(x) = (1 - x^2)^2/4$ and the corresponding one-dimensional Brownian dynamics problem

$$dX = (-X^3 + X)dt + \sqrt{2}dW, \qquad X(0) = 0, \tag{3.56}$$

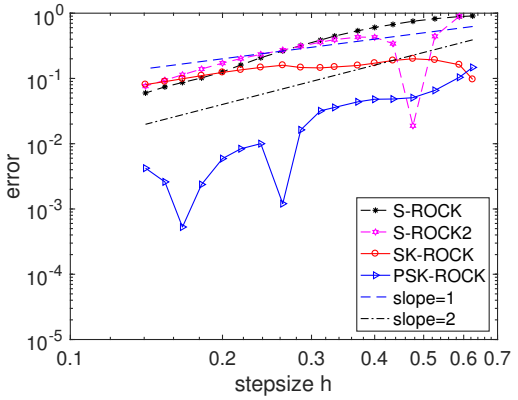
In Figure 3.8, we compare the performances of S-ROCK, S-ROCK2 considered in [12] (a method with weak order 2 for general SDEs), and the new SK-ROCK and PSK-ROCK methods at short time $T = 0.5$ (Figures 3.8(a)(b)) and long time $T = 10$ (Figures 3.8(c)(d)). As we focus on invariant measure convergence and not on strong convergence, we consider here discrete random increments with $\mathbb{P}(\xi_n = \pm\sqrt{3}) = 1/6, \mathbb{P}(\xi_n = 0) = 2/3$, which has the correct moments so that Theorem 3.5.4 remains valid. Our numerical tests indicate that it makes PSK-ROCK with modified stage (3.42) more stable. For a fair comparison, we use



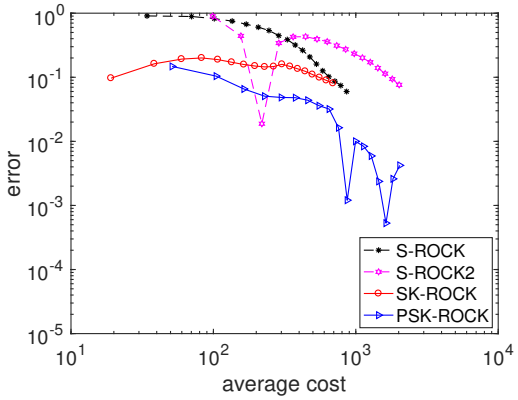
(a) Final time $T = 0.5$, $h = T/[2^{i/2}]$, $i = 1, \ldots, 14$.



(b) Final time $T = 0.5$, $h = T/[2^{i/2}]$, $i = 1, \ldots, 14$.



(c) Final time $T = 10$, $h = T/[15 \times 2^{i/8}]$, $i = 1, \ldots, 18$.



(d) Final time $T = 10$, $h = T/[15 \times 2^{i/8}]$, $i = 1, \ldots, 18$.

Figure 3.8: Second moment errors versus the average number of drift function evaluations for problem (3.56) using S-ROCK, S-ROCK2 and the new method SK-ROCK and its postprocessed version PSK-ROCK. We use discrete random increments and 10^8 samples.

the same discrete random increments for all schemes. We plot the second moment error versus the time stepsize h and versus the average cost which is the total number of function evaluations during the time integration divided by the total number number of samples. Indeed, the number of function evaluations depends on the trajectories because the stage parameter s is adaptive at each time step. For short time, we can see that the S-ROCK and the SK-ROCK method have order 1 (Figure 3.8(a)) and exhibit similar performance with nearly identical error versus cost curves in Figure 3.8(b), while PSK-ROCK is less advantageous for short time. This illustrates that the postprocessing has no advantage for short times. The S-ROCK2 method is the most accurate for small time steps, and it has order 2 as shown in Figures 3.8(a)(c), but at the same time it has a larger average cost as observed in Figures 3.8(b)(d) due to its smaller stability domain with size $\simeq 0.42 \cdot s^2$. For long time, the SK-ROCK and S-ROCK both exhibit order 1 of accuracy (Figure 3.8(c)), with an advantage in terms of error versus cost for the SK-ROCK method that is about 10 times more accurate for large time steps. In contrast, the postprocessed scheme PSK-ROCK exhibits order 2 of convergence (Figure 3.8(c)) which corroborates Theorem 3.5.4.

Since the postprocessing overcost is negligible (two additional vector field evaluations per timestep due to the modified stage K_1 in (3.42)), this makes PSK-ROCK the most efficient in terms of error versus cost, as shown in Figure 3.8(d). The S-ROCK2 method has order 2 here but with poor accuracy compared to the PSK-ROCK method with approximately the same cost. Note that typically the SK-ROCK method used $s = 1, 2, 3$ stages in contrast to the S-ROCK method using $s = 2, \ldots, 6$ stages per timesteps.

3.6.5 Stochastic heat equation with multiplicative space-time noise

Although our analysis applies only to finite dimensional systems of SDEs, we consider the following stochastic partial differential equation (SPDE) obtained by adding multiplicative noise to the heat equation,

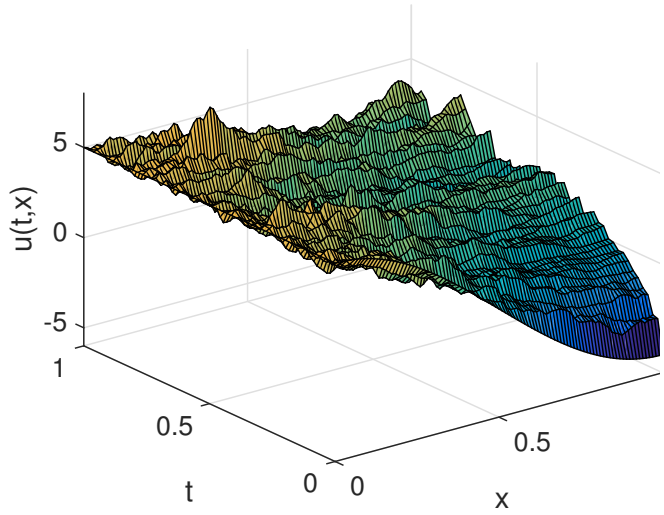
$$\begin{aligned} \frac{\partial u(t,x)}{\partial t} &= \frac{\partial^2 u(t,x)}{\partial x^2} + u(t,x)\dot{W}(t,x), \quad (t,x) \in [0,T] \times [0,1] \\ u(0,x) &= 5\cos(\pi x), \quad x \in [0,1], \\ u(t,0) &= 5, \quad \frac{\partial u(t,1)}{\partial x} = 0, \quad t \in [0,T], \end{aligned} \tag{3.57}$$

where $\dot{W}(t,x)$ denotes a space-time white noise that we discretize together with the Laplace operator with a standard finite difference formula [23]. We obtain the following stiff system of SDEs where $u(x_i,t) \approx u_i(t)$, with $x_i = i\Delta x$, $\Delta x = 1/N$,

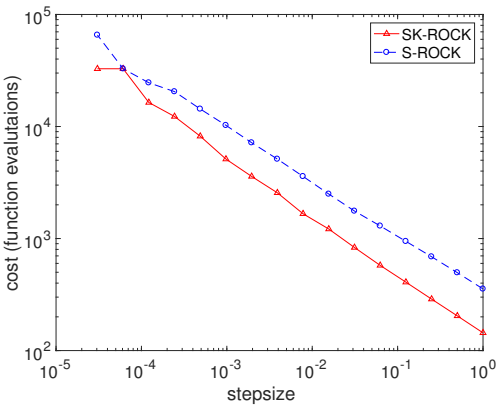
$$du_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} dt + \frac{u_i}{\sqrt{\Delta x}} dw_i, \quad i = 1, \ldots, N,$$

where the Dirichlet and the Neumann conditions impose $u_0 = 5$ and $u_{N+1} = u_{N-1}$, respectively. Here, $w_1, \ldots, w_N$ are independent standard Wiener processes and dw_i indicates Itô noise. In Figure 3.9(a), we plot one realization of the SPDE using space stepsize $\Delta x = 1/100$ and timestep size $\Delta t = 1/50$. Note that the Lipchitz constant associated to the space-discretization of (3.57) has size $\rho = 4\Delta x^{-2}$, and the stability condition is fulfilled for $s = 22$ stages. For comparison, the standard S-ROCK method would require $s = 46$ stages, while applying the standard Euler-Maruyama with a smaller stable timestep $\Delta t/s$ would require $s \geq \Delta t\rho/2 = 400$ intermediate steps. Notice that the initial condition in (3.57) satisfies the boundary conditions, which permits a smooth solution close to time $t = 0$. Taking alternatively an initial condition that does not satisfy the boundary conditions (for instance $u(x,0) = 1$) yields an inaccurate numerical solution with large oscillations close to the boundary $x = 0$. A simple remedy in such a case is to consider a larger damping parameter η, as described in Remark 3.6.1.

In Figure 3.9(b), we compare the number of vector field evaluations of the standard S-ROCK and new SK-ROCK methods when applied to the SPDE (3.57) with finite difference discretization with parameter $\Delta x = 1/100$. The better performance of SK-ROCK with damping $\eta = 0.05$ is due to its larger stability domain with size $\simeq 1.94 \cdot s^2$ compared to the size $\simeq 0.33 \cdot s^2$ for S-ROCK. Observing the ratio of the two costs in Figure 3.9(b), we see that the new SK-ROCK methods has a reduced cost for stabilization by an asymptotic factor of about $\sqrt{1.94/0.33} \simeq 2.4$ for large s and large stepsizes, which confirms the

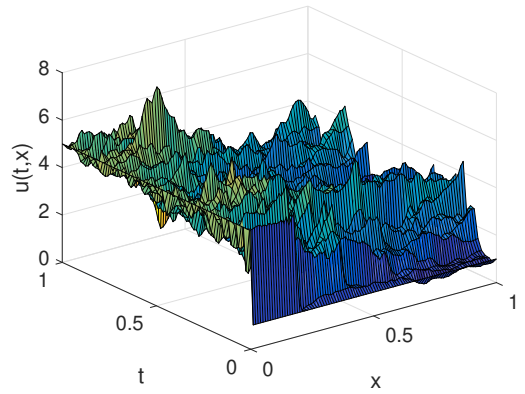


(a) One realization with SK-ROCK using $\Delta t = 1/50$, $\Delta x = 1/100$, $s = 22$.

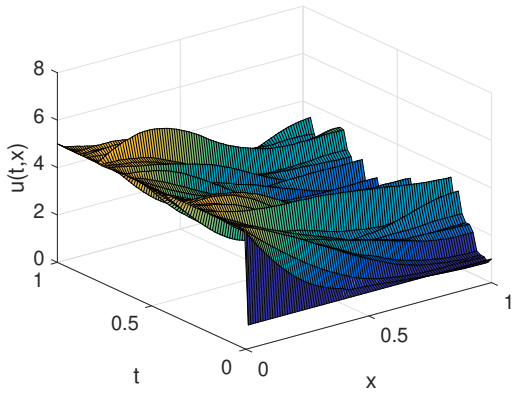


(b) Comparison of S-ROCK and SK-ROCK: cost (function evaluations) with respect to stepsize $\Delta t = 2^{-i}, i = 0, \ldots, 15$.

Figure 3.9: SPDE problem (3.57) using the space discretization stepsize $\Delta x = 1/100$.



(a) $\eta = 0.05$, $s = 22$.



(b) $\eta = 10$, $s = 44$.

Figure 3.10: SPDE problem (3.57) with the initial condition $u(0, x) = 1$. One realization with SK-ROCK using $\Delta t = 1/50$, $\Delta x = 1/100$ for different values of the damping parameter η.

stability analysis of Section 3.3. The convergence analysis of the SK-ROCK method for the stochastic heat equation is the topic of future work.

Remark 3.6.2. *Notice that SK-ROCK with $s = 1$ stage has the optimal mean-square stability length ($L = 2$ for $\eta = 0$) as defined in (3.20). In contrast, the S-ROCK method with $s = 1$ has the smaller stability length $L = 3/2$, while the standard Euler-Maruyama has $L = 0$. This explains why for the smallest considered stepsize $\Delta t = 2^{-15}$ in Figure 3.9(b), we have $s = 1$ for SK-ROCK while S-ROCK uses $s = 2$ stages.*

In Figure 3.10 we consider again one realization with SK-ROCK of the SPDE problem (3.57) but with a different initial condition $u(0,x) = 1$ not fulfilling the boundary conditions, i.e. that is outside the domain of the Laplace operator, as considered in [6]. We compare the result for the same sets of random numbers but for different values of the damping parameter η. We observe numerically high oscillations in time and space for the small damping value $\eta = 0.05$ in Figure (3.10a) while the larger damping $\eta = 10$ yields a smoother solution in Figure (3.10b). This illustrates again Remark 3.6.1 showing that the damping parameter η can be increased in the case of severely stiff problems, adjusting the stage parameter accordingly with (3.54).

3.7 Explicit stabilized method for advection-diffusion equations with optimal stability domain

We have seen in Chapter 2 that explicit stabilized methods have extended stability domains along the negative real axis, this property helps in reducing the cost of integration of ODEs for which the eigenvalues of the Jacobian matrix of the vector field are located very closely to the negative real axis, and the spectral radius is large in modulus (see Figure 2.3). These problems usually arise from the space discretization of second order parabolic (diffusion) PDEs, or diffusion dominated advection-diffusion equations (small Peclet number regime). In the case of advection-diffusion equations with large Peclet number (of size $\mathcal{O}(1)$ or more), these methods fail due to the stability domain limitation in the imaginary direction.

3.7.1 Stability of advection-diffusion problems

In this section we propose an explicit stabilized method for advection-diffusion problems, with large Peclet number. The new method is of order one of accuracy, but it has stability domain of optimal length in the real direction ($\approx 2s^2$) and increasing length as $\sqrt{2\Re(z)}$ in the imaginary direction, which makes it ideal for problems where the imaginary part of the eigenvalues is of size $\mathcal{O}(\sqrt{\text{real part}})$, such as advection-diffusion equations.

A partitioned Ruge-Kutta-Chebyshev method (PRKC) of order 2 was designed in [58] based on the RKC method (2.22) for the integration of ODEs that have moderately stiff (diffusion) and non-stiff terms (advection or costly reaction terms). PRKC has a limited stability for the advection term, and it shares with the standard RKC the same limited stability domain length over the negative real axis. In [10], the authors propose a partitioned implicit-explicit orthogonal Runge-Kutta method (called PIROCK) for the time integration of advection-diffusion-reaction problems with possibly severely stiff reaction terms and stiff stochastic terms. The diffusion terms are solved by the explicit second order orthogonal Chebyshev method (ROCK2). Applied to advection-diffusion problems, the method has order 2 of accuracy and can handle the large Peclet number regime but the length of its stability domain along the negative real axis is limited at most to $0.81s^2$. In addition, PIROCK relies on the ROCK2 method, for which no explicit formulas are available to compute the coefficients for stage-number s (see Section 2.1.3.3). Despite its order one of accuracy, the scheme presented in this section has two main advantages over PIROCK:

- A much longer stability domain over the negative real axis.
- Simple explicit formulas are available for the coefficients.

We consider problems of the form

$$\dot{y}(t) = F_D(y) + F_A(y), \qquad y(0) = y_0, \tag{3.58}$$

where F_D represents the diffusion term with eigenvalues close to the negative real axis, and F_A represents the advection term with eigenvalues close to the imaginary axis and symmetric with respect to the origin. The eigenvalues of the system lie in an ellipse in the left half plane, tangent to the imaginary axis, with center close to the negative real axis. Usually, such ODE arise when discretizing, in space, advection-diffusion equations of the form

$$\partial_t u(x,t) = d\Delta u(x,t) - a\partial_x u(x,t) \quad (+\text{initial and boundary conditions}), \tag{3.59}$$

where d and a are two positive parameters. The eigenvalues of the discrete Laplacian grow as $1/\Delta x^2$ while those of the advection operator ∂_x grow as $1/\Delta x$, which means that the ellipse containing the eigenvalues of the Jacobian of the obtained system, has the length of the minor axis proportional to the square root of the length of the major axis.

3.7.2 An optimal method of order 1 for advection-diffusion equations

Consider the linear test problem

$$\dot{y} = \lambda y + i\mu y, \qquad y(0) = y_0, \tag{3.60}$$

where $\lambda \in \mathbb{R}^-$, $\mu \in \mathbb{R}$, and $i = \sqrt{-1}$. Applying a Runge-Kutta method to the above equation, on gets an induction of the form

$$y_{n+1} = R(p,q)y_n, \tag{3.61}$$

with $p = h\lambda$ and $q = h\mu$. We define the stability domain of a Runge-Kutta method applied to (3.60) by

$$\mathcal{S} = \{(p,q) \in \mathbb{R}^2 \ ; \ |R(p,q)| \leq 1\}. \tag{3.62}$$

Equation (3.60) can be seen as the test equation for linear SDEs with the $i\mu$ replacing the noise. Hence, inspired by SK-ROCK, we consider the following stability polynomial

$$R(p,q) = A(p) + B(p)iq := \frac{T_s(\omega_0 + \omega_1 p)}{T_s(\omega_0)} + \frac{U_{s-1}(\omega_0 + \omega_1 p)}{U_{s-1}(\omega_0)}(1 + \frac{\omega_1}{2}p)iq, \tag{3.63}$$

where T_s and U_s are the first and the second kind Chebyshev polynomials of degree s (the number of stages), and the coefficients ω_0 and ω_1 are defined in (3.9). The stability condition $|R(p,q)| \leq 1$ is equivalent to $A(p)^2 + B(p)^2 q^2 \leq 1$ which is exactly the stability condition (3.22). By Theorem 3.3.2 and Remark 3.3.6, for all $\eta > 0$ and all $s \in \mathbb{N}$, $|R(p,q)| \leq 1$ for all $p \in [-2\omega_1^{-1}, 0]$ and $|q| \leq \sqrt{-2p}$ (See Figure 3.11).

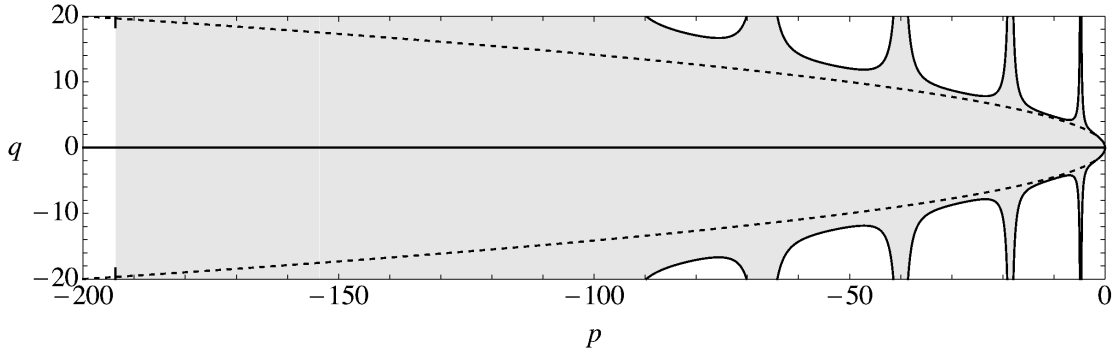


Figure 3.11: Stability domain of the new DA-ROCK method (3.64) in the $p-q$ plane for $s = 10$ and $\eta = 0.05$. The dashed lines correspond to $\pm\sqrt{-2p}$.

The new AD-ROCK method for advection-diffusion equations is defined in the same way as the SK-ROCK method (3.21), by replacing the noise term by the advection term,

$$\begin{aligned} K_0 &= y_0 \\ K_1 &= y_0 + \mu_1 h F_D(y_0 + \nu_1 h F_A(y_0)) + \kappa_1 h F_A(y_0) \\ K_i &= \mu_i h F_D(K_{i-1}) + \nu_i K_{i-1} + \kappa_i K_{i-2}, \quad i = 2, \dots, s, \\ y_1 &= K_s. \end{aligned} \tag{3.64}$$

All the coefficients are the same as those of the SK-ROCK method (3.21). Note that the method requires only 1 evaluation of the advection term per time step.

Assuming enough regularity on F_D and F_A, the convergence proof is straightforward and based on Lemma (3.4.2).

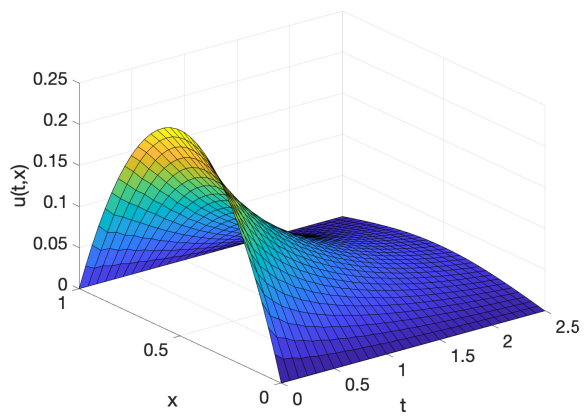
3.7.3 Numerical experiments

We consider the following nonlinear Burgers equation in 1D,

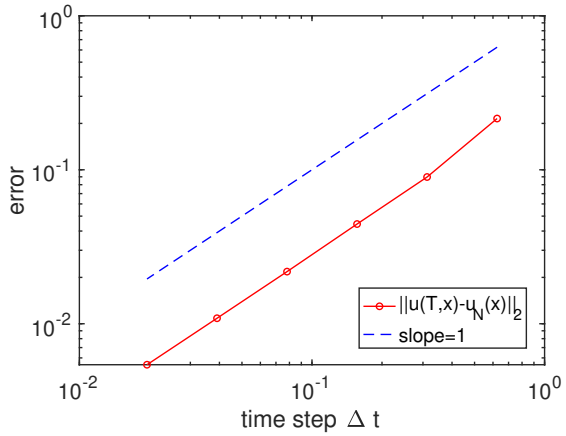
$$\begin{aligned} &\partial_t u = \mu \Delta u - \frac{\nu}{2} \partial_x (u^2) \qquad \text{in } (0,T] \times (0,1), \\ &u(0,x) = x(1-x) \quad \text{in } (0,1), \\ &u(t,0) = u(t,1) = 0, \end{aligned} \tag{3.65}$$

where $\mu, \nu > 0$, in dimension $d = 1$, and the final time is given by $T = 2.5$.

To perform the numerical experiment we choose $\mu = 0.1$ and $\nu = 2$, hence, a very large Peclet number equal to 20. We discretize (3.65) in space using finite differences with step size $\Delta x = 1/M$, where $M \in \mathbb{N}^*$, and we solve the obtained M-dimensional ODE using the AD-ROCK method (3.64) with time step T/N with $N \in \mathbb{N}^*$. In Figure 3.12a, we plot the solution obtained for M=N=30, which required $s = 5$ stages. We illustrate the convergence of the method in Figure 3.12b in which we plot the error at the final time in log scale. The slope 1 is clearly seen. For this convergence lot, we set $M = 500$ and $N \in \{2^i\}_{i=2,\dots 7}$. The corresponding numbers of internal stages are $s = 181, 128, 91, 64, 46, 32$ respectively.



(a) Solution.



(b) Convergence plot.

Figure 3.12: Solution and convergence plot of the DA-ROCK method applied to problem (3.65).

3.7.4 Conclusion

In this section, we profited from the SK-ROCK method to design an AD-ROCK method for advection-diffusion problems. The new method is a first order integrator with optimally large stability domain over the negative real axis that reduces efficiently the time step restriction induced by the stiffness of the diffusion term. In addition, the scheme enjoys a sufficient width in the imaginary direction enough to capture the imaginary parts of the eigenvalues caused by the advection term, even in the large Peclet number regime. Apart from the first order, the method has many advantages over other schemes from the literature [10, 58].

This approach could be extended to design higher order explicit stabilized methods for advection-diffusion PDEs, for example, by constructing partitioned schemes using second kind Chebyshev polynomials together with RKC or even nearly optimal ROCK polynomials.

Chapter 4

Explicit stabilized integrators for stiff optimal control problems

Note: This chapter is identical to the paper [14] in collaboration with Gilles Vilmart.

4.1 Introduction

In this chapter, we introduce and analyze numerical methods for the optimal control of systems of ordinary differential equations (ODEs) of the form

$$\min_u \ \Psi(y(T)); \qquad \dot{y}(t) := \frac{dy}{dt}(t) = f(u(t), y(t)), \quad t \in [0,T]; \quad y(0) = y^0, \tag{4.1}$$

where for a fixed final time $T > 0$ and a given initial condition $y^0 \in \mathbb{R}^n$, the function $y : [0,T] \to \mathbb{R}^n$ is the unknown state function, $u : [0,T] \to \mathbb{R}^m$ is the unknown control function. Here, $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ is the given vector field and $\Psi : \mathbb{R}^n \to \mathbb{R}$ is the given cost function, which are assumed to be C^∞ mappings. For simplicity of the presentation, we consider the case of autonomous problems (with f independent of time) but we highlight that our approach also applies straightforwardly to non-autonomous problems $\frac{dy}{dt}(t) = f(t, u(t), y(t))$.[2]

There are essentially two approaches for the numerical solution of optimal control problems: the direct approach, which consists in directly discretizing (4.1) and then applying a minimization method to the corresponding discrete minimization problem, and the indirect approach, which is based on Pontryagin's maximum principle, taking benefit of continuous optimality conditions (adjoint equation). A natural approach for the accurate numerical approximation of such optimal control problem (4.1) is to consider Runge-Kutta type schemes. It was shown in [27, Theorem 4.1] by studying the continuous and discrete optimality conditions that additional order conditions for the convergence rate are required in general by Runge-Kutta methods when applied to optimal control problems, compared to the integration of standard initial valued ordinary differential equations and conditions up to order 4 were derived. In [16], general order conditions were derived, in addition

[2] A standard approach is to consider the augmented system with $z(t) = t$, i.e. $\frac{dz}{dt} = 1,\ z(0) = 0$ and define $\tilde{y}(t) = (y(t), z(t))^T$, see e.g. [28, Chap. III] for details.

to identifying symplecticity properties. This result is related to the order of symplectic partitioned Runge-Kutta methods, and it implies in particular that applying naively a Runge-Kutta method to (4.1) yields in general an order reduction phenomenon. This analysis was then extended to other classes of Runge-Kutta type schemes in [35, 37, 30], see also [31, 13] in the context of hyperbolic problems and multistep methods. The use of symplectic integrators is motivated by the recent publication [42] which proves the convergence of forward-backward iterative algorithm (Algorithm 4.2.3 in the present chapter), to implement discretized optimal control problems, when using a symplectic Runge-Kutta method. The work done in [42] generalizes that of [41] in which the authors prove the global convergence of the algorithm in the continuous time case. We also mention the paper [57] where automatic differentiation can be efficiently applied for computing the gradient of the cost function under the assumption that optimal control order conditions are satisfied. In our algorithms, the Jacobian of the vector field is given as an input, however the idea of automatic differentiation could be coupled with our approach to compute derivatives automatically, but this is not the purpose of the present work.

In the case where the vector field f in (4.1) is stiff, due for instance to the multiscale nature of the model, or due to the spatial discretization of a diffusion operator in a partial differential equation (PDE) model, standard explicit integrators face in general a severe time step restriction making standard explicit methods unreasonable to be used due to their dramatic cost. A standard approach in this stiff case is to consider indirect implicit methods with good stability properties, as studied in [30] in the context of implicit-explicit (IMEX) Runge-Kutta methods for stiff optimal control problems. Note however that, already for initial value ODEs, such implicit methods can become very costly for nonlinear stiff problems in large dimension, requiring the usage of Newton-type methods and sophisticated linear algebra tools (preconditioners, etc.). Alternatively to using implicit methods, in this thesis we focus on fully explicit indirect methods, and introduce new families of explicit stabilized methods for stiff optimal control problems. The proposed methods rely on the so-called Runge-Kutta-Chebyshev methods of order one and its extension RKC of order two [54]. Such explicit stabilized methods are popular in the context of initial value problems of stiff differential equations, particularly in high dimensions in the context of diffusive PDEs, see e.g. the survey [4]. It was extended to the stochastic context first in [7, 8] and recently in [5] for the design of explicit stabilized integrators with optimally large stability domains in the context of mean-square stable stiff and ergodic problems.

This chapter is organized as follows. In Section 4.2, we recall standard tools on explicit stabilized methods and classical results on standard Runge-Kutta methods applied to optimal control problems. In Section 4.3, we introduce the new explicit stabilized schemes for optimal control problems and analyze their convergence and stability properties. Finally, Section 4.4 is dedicated to the numerical experiments where we illustrate the efficiency of the new approach.

4.2 Preliminaries

4.2.1 Discretization, order conditions, and symplecticity

Let us first recall the definition of a Runge-Kutta method for ordinary differential equations (ODEs),

$$\dot{y}(t) = F(y(t)), \qquad y(0) = y^0, \tag{4.2}$$

where $y : [0,T] \to \mathbb{R}^n$ is the unknown solution, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth vector field, and $y^0 \in \mathbb{R}^n$ is a given initial condition. We consider for simplicity a uniform discretization of the interval $[0,T]$ with $N+1$ points for $N \in \mathbb{N}$, and denote by $h = T/N$ the stepsize. For a given integer s and given real coefficients $b_i,\ a_{ij}\ (i,j = 1,\dots,s)$, an s-stage Runge-Kutta method, $y_k \approx y(t_k),\ t_k = kh$, to approximate the solution of (4.2), is defined, for all $k = 0,\dots,N-1$, by

$$y_{ki} = y_k + h\sum_{j=1}^{s} a_{ij}F(y_{kj}), \quad i = 1,\dots,s, \qquad y_{k+1} = y_k + h\sum_{i=1}^{s} b_i F(y_{ki}). \tag{4.3}$$

The coefficients are usually displayed in a Butcher tableau as follows

$$\begin{array}{c|c} & a_{ij} \\ \hline & b_i \end{array}, \tag{4.4}$$

and we will sometimes use the notation (a_{ij}, b_i). For more details about the order conditions of Runge-Kutta methods in the context of initial value ODEs, we refer for example to the book [28, Chap. III]. We denote by $y_{k+1} = \Phi_h(y_k)$ the numerical flow of (4.3), while the time adjoint method Φ_h^* of Φ_h is the inverse map of the original method with reversed time step $-h$, i.e., $\Phi_h^* := \Phi_{-h}^{-1}$ [28, Sect. II.3]. We recall that the time adjoint of an s-stage Runge-Kutta method (a_{ij}, b_i) (4.3) is again an s-stage Runge-Kutta method with the same order of accuracy and its coefficients (a_{ij}^*, b_i^*) are given by

$$a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j} \text{ and } b_i^* = b_{s+1-i}, \text{ where } i,j = 1,\dots,s.$$

If we discretize (4.1) using a Runge-Kutta discretization as above we naturally get the following discrete optimization problem,

$$\begin{aligned} &\min\ \ \Psi(y_N); \quad \text{subject to:} \\ &y_{k+1} = y_k + h\sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \qquad y_{ki} = y_k + h\sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}), \end{aligned} \tag{4.5}$$

where $i = 1\dots,s,\ k = 0,\dots,N-1,$ and $y_0 = y^0$. We denote by p_{ode} the order of accuracy of the method (4.3) applied to the ODE problem (4.2) and by p_{oc} the order of the method (4.5) for solving the optimal control problem (4.1). Note that we always have $p_{oc} \le p_{ode}$. In general, $p_{oc} < p_{ode}$ because additional order conditions, described in [27, 16], have to be satisfied.

Let us denote by $H(u,y,p) := p^T f(u,y)$ the pseudo-Hamiltonian of the system where p is the Lagrange multiplier (or the costate) associated to the state y. Applying Pontryagin's

maximum (or minimum) principle, the first order optimality conditions of (4.1) are given by the following boundary value problem,

$$\begin{aligned} \dot{y}(t) &= f(u(t), y(t)) = \nabla_p H(u(t), y(t), p(t)), \\ \dot{p}(t) &= -\nabla y f(u(t), y(t)) p = -\nabla_y H(u(t), y(t), p(t)), \\ 0 &= \nabla_u H(u(t), y(t), p(t)). \\ t &\in [0, T], \quad y(0) = y^0, \quad p(T) = \nabla \Psi(y(T)). \end{aligned} \tag{4.6}$$

Applying a Runge-Kutta integrator naively to (4.6) as an initial value system of ODEs combined with the classical methodology of shooting methods, would lead to severe instability due to the forward in time integration of the costate equation. For instance, for the optimal control of a diffusion PDE problem such as $\partial_t y(t,x) = \Delta y(t,x) + u(t,x)$, where Δ is the Laplace operator, $y(t,x)$ is the state function, and $u(t,x)$ is the control function (see also the diffusion-convection PDE problem considered in Sect. 4.4.2), then the costate equation takes the form of a heat equation with the wrong sign, $\partial_t p(t,x) = -\Delta p(t,x)$, which is naturally unstable if integrated forward in time. This makes classical shooting methods not applicable in the context of stiff dissipative optimal control problems considered in this thesis. Alternatively, we consider a forward-backward iterative algorithm as described below (Algorithm 4.2.3).

Introducing Lagrange multipliers for the finite dimensional optimization problem (4.5), and supposing that $b_i \neq 0$ for all $i = 1 \ldots, s$, a calculation [27, 16] yields the following discrete optimality conditions

$$\begin{aligned} y_{k+1} &= y_k + h \sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \qquad y_{ki} = y_k + h \sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}), \\ p_{k+1} &= p_k - h \sum_{i=1}^{s} \hat{b}_i \nabla_y H(u_{ki}, y_{ki}, p_{ki}), \; p_{ki} = p_k - h \sum_{j=1}^{s} \hat{a}_{ij} \nabla_y H(u_{kj}, y_{kj}, p_{kj}), \\ 0 &= \nabla_u H(u_{ki}, y_{ki}, p_{ki}), \quad k = 0, \ldots, N-1 \quad i = 1, \ldots, s, \\ y_0 &= y^0, \quad p_N = \nabla \Psi(y_N) \end{aligned} \tag{4.7}$$

where the coefficients $\hat{b}_i$ and $\hat{a}_{ij}$ are defined by the following relations which, as observed in [16], correspond to the symplecticity conditions of partitioned Runge-Kutta methods for ODEs,

$$\hat{b}_i := b_i, \qquad \hat{a}_{ij} := b_j - \frac{b_j}{b_i} a_{ji}, \qquad i = 1, \ldots s, j = 1, \ldots, s. \tag{4.8}$$

Note that the vectors p_k and p_{ki} are the Lagrange multipliers associated to y_k and y_{ki} respectively. Assuming that the Hessian matrix $\nabla_u^2 H(u, y, p) \in \mathbb{R}^{m \times m}$ is invertible along the trajectory of the exact solution, by the implicit function theorem there exists a C^∞ function ϕ such that $u = \phi(y, p)$ and then (4.7) is equivalent to a partitioned Runge-Kutta (PRK) method. As noticed in [16], if we consider the problem (4.6) as a Hamiltonian system, with the Hamiltonian function $\mathcal{H}(y, p) := H(\Psi(y, p), y, p)$, then the obtained PRK (4.7) scheme is symplectic thanks to the relations (4.8).

Theorem 4.2.1 (Theorem 4.1 in [27]). *Consider a Runge-Kutta method (a_{ij}, b_i) of order p_{ode} for ODEs, where $b_i \neq 0$, for all $i = 1, \ldots, s$, applied to the optimal control problem*

(4.1). *Consider the optimality conditions* (4.6)*, and assume that* $\nabla_u^2 H(u,y,p)$ *is invertible in a neighborhood of the solution, then we have the following theorem. If we discretize* (4.6) *using an s-stage partitioned Runge-Kutta method* $(a_{ij},b_i)-(\hat{a}_{ij},\hat{b}_i)$ *of order* $p^\star_{ode}$ *for ODEs (as partitioned RK method), and the condition* (4.8) *is satisfied, then the order* p_{oc} *of* (4.5) *satisfies* $p_{oc}=p^\star_{ode}\leq p_{ode}$ *and the schemes* (4.5) *and* (4.7) *are equivalent. In particular, for* $p_{ode}\geq 2$*, equivalently* $\sum_{i=1}^s b_i=1$ *and* $\sum_{i,j=1}^s b_i a_{ij}=\frac{1}{2}$*, we get* $p^\star_{ode}\geq 2$ *and* $p_{oc}\geq 2$*.*

The proof of Theorem 4.2.1 relies on the commutativity of the following diagram [16, Sect. 2] which means that methods (4.5) and (4.7), colorredwhere (4.7) is a symplectic partitioned Runge-Kutta method, yield exactly the same outputs (up to round-off errors) if derived for Runge-Kutta discretizations of (4.1) and (4.6) respectively. We also refer to the article [50] where the role of symplectic partitioned Runge-Kutta methods involved in this commutative diagram is discussed.

$$\begin{array}{ccc} (4.1) & \xrightarrow{\text{discretization}} & (4.5) \\ \text{optimality conditions}\downarrow & & \downarrow\text{optimality conditions} \\ (4.6) & \xrightarrow[\text{discretization}]{} & (4.7) \end{array}$$

Remark that in (4.7), if the method (a_{ij},b_i) is explicit, then $(\hat{a}_{ij},\hat{b}_i)$ is in contrast an implicit method. Hence it is useful to consider the costate equation backward in time and use the time adjoint of $(\hat{a}_{ij},\hat{b}_i)$ which turns out to be explicit as shown in Proposition 4.3.1 in Section 4.3. Indeed, consider method (4.7) and proceed as in [27, 30],

$$p_{k+1}+h\sum_{i=1}^s \hat{b}_i\nabla_y H(u_{ki},y_{ki},p_{ki})=p_{ki}+h\sum_{j=1}^s \hat{a}_{ij}\nabla_y H(u_{kj},y_{kj},p_{kj}),$$

we then deduce from the identity $b_j-\hat{a}_{ij}=\frac{b_j}{b_i}a_{ji}$ the following formulation where p_N serves to initialize the induction on $k=N-1,\ldots,0$,

$$p_k=p_{k+1}+h\sum_{i=1}^s b_i\nabla_y H(u_{ki},y_{ki},p_{ki}),\quad p_{ki}=p_{k+1}+h\sum_{j=1}^s \frac{b_j}{b_i}a_{ji}\nabla_y H(u_{kj},y_{kj},p_{kj}).$$

The above Runge-Kutta method $(\tilde{a}_{ij},\tilde{b}_i):=(\frac{b_j}{b_i}a_{ji},b_i)$ for the costate is in fact the time adjoint of $(\hat{a}_{ij},\hat{b}_i)$. Since the method $(\hat{a}_{ij},\hat{b}_i)$ is called in the literature the adjoint method in the sense of optimal control because it is applied to the adjoint equation (costate), we call the method $(\tilde{a}_{ij},\tilde{b}_i):=(\frac{b_j}{b_i}a_{ji},b_i)$ the **double adjoint** of (a_{ij},b_i), and we rewrite method (4.7) as

$$
\begin{aligned}
y_{k+1} &= y_k + h\sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \quad k = 0, \dots, N-1 \\
y_{ki} &= y_k + h\sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}), \quad k = 0, \dots, N-1 \quad i = 1, \dots, s \\
p_k &= p_{k+1} + h\sum_{i=1}^{s} \tilde{b}_i \nabla_y H(u_{ki}, y_{ki}, p_{ki}), \quad k = N-1, \dots, 0 \\
p_{ki} &= p_{k+1} + h\sum_{j=1}^{s} \tilde{a}_{ij} \nabla_y H(u_{kj}, y_{kj}, p_{kj}), \quad k = N-1, \dots, 0 \quad i = s, \dots, 1 \\
0 &= \nabla_u H(u_{ki}, y_{ki}, p_{ki}), \qquad k = 0, \dots, N-1 \quad i = 1, \dots, s \\
y_0 &= y^0, \qquad p_N = \nabla\Psi(y_N).
\end{aligned}
\tag{4.9}
$$

Note that we integrate the state forward in time (increasing indices k) and the costate backward in time (decreasing k).

An immediate consequence of Theorem 4.2.1 is that applying naively a Runge-Kutta method yields in general an order reduction, as stated in the following remark.

Remark 4.2.2. *Consider a Runge-Kutta method (a_{ij}, b_i) of order $p_{ode} = 2$ and define $(\tilde{a}_{ij}, \tilde{b}_i) := (a_{ij}, b_i)$, in general the obtained partitioned Runge-Kutta method (4.9) is not of order $p_{oc} = 2$. Indeed, the coupling order conditions $\sum_{i,j=1}^{s} b_i \hat{a}_{ij} = \frac{1}{2}$ and $\sum_{i,j=1}^{s} \hat{b}_i a_{ij} = \frac{1}{2}$ are not automatically satisfied in general. In particular, for (a_{ij}, b_i) being the standard order two RKC method studied in the next section below, it can be checked that $p_{oc} = 1$. This makes non trivial the construction of an explicit stabilized scheme of order 2 for optimal control problems, as described in section 4.3.3. We will see that the notion of double adjoint of a Runge-Kutta method, as described above, is an essential tool in our study.*

To implement (4.9) (equivalent to (4.7)), we shall use the following classical iterative algorithm which was proposed as a parallel algorithm with N sub-problems in [43, Algo. 4]. For simplicity of the presentation, we only recall the non parallel algorithm, but emphasize that the parallel version could also be used in our context with explicit stabilized schemes.

Algorithm 4.2.3. *(see for instance [43, Algo. 4]). First start with an initial guess for the internal stages of the control $U^0 = (u^0_{ki})^{i=1,\dots,s}_{k=0\dots,N-1}$ where $u^0_{ki} \in \mathbb{R}^m$ for all k and i. Denote by $y^l = (y^l_k)_{k=0,\dots,N-1}$ and $Y^l = (y^l_{ki})^{i=1,\dots,s}_{k=0\dots,N-1}$ the collection of the state values and its internal stages respectively at iteration l, and analogously we use the notations p^l and P^l for costate and its internal stages at iteration l. Suppose that at the iteration l, U^l is known. For the next iteration $l+1$, the computation of U^{l+1} is achieved as follows.*

1. *Compute Y^l, P^l, y^l, p^l as in (4.9), the computation is done forward in time for the state y_k and backward in time for the costate p_k.*

2. *Compute $\tilde{u}^{l+1}_{ki}$ solving the system $\nabla_u H(\tilde{u}^{l+1}_{ki}, y^l_{ki}, p^l_{ki}) = 0$, for all k and i using an analytical formula if available, or a Newton method for instance.*

3. *Denoting* $\tilde{U}^l = (\tilde{u}^l_{ki})^{i=1,\ldots,s}_{k=0\ldots,N-1}$, *define* U^{l+1} *by* $U^{l+1} = (1-\theta^l)U^l + \theta^l\tilde{U}^{l+1}$, *where* θ^l *is defined to minimize the scalar function* $\theta \mapsto \Psi(U^{l+1})$ *using a simple trisection method for instance, where* $\Psi : U \mapsto \Psi(y^N)$.

We stop when $\|U^{l+1} - U^l\| \leq tol$, *where tol is a prescribed small tolerance.*

In the recent paper [42], it was shown that the forward-backward sweep iteration defined in Algorithm 4.2.3 used in implementing discretized optimal control problems converges when using a symplectic Runge-Kutta discretization, which strengthens the interest of such symplectic methods.

For simplicity, we assume in the rest of the chapter that $y_0 = y^0$ always holds in (4.7).

4.2.2 Explicit stabilized methods

Stability is a crucial property of numerical integrators for solving stiff problems and we refer to the book [29]. A Runge-Kutta method is said to be stable if the numerical solution stays bounded along the integration process. Applying a Runge-Kutta method (4.3) to the linear test problem (with fixed parameter $\lambda \in \mathbb{C}$),

$$\dot{y} = \lambda y, \quad y(0) = y_0, \tag{4.10}$$

with stepsize h yields a recurrence of the form $y_{k+1} = R(h\lambda)y_k$ and by induction we get $y_k = R(h\lambda)^k y_0$. The function $R(z)$ is called the stability function of the method and the stability domain is defined as $\mathcal{S} := \{z \in \mathbb{C}; |R(z)| \leq 1\}$, and y_k remains bounded if and only if $h\lambda \in \mathcal{S}$. The same result also applies to the internal stages of the Runge-Kutta method, for all $i = 1, \ldots, s$, where s is the number of internal stages, $y_{ki} = R_i(h\lambda)y_k$, for some function R_i. Remark that $R(z)$ is a rational function for implicit methods, but in the case of explicit methods the stability function $R(z)$ reduces to a polynomial. The simplest Runge-Kutta type method to integrate ODEs (4.2) is the explicit Euler method $y_{k+1} = y_k + hf(y_k)$ with stability polynomial $R(z) = 1 + z$. However, its stability domain $\mathcal{S}$ is small (it reduces to the disc of center -1 and radius 1 in the complex plane) which yields a severe time step restriction and makes it very expensive for stiff problems.

4.2.2.1 Optimal first order Chebyshev methods

The idea of explicit stabilized methods (as introduced in [54], see the survey [4]) is to construct explicit Runge-Kutta integrators with extended stability domain that grows quadratically with the number of stages s of the method along the negative real axis, and hence allows to use large time steps typically for problems arising from diffusion partial differential equations. The family of methods considered in [54] is known as "Chebyshev methods" since its construction relies on Chebyshev polynomials $T_s(x)$ satisfying $T_s(\cos\theta) = \cos(s\theta)$. These polynomials allow us to obtain a two-step recurrence formula and hence low memory requirements and good internal stability with respect to round-off errors. The order one Chebyshev method for solving a stiff ODE (4.2) is defined as an explicit s-stage Runge-Kutta method by the recurrence

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_k + \mu_1 hF(y_{k0}), \\ y_{ki} &= \mu_i hF(y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i)y_{k,i-2}, \quad j = 2, \ldots, s \\ y_{k+1} &= y_{ks}, \end{aligned} \tag{4.11}$$

where

$$\omega_0 := 1 + \frac{\eta}{s^2}, \quad \omega_1 := \frac{T_s(\omega_0)}{T_s'(\omega_0)}, \quad \mu_1 := \frac{\omega_1}{\omega_0}, \tag{4.12}$$

where η is called the damping parameter and is used to make the stability of the method robust with respect to small perturbations as described below. Finally, for all $i = 2, \ldots, s$,

$$\mu_i := \frac{2\omega_1 T_{i-1}(\omega_0)}{T_i(\omega_0)}, \quad \nu_i := \frac{2\omega_0 T_{i-1}(\omega_0)}{T_i(\omega_0)}. \tag{4.13}$$

One can easily check that the (family) of methods (4.11) has the same first order of accuracy as the explicit Euler method (recovered for $s = 1$). Note that instead of the standard Runge-Kutta method formulation (4.3) with coefficients (a_{ij}, b_i), the one step method $y_{k+1} = \Phi_h(y_k)$ in (4.11) should be implemented using a recurrence relation (indexed by j) inspired from the relation (4.14) on Chebyshev polynomials

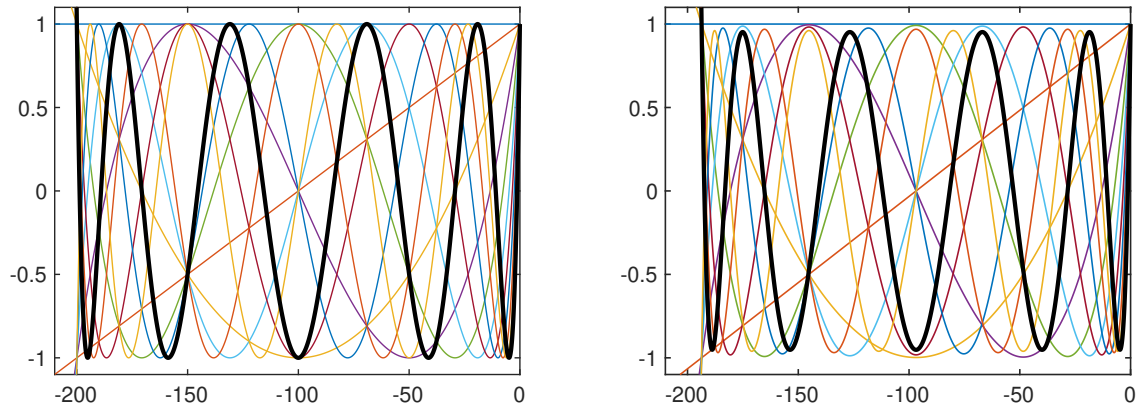
$$T_j(z) = 2zT_{j-1}(z) - T_{j-2}(z), \tag{4.14}$$

where $T_0(z) = 1, T_1(z) = z$. This implementation (4.11) yields a good stability [54] of the scheme with respect to round-off errors. The most interesting feature of this scheme is its stability behavior. Indeed, the method (4.11) applied to (4.10) yields, with $z = \lambda h$, $y_{k+1} = R_s^\eta(z) y_k = \frac{T_s(\omega_0+\omega_1 z)}{T_s(\omega_0)} y_k$. A large real negative interval $(-C_\eta s^2, 0)$ is included in the stability domain of the method $\mathcal{S} := \{z \in \mathbb{C}; |R_s^\eta(z)| \leq 1\}$. For the internal stages, we have analogously $y_{ki} = R_{s,i}^\eta(z) y_k = \frac{T_i(\omega_0+\omega_1 z)}{T_i(\omega_0)} y_k$. The constant $C_\eta = 2 - 4/3\,\eta + \mathcal{O}(\eta^2)$ depends on the so-called damping parameter η and for $\eta = 0$, it reaches the maximal value $C_0 = 2$ (also optimal with respect to all possible stability polynomials for explicit schemes of order 1). Hence, given the stepsize h, for dissipative vector fields with a Jacobian having large real negative eigenvalues (such as diffusion problems) with spectral radius $\lambda_{\max}$ at y_n, the parameter s for the next step y_{n+1} can be chosen adaptively as[1]

$$s := \left[\sqrt{\frac{h\lambda_{\max} + 1.5}{2 - 4/3\,\eta}} + 0.5 \right], \tag{4.15}$$

see [3] in the context of stabilized schemes of order two with adaptive stepsizes. The method (4.11) is much more efficient as its stability domain increases *quadratically* with the number s of function evaluations while a composition of s explicit Euler steps (same cost) has a stability domain that only increases *linearly* with s. In Figure 4.1 we plot the internal stages for $s = 10$ and different values $\eta = 0$ and $\eta = 0.05$, respectively. We observe that in the absence of damping ($\eta = 0$), the stability function (here a polynomial) is bounded by 1 in the large real interval $[-2s^2, 0]$ of width $2 \cdot 10^2 = 200$. However, for all z that are local extrema of the stability function, where $|R_s^\eta(z)| = 1$, the stability domain is very narrow in the complex plane. Here the importance of some damping appears, to make the scheme robust with respect to small perturbations of the eigenvalues. A typical recommended value for the damping parameter is $\eta = 0.05$, see [55, 4]. The advantage of this damping is that the stability polynomial is now **strictly** bounded by 1 and the stability domain includes a neighborhood of the negative interval $(-C_\eta s^2, 0)$. This improvement costs a slight reduction of the stability domain length from $2s^2$ to $C_\eta s^2$ where $C_\eta \geq 2 - \frac{4}{3}\eta$.

[1]The notation $[x]$ stands for the integer rounding of real numbers.



(a) $\eta = 0$ (b) $\eta = 0.05$

Figure 4.1: Internal stages (thin curves) and stability polynomials (bold curves) of the Chebyshev method (4.11) for $s = 10$ with and without damping.

4.2.2.2 Second order RKC methods

To design a second order method, we need the stability polynomial to satisfy[2]

$$R(z) = 1 + z + \frac{z^2}{2} + \mathcal{O}(z^3).$$

In [15], Bakker introduced a correction to the first order shifted Chebyshev polynomials to get the following second order polynomial

$$R_s^\eta(z) = a_s + b_s T_s(\omega_0 + \omega_2 z), \tag{4.16}$$

where,

$$a_s = 1 - b_s T_s(\omega_0), \quad b_s = \frac{T_s''(\omega_0)}{(T_s'(\omega_0)^2)}, \quad \omega_0 = 1 + \frac{\eta}{s^2}, \quad \omega_2 = \frac{T_s'(\omega 0)}{T_s''(\omega_0)}, \quad \eta = 0.15. \tag{4.17}$$

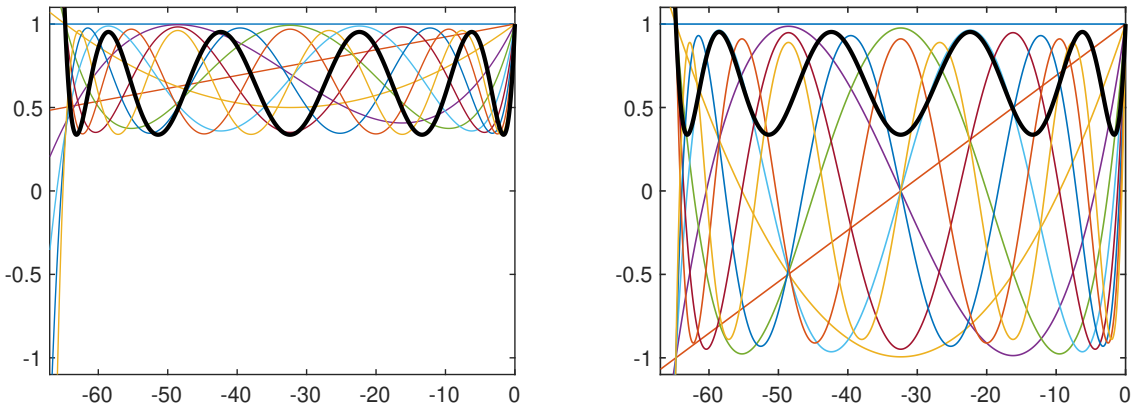
For each s, $|R_s^\eta(z)|$ remains bounded by $a_s + b_s = 1 - \eta/3 + \mathcal{O}(\eta^2)$ for z in the stability interval (except for a small interval near the origin). The stability interval along the negative real axis is approximately $[-0.65s^2, 0]$, and covers about 80% of the optimal stability interval for second order stability polynomials, and the formula now for calculating s for a given time step h is

$$s := \left\lceil \sqrt{\frac{h\lambda_{\max} + 1.5}{0.65}} + 0.5 \right\rceil. \tag{4.18}$$

Using the recurrence relation of the Chebyshev polynomials, the RKC method as introduced in [54] is defined by

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_{k0} + h b_1 \omega_2 f(y_{k0}), \\ y_{ki} &= y_{k0} + \mu_i' h (f(y_{k,i-1}) - a_{i-1} f(y_{k0})) + \nu_i' (y_{k,i-1} - y_{k0}) \\ &\quad + \kappa_i' (y_{k,i-2} - y_{k0}), \\ y_{k+1} &= y_{ks}, \end{aligned} \tag{4.19}$$

[2]Indeed, up to order two, the order conditions for nonlinear problems are the same as the order conditions for linear problems [28, Chap. III].



(a) Classical RKC (4.19). (b) New RKC (4.21).

Figure 4.2: Internal stages (thin curves) and stability polynomials (bold curves) of the classical (4.19) and the new (4.21) RKC implementations for $s = 10$ internal stages.

where,

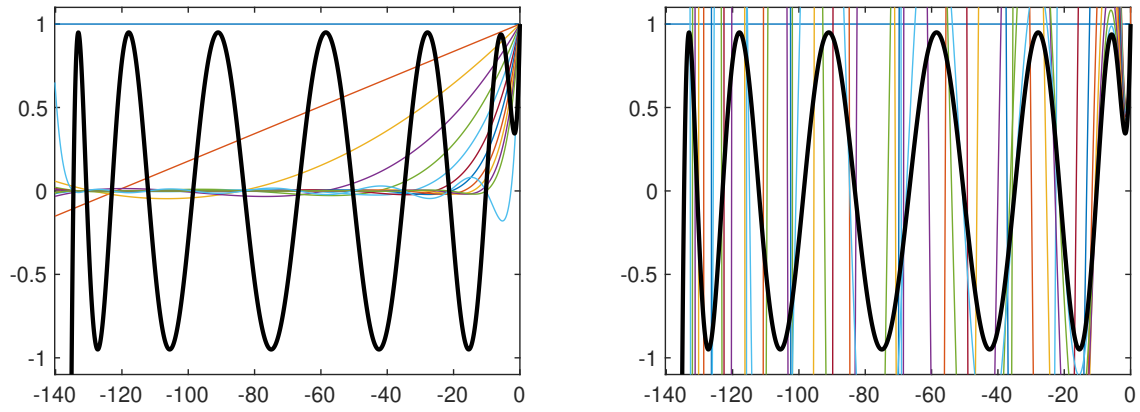
$$\mu_i' = \frac{2b_i\omega_2}{b_{i-1}}, \quad \nu_i' = \frac{2b_i\omega_0}{b_{i-1}}, \quad \kappa_i' = -\frac{b_i}{b_{i-2}}, \quad b_i = \frac{T_i''(\omega_0)}{(T_i'(\omega_0)^2)}, \quad a_i = 1 - b_i T_i(\omega_0), \tag{4.20}$$

for $i = 2, \ldots, s$. As in (4.16), the stability functions of the internal stages are given by $R_i^\eta(z) = a_i + b_i T_i(\omega_0 + \omega_2 z)$, where $i = 0, \ldots, s-1$, and the parameters a_i and b_i are chosen such that the above stages are consistent $R_i^\eta(z) = 1 + \mathcal{O}(z)$. The parameters b_0 and b_1 are free ($R_0^\eta(z)$ is constant and only order 1 is possible for R_1^η(z)) and the values $b_0 = b_1 = b_2$ are suggested in [52]. In this chapter, to facilitate the analysis of the internal stability of the optimal control methods, making the internal stages of the RKC method analogous to the Chebyshev method (4.11) of order one, we introduce a new implementation of RKC method

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_k + \mu_1 h F(y_{k0}), \\ y_{ki} &= \mu_i h F(y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \ldots, s \\ y_{k+1} &= a_s y_{k0} + b_s T_s(\omega_o) y_{ks}, \end{aligned} \tag{4.21}$$

where $\mu_1 = \frac{\omega_2}{\omega_0}$, a_s, b_s are given in (4.17), and the parameters μ_i and ν_i are defined by (analogously to (4.13), using ω_2 instead of ω_1), $\mu_i = \frac{2\omega_2 T_{i-1}(\omega_0)}{T_i(\omega_0)}$, $\nu_i = \frac{2\omega_0 T_{i-1}(\omega_0)}{T_i(\omega_0)}$, for $i = 2, \ldots, s$. This new formulation (4.21) yields the same stability function $R_s^\eta(z)$ in (4.16) but different internal stages, and it will be helpful when we introduce the double adjoint of RKC in Section 4.3. We recall that for an accurate implementation, one should not use the standard Runge-Kutta formulations with coefficients (a_{ij}, b_i) for (4.11) and (4.19) since they are unstable due to the accumulated round-off error for large values of s. In contrast, the *low memory* induction formulations (4.11) and (4.19) are easy to implement and very stable with respect to round-off errors [54].

Note that (4.21) is not the same Runge-Kutta method as the standard RKC (4.19) from [54], it has different internal stages but the same stability function $R_s^\eta(z)$ in (4.16) and hence order $p_{ode} = 2$. In Figure 4.2 we can see that the internal stages of the new formulation



(a) Classical ROCK2.

(b) Double adjoint of ROCK2.

Figure 4.3: Internal stages (thin curves) and stability polynomials (bold curves) of the ROCK2 method [9] and its double adjoint for optimal control for $s = 13$ stages.

(4.21) of RKC have an analogous behavior compared to those of the first order Chebyshev method oscillating around zero (see Figure 4.1), in contrast to those of the standard RKC (4.19), oscillating around the value $a_s > 0$. In addition, comparing Figures 4.2b and 4.1b, we see that the internal stages of the new RKC method (4.21) are the same as the order one Chebyshev method (4.11) up to a horizontal rescaling ω_2/ω_1. This is because the s internal stages of the methods have the stability function $T_i(\omega_0 + \omega_j z)/T_i(\omega_0),\ i = 1, \dots, s$ for $j = 1, 2$ respectively.

Notice however that this modification of the standard RKC method (4.19) deteriorates the order two of accuracy of the internal stages of the method, useful for PDEs with non homogeneous boundary conditions [34, Chap. V].

Remark 4.2.4. *Analogously to the standard RKC method* (4.19)*, the new RKC formulation* (4.21) *can be equipped with an error estimator to allow a variable time step control. Since the new formulation* (4.21) *has the same stability function, one can use the same error estimator as proposed in [51, Sect. 3.1]. In this chapter we consider only a constant time step for simplicity of the presentation but emphasize that a variable time step h_n can be used for the new optimal control method of order two.*

Remark 4.2.5. *Second order Runge-Kutta Orthogonal Chebyshev (ROCK2) methods, as introduced in [9], are second order explicit stabilized methods for which the stability domain contains an interval that covers around 98% of the optimal one for second order explicit methods. It would be interesting to extend such second order methods with nearly optimally large stability domain to the context of optimal control problems. It turns out however that such an extension based on ROCK2 (or its order four extension ROCK4 [3]) is difficult and not analyzed in the present thesis. This difficulty arises from the severe instability of the internal stages of the double adjoint of the standard ROCK2 method (see Figure 4.3) which would introduce large round-off errors for stiff problems (large values of s), making the obtained optimal control method not reliable.*

4.3 Explicit stabilized methods for optimal control

In this section, we derive new two term recurrence relations of the double adjoints of Chebyshev method (4.11) and RKC method (4.21) that are numerically stable. Indeed, one cannot rely on standard Runge-Kutta coefficients for the implementation of explicit stabilized schemes.

4.3.1 Double adjoint of a general Runge-Kutta method

Recall from (4.9) that the Butcher tableau of the double adjoint $(\tilde{a}_{ij}, \tilde{b}_i)$ of (4.4) is

$$\begin{array}{c|c} \tilde{a}_{ji} \\ \hline \tilde{b}_i \end{array} := \begin{array}{c|c} \frac{b_j}{b_i} a_{ji} \\ \hline b_i \end{array}. \tag{4.22}$$

Proposition 4.3.1. *If a Runge-Kutta method (a_{ij}, b_i) (4.4) is explicit, then its double adjoint (4.22) is explicit as well.*

Proof. For an explicit Runge-Kutta method we have that $a_{ij} = 0$ for all $j \geq i$ i.e the matrix (a_{ij}) is strictly lower triangular. Permuting the internal stages in (4.22) for $i, j = s, \dots, 2, 1$ does not modify the method but yields the following Butcher tableau

$$\begin{array}{c|c} \frac{b_{s+1-j}}{b_{s+1-i}} a_{s+1-j,s+1-i} \\ \hline b_{s+1-i} \end{array} \tag{4.23}$$

which is strictly lower triangular, and thus the method (4.22) is again explicit. □

An immediate consequence of Proposition 4.3.1 is that for explicit methods, the stability function of the double adjoint is again a polynomial. In fact it turns out, as stated in Theorem 4.3.2 below, that for any Runge-Kutta method, the double adjoint $(\tilde{a}_{ij}, \tilde{b}_i)$ has exactly the same stability function as (a_{ij}, b_i). Note however that this result does not hold in general for the internal stages (see Remark 4.2.5 about ROCK2).

Theorem 4.3.2. *A Runge-Kutta method (a_{ij}, b_i) and its double adjoint $(\tilde{a}_{ij}, \tilde{b}_i)$ in (4.22) share the same stability function $R(z)$.*

Proof. Let $A = (a_{ij})$, $A_d = (a_{ji}\, b_j / b_i)$, and $b = (b_i)$, $i, j = 1 \dots s$. We recall the formula for the stability function of the Runge-Kutta method (a_{ij}, b_i),

$$R(z) = \frac{\det(I - zA + z\mathbb{1}b^T)}{\det(I - zA)}, \tag{4.24}$$

where $\mathbb{1} \in \mathbb{R}^s$ is the line vector of size s containing only ones. Using a simple calculation, one can show that $A_d^T = DAD^{-1}$ where $D = \mathrm{diag}(b_i)$. This implies that $I - zA_d^T = D(I - zA)D^{-1}$, and since $I - zA_d^T = (I - zA_d)^T$, thus $\det(I - zA) = \det(I - zA_d)$. Using the same diagonal matrix D we have $D\mathbb{1}b^T D^{-1} = \mathbb{1}^T b$, hence $I - zA_d^T + z\mathbb{1}b = D(I - zA + z\mathbb{1}b)D^{-1}$ and $\det(I - zA_d^T + z\mathbb{1}b) = \det(I - zA + z\mathbb{1}b)$ and hence the stability function of $(\tilde{a}_{ij}, \tilde{b}_i)$ is again (4.24). □

4.3.2 Chebyshev method of order one for optimal control problems

For clarity of presentation, we first study Chebyshev method of order one for optimal control problems before introducing the second order RKC method. Applying the order one Chebyshev method (4.11) to the problem (4.1) we get

$$\begin{aligned} &\min \Psi(y_N),\ \textit{such that}\\ &y_{k0} = y_k, \quad y_{k1} = y_{k0} + \mu_1 h f(u_{k0}, y_{k0}),\\ &y_{ki} = \mu_i h f(u_{k,i-1}, y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \dots, s\\ &y_{k+1} = y_{ks}, \end{aligned} \tag{4.25}$$

where, $k = 0, \dots, N-1,\ \eta > 0$ is fixed, and the parameters $\mu_i,\ \nu_i$ are defined as in (4.12) and (4.13).

For the implementation of Algorithm 4.2.3 based on the order one Chebyshev method (4.25) for the state equation, the costate equation can be implemented efficiently using the recurrence relations given by the following theorem.

Theorem 4.3.3. *The double adjoint of scheme* (4.25) *is given by the recurrence*

$$\begin{aligned} p_N &= \nabla\Psi(y_N), \quad p_{ks} = p_{k+1}\\ p_{k,s-1} &= p_{ks} + \frac{\mu_s}{\nu_s} h \nabla_y H(u_{k,s-1}, y_{k,s-1}, p_{ks})\\ p_{k,s-j} &= \frac{\mu_{s-j+1}\alpha_{s-j+1}}{\alpha_{s-j}} h \nabla_y H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1})\\ &\quad + \frac{\nu_{s-j+1}\alpha_{s-j+1}}{\alpha_{s-j}} p_{k,s-j+1}\\ &\quad + \frac{(1-\nu_{s-j+2})\alpha_{s-j+2}}{\alpha_{s-j}} p_{k,s-j+2}, \quad j = 2\dots, s-1,\\ p_{k0} &= \mu_1\alpha_1 h \nabla_y H(u_{k0}, y_{k0}, p_{k1}) + \alpha_1 p_{k1} + (1-\nu_2)\alpha_2 p_{k2}\\ p_k &= p_{k0}\\ &\nabla_u H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) = 0, \quad j = 1, \dots, s. \end{aligned} \tag{4.26}$$

where $k = N-1, \dots, 2, 1, 0$ *and the coefficients* α_j *are defined by induction as*

$$\begin{aligned} \alpha_s &= 1, \quad \alpha_{s-1} = \nu_s,\\ \alpha_{s-j} &= \nu_{s-j+1}\alpha_{s-j+1} + (1-\nu_{s-j+2})\alpha_{s-j+2}, \quad j = 2\dots s-1. \end{aligned} \tag{4.27}$$

The proof of Theorem 4.3.3 uses similar arguments to the proof of Theorem 4.2.1, with the exception that we now rely on the recurrence formula (4.25) instead of the standard Runge-Kutta formulation (4.6) to avoid numerical instability.

Proof of Theorem 4.3.3. The Lagrangian associated to the discrete optimization problem (4.25) is

$$\begin{aligned}\mathcal{L} = \Psi(y_N) + p_0 \cdot (y^0 - y_0) + \sum_{k=0}^{N-1} \Big\{ & p_{k+1} \cdot (y_{ks} - y_{k+1}) - p_{k0} \cdot (y_k - y_{k0}) \\ & + p_{k1} \cdot (y_{k0} + \mu_1 h f(u_{k0}, y_{k0}) - y_{k1}) \\ & + \sum_{i=2}^{s} p_{ki} \cdot (\mu_i h f(u_{k,i-1}, y_{k,i-1}) + \nu_i y_{k,i-1} \\ & + (1-\nu_i) y_{ki-2} - y_{ki}) \Big\}.\end{aligned}$$

Here p_{k+1}, p_{ki}, and p_0 are the Lagrange multipliers. The optimality necessary conditions are thus given by

$$\frac{\partial \mathcal{L}}{\partial y_k} = 0, \quad \frac{\partial \mathcal{L}}{\partial y_{ki}} = 0, \quad \frac{\partial \mathcal{L}}{\partial p_k} = 0, \quad \frac{\partial \mathcal{L}}{\partial p_{ki}} = 0, \quad \frac{\partial \mathcal{L}}{\partial u_{ki}} = 0, \tag{4.28}$$

where $k = 0, \dots, N-1$ and $i = 0, \dots, s$. By a direct calculation, we obtain the following system,

$$\begin{aligned} y_{k0} &= y_k, \quad y_{k1} = y_{k0} + \mu_1 h f(u_{k0}, y_{k0}), \\ y_{ki} &= \mu_i h f(u_{k,i-1}, y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \dots, s, \\ y_{k+1} &= y_{ks}, \\ p_N &= \nabla \Psi(y_N), \quad p_{ks} = p_{k+1}, \\ p_{k,s-1} &= \mu_s h \nabla_y H(u_{k,s-1}, y_{k,s-1}, p_{ks}) + \nu_s p_{ks}, \\ p_{k,s-j} &= \mu_{s-j+1} h \nabla_y H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) + \nu_{s-j+1} p_{k,s-j+1} \\ &\quad + (1-\nu_{s-j+2}) p_{k,s-j+2}, \quad j = 2, \dots, s-1, \\ p_{k0} &= \mu_1 h \nabla_y H(u_{k0}, y_{k0}, p_{k1}) + p_{k1} + (1-\nu_2) p_{k2}, \\ p_k &= p_{k0}, \\ \nabla_u H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) &= 0, \quad j = 1, \dots, s, \end{aligned} \tag{4.29}$$

where $k = 0, \dots, N-1$. In the above system, observe that the steps p_{ki} of the double adjoint are not internal stages of a Runge-Kutta method, that is because they are not $\mathcal{O}(h)$ perturbations of the p_{k+1}, i.e. $p_{ki} \neq p_{k+1} + \mathcal{O}(h)$, for instance, already for the first step $p_{k,s-1} = \nu_s p_{k+1} + \mathcal{O}(h)$ with $\nu_s = 2 + \mathcal{O}(\eta)$. Since the pseudo-Hamiltonian $H(u, y, p)$ is linear in p, we can rescale the internal stages of the costate by a factor α_j such that for $\hat{p}_{kj} := \alpha_j^{-1} p_{kj}$, we obtain $\hat{p}_{kj} = p_{k+1} + \mathcal{O}(h)$. We define

$$\hat{p}_{ks} := p_{ks}, \quad \hat{p}_{k,s-1} := \frac{p_{k,s-1}}{\nu_s} = \hat{p}_{ks} + \frac{\mu_s}{\nu_s} h \nabla_y H(u_{k,s-1}, y_{k,s-1}, \hat{p}_{ks}).$$

Substituting $\hat{p}_{k,s-2}$ in (4.29), we obtain

$$p_{k,s-2} = \mu_{s-1}\nu_s h \nabla_y H(u_{k,s-2}, y_{k,s-2}, \hat{p}_{k,s-1}) + \nu_{s-1}\nu_s \hat{p}_{k,s-1} + (1-\nu_s)\hat{p}_{ks},$$

the quantities $\hat{p}_{k,s-1}$ and $\hat{p}_{ks}$ are equal to $p_{k+1} + \mathcal{O}(h)$, hence $p_{k,s-2} = (\nu_{s-1}\nu_s + 1 - \nu_s) p_{k+1} + \mathcal{O}(h)$, this implies that $\alpha_{s-2} = \nu_{s-1}\nu_s + 1 - \nu_s$ and therefore

$$\hat{p}_{k,s-2} := \frac{p_{k,s-2}}{\nu_{s-1}\nu_s + 1 - \nu_s} = \frac{p_{k,s-2}}{\nu_s(\nu_{s-1} - 1) + 1}.$$

Following this procedure for $p_{k,s-j}, j = 2, \ldots, s-1$, we arrive at the Runge-Kutta formulation (4.26) of the double adjoint of scheme (4.25), where we go back to the notation p_{ki} instead of $\hat{p}_{ki}$. □

Remark 4.3.4. *A straightforward calculation yields that without damping (for $\eta = 0$), we have $\alpha_{s-j} = j + 1$, and method* (4.25)-(4.26) *reduces to the following recurrence*

$$
\begin{aligned}
y_{k0} &= y_k, \quad y_{k1} = y_{k0} + \frac{h}{s^2} f(u_{k0}, y_{k0}),\\
y_{ki} &= \frac{2h}{s^2} f(u_{k,i-1}, y_{k,i-1}) + 2y_{k,i-1} - y_{k,i-2}, \quad i = 2, \ldots, s,\\
y_{k+1} &= y_{ks},\\
p_N &= \nabla\Psi(y_N), \quad p_{ks} = p_{k+1},\\
p_{k,s-1} &= p_{ks} + \frac{h}{s^2} \nabla_y H(u_{k,s-1}, y_{k,s-1}, p_{ks}),\\
p_{k,s-j} &= \frac{2j}{(j+1)s^2} h \nabla_y H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) + \frac{2j}{j+1} p_{k,s-i+1}\\
&\quad + \frac{1-j}{j+1} p_{k,s-j+2}, \quad j = 2, \ldots, s-1,\\
p_{k0} &= \frac{h}{s} \nabla_y H(u_{k0}, y_{k0}, p_{k1}) + s p_{k1} + (1-s) p_{k2},\\
p_k &= p_{k0},\\
&\nabla_u H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) = 0, \quad j = 1, \ldots, s.
\end{aligned}
\tag{4.30}
$$

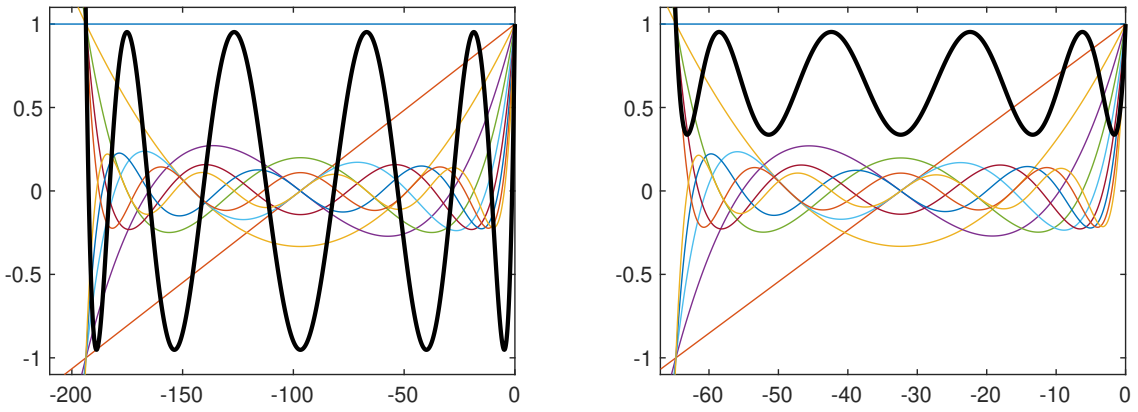
where $k = 0, \ldots, N-1$. In Section 4.3.4, we shall study the stability of (4.30) *(without damping) and of* (4.25)-(4.26) *(with damping).*

4.3.3 RKC method of order 2

We consider the new implementation (4.21) of the RKC method applied to (4.1) for which the internal stages behave similarly to that of the order one method, given by

$$
\begin{aligned}
&\min \Psi(y_N), \ \textit{such that}\\
&y_{k0} = y_k, \quad y_{k1} = y_{k0} + \mu_1 h f(u_{k0}, y_{k0}),\\
&y_{ki} = \mu_i h f(u_{k,i-1}, y_{k,i-1}) + \nu_i y_{k,i-1} + (1-\nu_i) y_{k,i-2}, \quad i = 2, \ldots, s\\
&y_{k+1} = a_s y_{k0} + b_s T_s(\omega_0) y_{ks},
\end{aligned}
\tag{4.31}
$$

where, $k = 0, \ldots, N-1$, $\eta = 0.15$, and again all the parameters are defined as for the Chebyshev method using ω_2 instead of ω_1. The order two RKC method with formulation (4.31) for the state equation can be implemented using Algorithm 4.2.3, the costate equation being implemented using the recurrence relations given by the following Theorem 4.3.5. Its proof is analogous to that of Theorem 4.3.3 and thus omitted.



(a) Double adjoint (4.26) for $\eta = 0.05$ (b) Double adjoint (4.32) for $\eta = 0.15$

Figure 4.4: Internal stages (thin curves) and stability polynomial (bold curve) of the double adjoint of the Chebyshev method (4.26) of order one and the RKC method (4.32) of order two for $s = 10$ internal stages.

Theorem 4.3.5. *The double adjoint of the scheme* (4.31) *is given by the recurrence*

$$\begin{aligned}
p_N &= \nabla\Psi(y_N), \quad p_{ks} = p_{k+1},\\
p_{k,s-1} &= p_{ks} + \frac{\mu_s}{\nu_s} h\nabla_y H(u_{k,s-1}, y_{k,s-1}, p_{ks}),\\
p_{k,s-j} &= \frac{\mu_{s-j+1}\alpha_{s-j+1}}{\alpha_{s-j}} h\nabla_y H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1})\\
&\quad + \frac{\nu_{s-j+1}\alpha_{s-j+1}}{\alpha_{s-j}} p_{k,s-j+1},\\
&\quad + \frac{(1-\nu_{s-j+2})\alpha_{s-j+2}}{\alpha_{s-j}} p_{k,s-j+2}, \quad j = 2, \dots, s-1,\\
p_{k0} &= \mu_1\alpha_1 h\nabla_y H(u_{k0}, y_{k0}, p_{k1}) + \alpha_1 p_{k1} + (1-\nu_2)\alpha_2 p_{k2} + a_s p_{k+1},\\
p_k &= p_{k0},\\
\nabla_u H(u_{k,s-j}, y_{k,s-j}, p_{k,s-j+1}) &= 0, \quad j = 1, \dots, s,
\end{aligned} \tag{4.32}$$

where the coefficients α_j are defined using the induction

$$\begin{aligned}
\alpha_s &= b_s T_s(\omega_0), \quad \alpha_{s-1} = \nu_s\alpha_s,\\
\alpha_{s-j} &= \nu_{s-j+1}\alpha_{s-j+1} + (1-\nu_{s-j+2})\alpha_{s-j+2}, \quad j = 2 \dots s-1.
\end{aligned} \tag{4.33}$$

In Figure 4.4, we plot the stability function and the internal stages of the double adjoint (4.26) of Chebyshev (4.11) and the double adjoint (4.32) of RKC (4.21). Comparing with Figures 4.1b and 4.2b, we observe that the internal stages are not the same for the double adjoint methods compared to the (4.11) and (4.21), while the stability function itself is identical as shown in Theorem 4.3.2 for a general Runge-Kutta method.

4.3.4 Stability and convergence analysis

In this section, we study the stability of the double adjoint of the Chebyshev method (4.25) and the RKC method (4.31). We recall that for the Chebyshev method of order one (resp. RKC of order two) the stability domain contains the interval $[-\beta(s,\eta),0]$ where $\beta_{Cheb}(s,\eta)\approx(2-4\eta/3)s^2$ (resp. $\beta_{RKC}(s,\eta)\approx 0.653s^2$ for $\eta=0.15$).

Theorem 4.3.6. *Consider the Chebyshev (4.25) and the RKC (4.31) methods. For $\eta=0$, the stability functions of the internal stages $R_{s,i}(z)$ of the Chebyshev (resp. RKC) double adjoint (4.26) (resp. (4.32)), are bounded by 1 for all $z\in[-2s^2,0]$ (resp. $[-\frac{2}{3}s^2+\frac{2}{3},0]$) and all $s\in\mathbb{N}$.*

The proof of Theorem 4.3.6 relies on the following lemma.

Lemma 4.3.7. *Let $s\geq 1$, and consider the double sequence $\tilde\gamma_j^i$ indexed by i and j,*

$$\begin{aligned}&\tilde\gamma_j^i=0\ \forall\ j>i, \qquad i=0,\dots,s-1,\\ &\tilde\gamma_0^0=1,\quad \tilde\gamma_0^1=0,\quad \tilde\gamma_1^1=2,\end{aligned} \tag{4.34}$$

$$\begin{aligned}&\tilde\gamma_0^i=\tilde\gamma_1^{i-1}-\tilde\gamma_0^{i-2},\quad \tilde\gamma_1^i=2\tilde\gamma_0^{i-1}+\tilde\gamma_2^{i-1}-\tilde\gamma_1^{i-2} \qquad i=2,\dots,s-1,\\ &\tilde\gamma_j^i=\tilde\gamma_{j-1}^{i-1}+\tilde\gamma_{j+1}^{i-1}-\tilde\gamma_j^{i-2} \qquad i=2,\dots,s-1,\ j=2,\dots,i,\end{aligned} \tag{4.35}$$

Then,

$$\begin{aligned}&\tilde\gamma_j^i=0\ \forall\ j>i,\\ &\tilde\gamma_0^i=\begin{cases}1 & \text{if } i \text{ is even}\\ 0 & \text{if } i \text{ is odd}\end{cases} \qquad \tilde\gamma_j^i=\begin{cases}2 & \text{if } i-j \text{ is even}\\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{4.36}$$

where $i=0,\dots,s-1,\ j=0,\dots,i$.

Proof. It can be checked that the coefficients defined in (4.36) verify the induction (4.35). Hence using the fact that they have the same initial terms (4.34), we conclude that they coincide by induction on i and j. □

Proof of Theorem 4.3.6. We first consider the Chebyshev method without damping applied to the linear test problem $y'=\lambda y$, $\lambda\in\mathbb{C},\ t\in(0,T],\ y(0)=1$, with a uniform subdivision $x_0=0<x_1<\cdots<x_N=T$ of stepsize h. Using Remark 4.3.4, we obtain for $k=0$:

$$\begin{aligned}&y_{k0}=1,\quad y_{k1}=y_{k0}+\frac{h\lambda}{s^2}y_{k0},\\ &y_{ki}=\frac{2ih\lambda}{(i+1)s^2}y_{k,i-1}+\frac{2i}{i+1}y_{k,i-1}+\frac{1-i}{i+1}y_{k,i-2},\quad i=2,\dots,s-1,\\ &y_{ks}=\frac{h\lambda}{s}y_{k,s-1}+sy_{k,s-1}+(1-s)y_{k,s-2},\\ &y_1=y_{ks}.\end{aligned} \tag{4.37}$$

First, notice that since $\frac{2i}{i+1}+\frac{1-i}{i+1}=1$, we have that for all $i=0,\dots,s$, $y_{ki}=(1+\mathcal{O}(h))$ (by induction). Setting $z=h\lambda$, it is sufficient to prove the identity

$$y_{ki}=\sum_{j=0}^{i}\gamma_j^iT_j(1+\frac{z}{s^2}) \tag{4.38}$$

where y_{ki} is a convex combination of the polynomials $T_j(1+\frac{z}{s^2})$,

$$\sum_{j=0}^{i} \gamma_j^i = 1 \text{ and } \gamma_j^i \geq 0 \ \forall \ i,j = 1,\dots s-1, \tag{4.39}$$

because $|T_j(1+\frac{z}{s^2})| \leq 1$ for all $j = 0,\dots,s$ and $z \in [-\beta(s,0),0] = [-2s^2,0]$.
Since the Chebyshev polynomials form a basis of the vector space of polynomials, this already justifies the existence of the expansion (4.38) with some real coefficients γ_j^i. The identity $\sum_{j=0}^{i} \gamma_j^i = 1$ follows from the fact that $y_{ki} = 1+\mathcal{O}(h)$ and $T_j(1+\frac{z}{s^2}) = 1+\mathcal{O}(h)$ for all $i,j = 0,\dots,s$. Now we can calculate these coefficients for the first two internal stages, $R_{s,0}(z) = y_{k0} = y_k = 1 = T_0(1+\frac{z}{s^2})$, thus $\gamma_0^0 = 1$. Analogously,

$$R_{s,1}(z) = y_{k1} = y_{k0} + \frac{h\lambda}{s^2} y_{k0} = 1 + \frac{z}{s^2} = T_1(1+\frac{z}{s^2}) \text{ we obtain } \gamma_0^1 = 0,\ \gamma_1^1 = 1.$$

It remains to prove the positivity of the coefficients γ_j^i. Coupling (4.38) and (4.39), we obtain

$$\begin{aligned}
R_{s,i}(z) = y_{ki} &= \frac{2i}{i+1}(1+\frac{z}{s^2}) y_{k,i-1} + \frac{1-i}{i+1} y_{k,i-2} \\
&= \frac{2i}{i+1}(1+\frac{z}{s^2}) \sum_{j=0}^{i-1} \gamma_j^{i-1} T_j(1+\frac{z}{s^2}) + \frac{1-i}{i+1} \sum_{j=0}^{i-2} \gamma_j^{i-2} T_j(1+\frac{z}{s^2}) \\
&= \frac{2i}{i+1} \gamma_0^{i-1} T_1(1+\frac{z}{s^2}) + \sum_{j=2}^{i} \frac{i}{i+1} \gamma_{j-1}^{i-1} T_j(1+\frac{z}{s^2}) \\
&+ \sum_{j=2}^{i} \left(\frac{i}{i+1} \gamma_{j-1}^{i-1} - \frac{i-1}{i+1} \gamma_{j-2}^{i-2} \right) T_{j-2}(1+\frac{z}{s^2})
\end{aligned}$$

where we used (4.14). By comparison with (4.38), we obtain the following recurrence

$$\begin{aligned}
&\gamma_j^i = 0 \ \forall \ j > i, \qquad i = 0,\dots,s-1, \\
&\gamma_0^0 = 1, \quad \gamma_0^1 = 0, \quad \gamma_1^1 = 1, \quad \gamma_0^i = \frac{i}{i+1}\gamma_1^{i-1} - \frac{i-1}{i+1}\gamma_0^{i-2} \qquad i = 2,\dots,s-1, \\
&\gamma_1^i = \frac{2i}{i+1}\gamma_0^{i-1} + \frac{i}{i+1}\gamma_2^{i-1} - \frac{i-1}{i+1}\gamma_1^{i-2} \qquad i = 2,\dots,s-1, \\
&\gamma_j^i = \frac{i}{i+1}\gamma_{j-1}^{i-1} + \frac{i}{i+1}\gamma_{j+1}^{i-1} - \frac{i-1}{i+1}\gamma_j^{i-2} \qquad i = 2,\dots,s-1,\ j = 2,\dots,i.
\end{aligned}$$

Now defining $\tilde{\gamma}_j^i = (i+1)\gamma_j^i$, the above induction relations simplify to (4.34) and (4.35). The positivity of $\tilde{\gamma}_j^i$, and hence of γ_j^i, follows from Lemma 4.3.7. For $i = s$, the stability is a consequence of Theorem 4.3.2.

Analogously, the RKC method reads for $\eta = 0$,

$$\begin{aligned} y_{k0} &= 1, \quad y_{k1} = y_{k0} + \frac{3h\lambda}{s^2-1} y_{k0}, \\ y_{ki} &= \frac{6ih\lambda}{(i+1)(s^2-1)} y_{k,i-1} + \frac{2i}{i+1} y_{k,i-1} + \frac{1-i}{i+1} y_{k,i-2}, \quad i = 2, \ldots, s-1, \\ y_{ks} &= \frac{h\lambda}{s} y_{k,s-1} + \frac{s^2-1}{3s} y_{k,s-1} - \frac{s^3-s^2-s+1}{3s^2} y_{k,s-2} + \frac{2s^2+1}{3s^2} y_0, \\ y_1 &= y_{ks}, \end{aligned} \tag{4.40}$$

and we follow the same methodology as above. Note that for RKC we have $|T_j(1+\frac{3}{s^2-1}z)| \leq 1$ for all $j = 0, \ldots, s$ and $z \in [-\beta(s,0), 0] = [-\frac{2}{3}(s^2-1), 0]$. Using the same notations we search for coefficients satisfying the following

$$y_{ki} = \sum_{j=0}^{i} \gamma_j^i T_j(1 + \frac{3}{s^2-1} z), \text{ where } \sum_{j=0}^{i} \gamma_j^i = 1 \text{ and } \gamma_j^i \geq 0 \ \forall \ i, j = 1, \ldots s-1. \tag{4.41}$$

Remark that $R_{s,0}(z) = y_{k0} = y_k = 1 = T_0(1 + \frac{3}{s^2-1}z)$, hence $\gamma_0^0 = 1$. Analogously, $R_{s,1}(z) = y_{k1} = y_{k0} + \frac{3h\lambda}{s^2-1} y_{k0} = 1 + \frac{3z}{s^2-1} = T_1(1+\frac{3}{s^2-1}z)$, and we deduce that $\gamma_0^1 = 0$, $\gamma_1^1 = 1$. Again we find a relation between these new coefficients to prove their positivity using $y_{ki} = \frac{2i}{i+1}(1 + \frac{3}{s^2-1}z) y_{k,i-1} + \frac{1-i}{i+1} y_{k,i-2}$. We get a recurrence of the same form as in the Chebyshev double adjoint method (4.37) but with different parameter, proceeding in the same we obtain exactly the same coefficients γ_j^i, which concludes the proof. □

Remark 4.3.8. *For the case of positive damping $\eta > 0$, the coefficients get very complicated and it is difficult to find a recurrence relation between them in order to prove their positivity. However, observing that all the coefficients in the recurrence relations of the internal stages of the methods are continuous functions of η, then for all s, there exists $\eta_0(s)$ such that the internal stages are stable (bounded) for all $\eta \in [0, \eta_0(s)]$. Numerical investigations suggest that Theorem 4.3.6 remains valid for all $\eta > 0$ i.e the methods remain stable with the stability functions of the internal stages bounded by 1, for all integers $s \geq 1$ for Chebyshev and $s \geq 2$ for RKC, and all $\eta > 0$. We have verified this numerically for $s \leq 200$.*

We conclude this section by the following convergence theorem for the new explicit stabilized methods for stiff optimal control problems.

Theorem 4.3.9. *The method* (4.25)-(4.26) *(resp.* (4.31)-(4.32)*) has order 1 (resp. 2) of accuracy for the optimal control problem* (4.1)*.*

Proof. The proof follows immediately from Theorem 4.2.1 with $p_{oc} = p_{ode} = 2$ for the RKC method. □

Remark 4.3.10. *The proposed explicit stabilized integrators for optimal control problems could be combined with the idea of implicit-explicit (IMEX) integrators as proposed in [30], where RKC type methods would replace the implicit part in the IMEX integrator. This idea is already proposed in [58, 10] in the context of advection-diffusion-reaction problems. In [58], the diffusion part is discretized with an RKC method which typically has a large number of internal stages, and the advection-reaction part is integrated using a 4-stage explicit*

Runge-Kutta method. In [10], the method integrates the diffusion term using ROCK2 method, the advection term using a 3-stage explicit method, and the nonlinear reaction term is solved implicitly. Such an extension is however out of the scope of the thesis.

4.4 Numerical experiments

In this Section, we illustrate numerically our theoretical findings of convergence and stability of the new fully explicit methods for stiff optimal control problems, first on a stiff three dimensional problem and second on a nonlinear advection-diffusion PDE (Burgers equation).

4.4.1 A linear quadratic stiff test problem

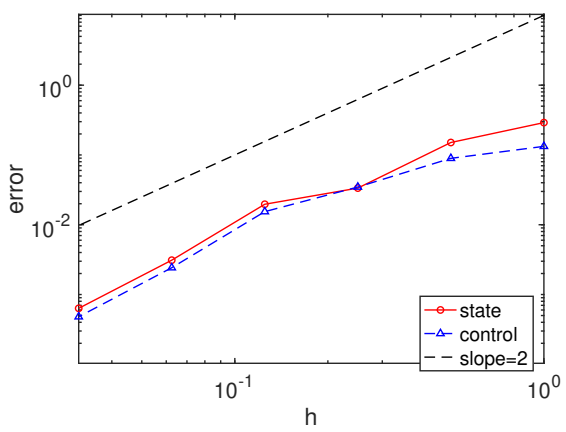
We start this section by a simple test problem taken from [27]:

$$\begin{aligned} &\min \frac{1}{2}\int_0^1 (u^2(t) + 2x^2(t))dt \text{ subject to} \\ &\dot{x}(t) = \frac{1}{2}x(t) + u(t),\ t \in [0,1],\ x(0) = 1. \end{aligned} \tag{4.42}$$

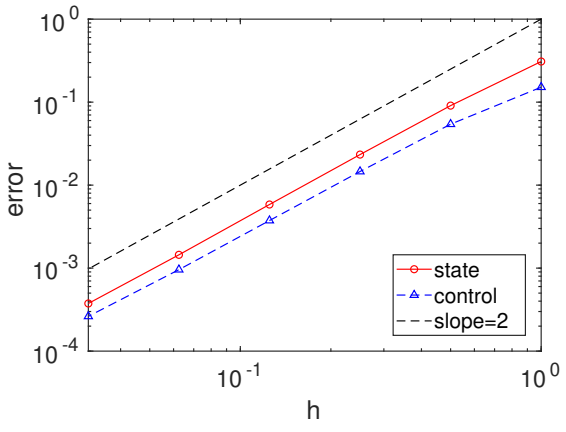
The optimal solution (u^*, x^*) is given by $u^*(t) = \frac{2(e^{3t}-e^3)}{e^{3t/2}(2+e^3)}$, $x^*(t) = \frac{2e^{3t}+e^3}{e^{3t/2}(2+e^3)}$. As studied in [30] we modify problem (4.42) into a singularly perturbed (stiff) problem to illustrate the good stability properties of our new method. For a fixed $\varepsilon > 0$, we consider the following stiff optimal control problem,

$$\begin{aligned} &\min c(1) \text{ subject to} \\ &\quad \dot{c}(t) = \frac{1}{2}(u^2(t) + x^2(t) + 4z^2(t)),\ c(0) = 0, \\ &\quad \dot{x}(t) = z(t) + u(t),\ x(0) = 1, \\ &\quad \dot{z}(t) = \frac{1}{\varepsilon}\left(\frac{1}{2}x(t) - z(t)\right),\ z(0) = \frac{1}{2}, \end{aligned} \tag{4.43}$$

Figure 4.5 shows the convergence behavior, using the new RKC method (4.31)-(4.32), of the error in infinity norm between the solutions of the stiff problem (4.43) for $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$ and different sizes of the time step $h_i = 2^{-i}, i = 0, \ldots, 5$ and the reference solution is obtained with $h = 2^{-7}$. We observe lines of slope 2 which confirms the theoretical order two of accuracy of the scheme (Theorem 4.3.9). In the stiff case ($\varepsilon = 10^{-3}$), the method uses $s = 4$ to calculate the reference solution and $s = 40, 28, 20, 14, 10, 7$ respectively for the different time steps used to illustrate the convergence, these values coincide with the theoretical values that can be obtained using (4.18). Analogously to the case of stiff ODEs, the cost of scheme (4.31)-(4.32) is $\mathcal{O}(\varepsilon^{-\frac{1}{2}})$ function evaluations of f, while using Euler method with its double adjoint would cost $\mathcal{O}(\varepsilon^{-1})$.



(a) $\varepsilon = 10^{-1}$.



(b) $\varepsilon = 10^{-3}$.

Figure 4.5: Convergence plot of RKC (4.31)-(4.32) applied to problem (4.43).

4.4.2 Optimal control of Burgers equation

To illustrate the performance of the new method, we consider the following optimal control problem of a nonlinear advection-diffusion PDE corresponding to the Burgers equation

$$\begin{aligned} \min_{u\in L^2([0,T];L^2(\Omega))} & J(u) = \frac{1}{2}\|y(T) - y^{target}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\int_0^T \|u(t)\|^2_{L^2(\Omega)} \\ & \text{subject to} \\ & \partial_t y(t,x) = \mu\Delta y(t,x) - \frac{\nu}{2}\partial_x(y^2(t,x)) + u(t,x) \quad \text{in } (0,T)\times\Omega, \\ & y(0,x) = g(x) \quad \text{in } \Omega, \\ & y(t,x) = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.44}$$

where $\mu, \nu > 0$, in dimension $d = 1$ with domain $\Omega = (0,1)$ and the final time is given by $T = 2.5$. Here the control u is a part of the source that we want to adjust in order to achieve a given final state $y^{target} : \Omega \to \mathbb{R}$.

We use a standard central finite difference space discretization for the state equation, and the trapezoid rule to discretize in space the norm $L^2(\Omega)$. We consider $M+2$ points in space $x_m = m\Delta x$, with grid mesh size $\Delta x = \frac{1}{M+1}$, and we denote by $y_m(t)$ the approximation to $y(t, x_m)$, and define the vector $Y(t) = (y_0(t), y_1(t), \ldots, y_{M+1}(t)) \in \mathbb{R}^{M+2}$. Similar notations are used for U and P. We obtain the following optimal control problem

semi-discretized in space,

$$\min_{U(t)\in\mathbb{R}^{M+2}} \quad \Psi(c(T),Y(T)) = \frac{1}{2(M+1)}{\sum_{m=0}^{M+1}}'(y_m(T)-y^{target}(x_m))^2+\alpha c(T)$$

subject to

$$\begin{aligned}
\dot{c}(t) &= \frac{1}{2(M+1)}{\sum_{m=0}^{M+1}}' u_m^2(t), \quad c(0)=0,\\
\dot{y}_m(t) &= F_m(U(t),Y(t)) := \frac{\mu}{\Delta x^2}(y_{m+1}-2y_m+y_{m-1})\\
&\qquad - \frac{\nu}{4\Delta x^2}(y_{m+1}^2-y_{m-1}^2)+u_m,\\
y_m(0) &= g(x_m),\ m=0,\dots,M+1,
\end{aligned} \tag{4.45}$$

where $m=0,\dots M+1$ and the primed sum denotes a normal sum where the first and the last term are divided by 2 and we define $y_0=y_{M+1}=0$ to take into account the homogeneous Dirichlet boundary conditions. The function $F:\mathbb{R}^{M+2}\to\mathbb{R}^{M+2}$ with components $F_m:\mathbb{R}^{M+2}\to\mathbb{R}$ is obtained from the standard central finite difference discretization of the right hand side of the state equation (4.44), and adapted to the boundary conditions. The corresponding adjoint system is

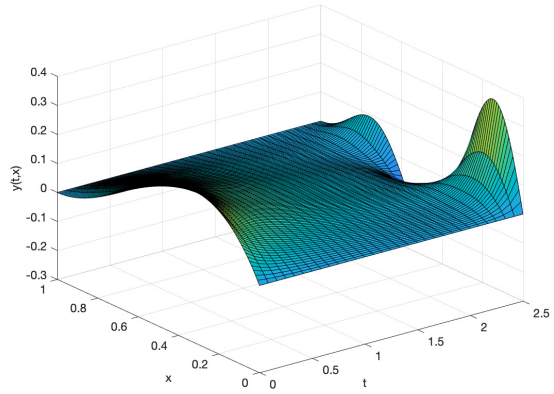
$$\begin{aligned}
\dot{p}_c(t) &= 0, \quad p_c(T)=\nabla_c\Psi=\alpha,\\
\dot{P}(t) &= -\nabla_Y F(U(t),Y(t))P,
\end{aligned} \tag{4.46}$$

where P is a vector of length $M+2$ containing the costate values p_m, $m=0,\dots,M+1$. In all our experiments we take $\mu=0.1$, $\nu=0.02$, $g(x)=\frac{3}{2}x(1-x)^2$, and $y^{target}(x)=\frac{1}{2}\sin(10x)(1-x)$.

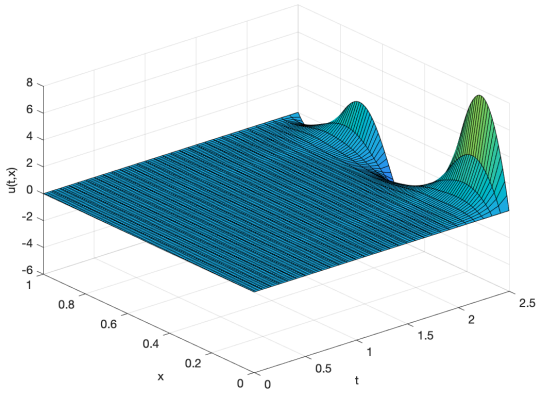
In Figure 4.6 we plot the optimal control function (Fig. 4.6b) and the corresponding state function (Fig. 4.6a) obtained using scheme (4.31)-(4.32). When we use a small value for α in the model, we allow larger control values and thus a final state very close to the target (Fig. 4.6c), otherwise the control will be more limited and then the final state will not be that close to the target (Fig. 4.6d). Note that the method required $s=24$ stages for $\Delta x=1/100$ and $\Delta t=2.5/30$, while using an Euler method with its double adjoint would require $\Delta t\le\Delta t_{max,Euler}:=\Delta x^2/2$ at most. Hence, the standard Euler method would be $\Delta t/(s\Delta t_{max,Euler})\simeq 70$ times more expensive in terms of number of function evaluations for $\Delta x=1/100$, a factor that grows arbitrarily as $\Delta x\to 0$.

In Figure 4.7, we plot the convergence curves for the state and control functions of the new RKC method (4.31)-(4.32) applied to the diffusion problem discretized in space (4.45), where the number of stages s is computed adaptively using (4.18). We recover again lines of slope two, which corroborate the order two of the method. Although our convergence analysis is valid only in finite dimensions (Theorem 4.3.9), this suggests that the convergence of order two persists in the PDE case. For comparison, we also included the results for the following standard diagonally implicit Runge-Kutta method of order two, inspired from [30, Table 5.1] in the context of stiff optimal control problems, and given by the following Butcher tableau where $\gamma=1-\sqrt{2}/2$ (making the method L-stable),

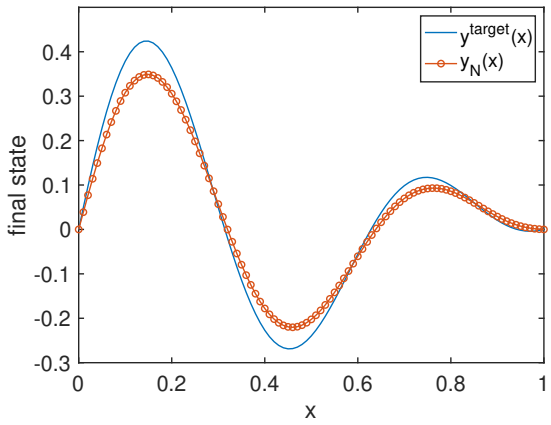
$$\begin{array}{c|cc}
\gamma & & \\
1-2\gamma & \gamma & \\
\hline
 & 1/2 & 1/2
\end{array} \tag{4.47}$$



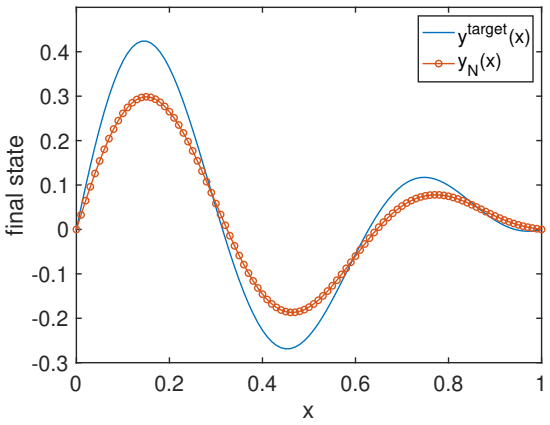
(a) State.



(b) Optimal control



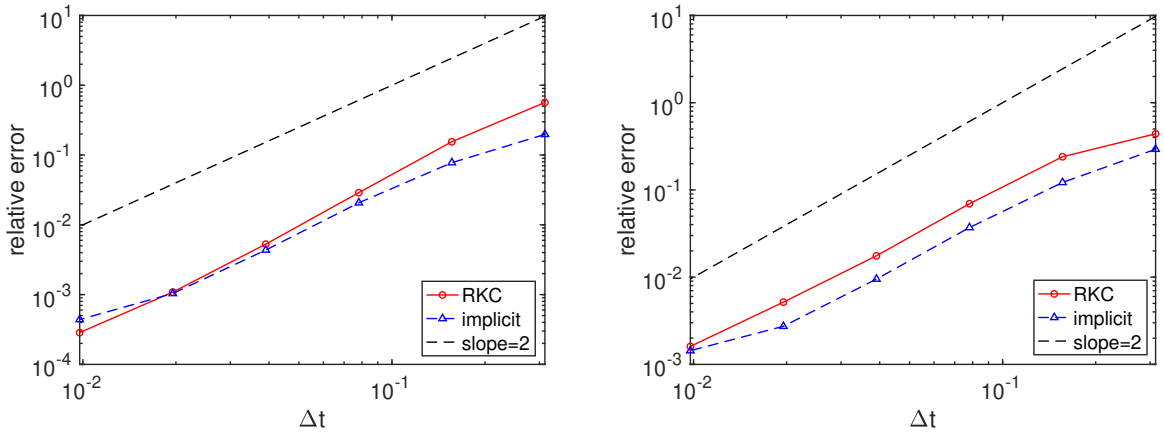
(c) Final and target states for $\alpha = 0.01$.



(d) Final and target states for $\alpha = 0.02$.

Figure 4.6: State, final state, and control, of problem (4.44). Figures (a), (b), and (c) are obtained using $\Delta x = 1/100,\ \Delta t = T/30$, and $s = 24$ stages, and $\alpha = 0.01$. Figure (d) uses the same $\Delta x,\ g$ and y^{target} but $\alpha = 0.02$.

Although for a fixed timestep, the second order implicit method (4.47) appears about two times more accurate than the RKC method (4.31)-(4.32) for the control and almost of the same accuracy for the state, we emphasize that these convergence plots do not take into account the extra cost of the implicitness of method (4.47). Indeed, the cost and difficulty of the implementation of the implicit methods (nonlinear iterations, preconditioners, etc.) would typically deteriorate in larger dimensions and for a nonlinear diffusion operator, as it is already the case for initial value PDEs [4], while as an explicit stabilized scheme, the RKC method (4.31)-(4.32) can be conveniently implemented in the spirit of the simplest explicit Euler method.



(a) Error in the state.

(b) Error in the control.

Figure 4.7: Convergence plot of the RKC method (4.31)-(4.32) and the implicit method (4.47) applied to problem (4.45) for many time steps $\Delta t_i = T/2^i,\ i = 3, \ldots, 8$, $\Delta x = 1/100$, and $\alpha = 0.02$, The reference solution is obtained using $\Delta t = T/2^{12},\ s = 3$.

Chapter 5

Conclusion and outlook

We have designed new and explicit stabilized schemes for different types of problems, and we think that we now have enough material to look at many new interesting ideas related to our work. Before concluding we would like to present one idea that we looked at, and we think it is very interesting and promising.

5.1 Towards explicit implementation of implicit methods using optimization techniques and explicit stabilized integrators

We aim to implement excellent well established higher order implicit Runge-Kutta methods explicitly by combining some optimization techniques with an explicit stabilized method of order 1. We will explain the methodology in the rest of the current section.

In the recent paper [24], the authors proposed an explicit stabilized version of gradient descent (GD) to replace the standard GD method (5.3) for stiff optimization problems in large dimension to avoid step size restriction. Here we mean by stiff optimization problems the case where the gradient of the objective function is stiff. In order to solve the problem

$$\min_{x\in\ \mathbb{R}^d} f(x), \tag{5.1}$$

where $d \in \mathbb{N}\backslash\{0\}$ and $f : \mathbb{R}^d \to \mathbb{R}$ a continuously differentiable function, with L-Lipschitz gradient, and such that the real parts of the eigenvalues of its Hessian matrix $\mathcal{H}f(x)$ are strictly positive for all $x \in \mathbb{R}^d$, the authors of [24] consider its gradient flow

$$\dot{x} = -\nabla f(x), \qquad x(0) = x_0 \in \mathbb{R}^d, \tag{5.2}$$

where x_0 is an initialization. Instead of applying an explicit Euler method to (5.2), which is equivalent to GD, they apply an explicit stabilized method. Indeed, if we discretize (5.2) using an explicit Euler method with step size h, we get the famous gradient descent (GD) method

$$x_{n+1} = x_n - h\nabla f(x_n), \qquad n = 0, 1, 2, ... \tag{5.3}$$

where x_n is the numerical approximation of $x(t_n)$. For stiff gradients, this method faces severe restriction on h in order to be stable. Another approach is to discretize (5.2) using an implicit Euler method with a step size h

$$x_{n+1} = x_n - h\nabla f(x_{n+1}), \qquad n = 0, 1, 2, ... \tag{5.4}$$

It is easy to verify that x_{n+1} in (5.4) is the unique minimizer of the problem

$$\min_{x\in\ \mathbb{R}^d} hf(x) + \frac{1}{2}\|x - x_n\|_2^2. \tag{5.5}$$

It is reasonable to think of applying a GD method with step size τ to the gradient flow of (5.5) to avoid implicitness, and reduce the step size restriction, but this will not reduce enough the restriction on τ especially if h is large. This pushes us to think of applying an explicit stabilized method to (5.5) with a step size $\tau = \mathcal{O}(1)$ and a suitable number of stages s, and iterate until convergence. The obtained algorithm is in fact explicit. We look at our stiff ODE as a gradient flow of some optimization problem, for which, the objective function needs not to be known. We aim to apply this idea to a powerful implicit method of higher order and perfect stability properties such as RADAU methods. If it works with the fifth order RADAU IIA for example, we obtain an explicit implementation of an excellent stiffly accurate method of order 5.

In a general framework, let d be a positive integer and consider the ODE

$$\dot{y} = f(y), \qquad y(0) = y_0, \tag{5.6}$$

where $y_0 \in \mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R}^d$ is an L-Lipschitz vector field. Suppose that the eigenvalues of the Jacobian matrix of f have strictly negative real parts for all $y \in \mathbb{R}^d$. Consider a stiffly accurate implicit method (for example RADAU IIA of order 5) with m stages, and with matrix of coefficients $\mathcal{A}$. After some calculations we arrive at the minimization problem

$$\min_{\mathcal{Y}\in\mathbb{R}^{d\times m}} hF(\mathcal{Y}) + \frac{1}{2}\|\mathcal{Y} - \mathcal{Y}_0\|_2^2, \tag{5.7}$$

where $\mathcal{Y}$ is a vector containing all the internal stages of the considered method. The function F needs not to be known and it is such that $-\nabla F(\mathcal{Y}) = \mathcal{A} \otimes I_d \begin{pmatrix} f(Y_1) \\ f(Y_2) \\ \vdots \\ f(Y_m) \end{pmatrix}$. The corresponding gradient flow is

$$\dot{\mathcal{Y}} = h\mathcal{A} \otimes I_d \begin{pmatrix} f(Y_1) \\ f(Y_2) \\ \vdots \\ f(Y_m) \end{pmatrix} - \mathcal{Y} + \mathcal{Y}_0, \qquad \mathcal{Y}(0) = \mathcal{Y}_0. \tag{5.8}$$

The idea now is to apply an explicit stabilized integrator to (5.8) with a time step $\tau = \mathcal{O}(1)$, and iterate.

Consider the test problem $\dot{y} = -\lambda y$, $y(0) = y_0$, where $\Re\lambda < 0$, which leads to

$$\dot{\mathcal{Y}} = -h\lambda\mathcal{A}\mathcal{Y} - \mathcal{Y} + \mathcal{Y}_0, \qquad \mathcal{Y}(0) = \mathcal{Y}_0. \tag{5.9}$$

In the case where the eigenvalues of $\mathcal{A}$ are real positive, we can apply the standard Chebyshev method (2.14). For some implicit methods (RADAU IIA for example) the eigenvalues of $\mathcal{A}$ are complex, which makes the standard Chebyshev method useless.

By diagonalizing A, we compute A^r and A^i such that $A = A^r + iA^i$ and the eigenvalues of A^r are the real parts of the eigenvalues of A, and the eigenvalues of A^i are the imaginary parts of the eigenvalues of A.

The above splitting of the matrix A requires an explicit stabilized method that is stable enough in the imaginary direction, and this can be done by stabilizing the part of the vector field with almost pure imaginary eigenvalues (advection-like) using a new polynomial. The overall stability function will look like

$$Q_s(h\tau\lambda A^r) + B_s(h\tau\lambda A^r)(h\tau\lambda i A^i). \tag{5.10}$$

For optimization purposes, we ask that the above partition preserves the steady state of the system in both linear and nonlinear vector fields cases. We consider

$$Q_s(p) = \frac{T_s(\omega_0 + \omega_1' p)}{2T_s(\omega_0)} + \frac{1}{2}$$

and $B_s(p) = (Q_s(p) - 1)/p$, where $\omega_0 = 1 + \eta/s^2$, $\omega_1' = \frac{\mathbf{2T_s(\omega_0)}}{\mathbf{T_s'(\omega_0)}}$ and $\eta = \mathbf{0.1}$, and the overall stability function

$$R_s(p,q) = \frac{T_s(\omega_0 + \omega_1' p)}{2T_s(\omega_0)} + \frac{1}{2} + \frac{1}{p}\left(\frac{T_s(\omega_0 + \omega_1' p)}{2T_s(\omega_0)} - \frac{1}{2}\right) iq, \tag{5.11}$$

where $p = h\tau\lambda A^r$ and $q = h\tau\lambda A^i$. To motivate the above choice, consider the test equation $\dot{y} = Uy + Vy$, $y(0) = y_0$, such that $Uy_0 + Vy_0 = 0$, we want $y_1 = y_0$, i.e.

$$\begin{aligned} Q_s(hU)y_0 + B_s(hU)hVy_0 &= y_0, \\ -B_s(hU)hVy_0 &= (Q_s(hU) - I)y_0, \\ B(hU)hUy_0 &= (Q_s(hU) - I)y_0, \quad \text{since } -Vy_0 = Uy_0, \\ B_s(p) &= (Q_s(p) - 1)/p, \end{aligned}$$

where $p = hU$.

For problems of the form

$$\dot{y} = f(y) + g(y), \qquad y(0) = y_0, \tag{5.12}$$

where g contains the imaginary parts of the eigenvalues appearing in the case of RADAU IIA for example, we consider the following integrator

$$\begin{aligned} K_0 &= y_0, \qquad K_1 = y_0 + \frac{\omega_1'}{2\omega_0} h(f(y_0) + g(y_0)), \\ K_j &= \mu_j h(f(K_{j-1}) - \frac{1}{2}f(y_0) + \frac{1}{2}g(y_0)) + \nu_j K_{j-1} + (1 - \nu_j)K_{j-2}, \quad j = 2, \ldots, s, \\ y_1 &= K_s, \end{aligned} \tag{5.13}$$

where,

$$\mu_j = \frac{2\omega_1' T_{j-1}(\omega_0)}{T_j(\omega_0)}, \qquad \nu_j = \frac{2\omega_0 T_{j-1}(\omega_0)}{T_j(\omega_0)}.$$

Lemma 5.1.1. *Applied to the linear test problem $\dot{y} = \lambda y + i\sigma y$ the scheme* (5.13) *yields the following stability function*

$$R_s(p,q) = \frac{T_s(\omega_0 + \omega_1' p)}{2T_s(\omega_0)} + \frac{1}{2} + \frac{1}{p}\left(\frac{T_s(\omega_0 + \omega_1' p)}{2T_s(\omega_0)} - \frac{1}{2}\right) iq, \tag{5.14}$$

where $p = h\lambda$ and $q = h\sigma$.

Theorem 5.1.2. *The method* (5.13) *is of order 1 of accuracy for ODEs of the form* (5.12)*, with f and g are Lipschitz continuous.*

Proposition 5.1.3 (steady state conservation)**.** *Let $\dot{y} = f(y) + g(y)$ and let $y(0) = y_0$ such that $f(y_0) + g(y_0) = 0$. Then, $y_1 = y_0$, with y_1 defined as in* (5.13)*.*

Proofs of Lemma 5.1.1 and Theorem 5.1.2 are similar to the proofs of Lemma 3.3.1 and Theorem 3.4.1 respectively. The proof of Proposition 5.1.3 is straightforward by induction.

5.2 Conclusion

In this thesis, we have constructed novel and efficient explicit stabilized methods for different types of stiff problems by combining techniques from stiff integration and geometric numerical integration. Part of the work is presented in two research articles [5, 14].

A remarkable feature of the explicit stabilized schemes proposed in Chapter 3 (stiff stochastic problems) and Chapter 4 (stiff optimal control problems) is that their extended size of stability domain is not only optimally large in many situations, but it is also based on rigorous estimates and it does not rely on empirical estimates as proposed in the past literature. This asset makes the new schemes promising in the context of variance reduction techniques (for instance combined with the multilevel Monte-Carlo method for stiff stochastic problems [6]) or parallel computing (for instance to be combined with the parareal algorithm for dissipative problems, in particular for optimal control [43]), where the ability of the integrators to apply large time steps with reliable stability is a key ingredient.

The SK-ROCK method with optimal stability domain presented in Chapter 3 is used in the article [47] to construct a highly efficient proximal Markov chain Monte Carlo methodology to perform Bayesian computation in imaging problems. Instead of the conventional Euler-Maruyama approximation that underpins existing proximal Monte Carlo methods, the authors use SK-ROCK method to significantly accelerate the convergence speed, similarly to accelerated gradient optimization methods. A multirate version of SK-ROCK is also used in the thesis [48] to solve stiff multirate stochastic differential equations.

Recently, the idea of stabilized Chebyshev methods was used in [21] in the context of ODEs arising from the discretization of wave equations. Based on Chebyshev polynomials, and in the spirit of explicit stabilized methods from the literature, the authors construct a stabilized version of leapfrog method for linear and semilinear second-order differential equations.

To conclude, we see that explicit stabilized methods are very useful in many contexts and not only for stiff dissipative problems, and we believe that they could still be used for many other classes of stiff problems and hence would give rise to many new interesting research ideas.

Bibliography

[1] A. Abdulle. On roots and error constants of optimal stability polynomials. *BIT Numerical Mathematics*, 40(1):177–182, 2000.

[2] A. Abdulle. *Chebyshev methods based on orthogonal polynomials*. PhD Thesis, University of Geneva, Department of Mathematics. University of Geneva, 2001.

[3] A. Abdulle. Fourth order Chebyshev methods with recurrence relation. *SIAM J. Sci. Comput.*, 23(6):2041–2054, 2002.

[4] A. Abdulle. *Explicit Stabilized Runge–Kutta Methods*, pages 460–468. Encyclopedia of Applied and Computational Mathematics, Springer Berlin Heidelberg, 2015.

[5] A. Abdulle, I. Almuslimani, and G. Vilmart. Optimal explicit stabilized integrator of weak order 1 for stiff and ergodic stochastic differential equations. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):937–964, 2018.

[6] A. Abdulle and A. Blumenthal. Stabilized multilevel Monte Carlo method for stiff stochastic differential equations. *J. Comput. Phys.*, 251:445–460, 2013.

[7] A. Abdulle and S. Cirilli. S-ROCK: Chebyshev methods for stiff stochastic differential equations. *SIAM J. Sci. Comput.*, 30(2):997–1014, 2008.

[8] A. Abdulle and T. Li. S-ROCK methods for stiff Ito SDEs. *Commun. Math. Sci.*, 6(4):845–868, 2008.

[9] A. Abdulle and A. Medovikov. Second order chebyshev methods based on orthogonal polynomials. *Numer. Math.*, 90(1):1–18, 2001.

[10] A. Abdulle and G. Vilmart. PIROCK: a swiss-knife partitioned implicit-explicit orthogonal Runge-Kutta Chebyshev integrator for stiff diffusion-advection-reaction problems with or without noise. *J. Comput. Phys.*, 242:869–888, 2013.

[11] A. Abdulle, G. Vilmart, and K. C. Zygalakis. Mean-square A-stable diagonally drift-implicit integrators of weak second order for stiff Itô stochastic differential equations. *BIT*, 53(4):827–840, 2013.

[12] A. Abdulle, G. Vilmart, and K. C. Zygalakis. Weak second order explicit stabilized methods for stiff stochastic differential equations. *SIAM J. Sci. Comput.*, 35(4):A1792–A1814, 2013.

[13] G. Albi, M. Herty, and L. Pareschi. Linear multistep methods for optimal control problems and applications to hyperbolic relaxation systems. *Appl. Math. Comput.*, 354:460–477, 2019.

[14] I. Almuslimani and G. Vilmart. Explicit stabilized integrators for stiff optimal control problems. *Under revision for SIAM J. Sci. Comput.*, arXiv: 1910.10584, 2020.

[15] M. Bakker. Analytical aspects of a minimax problem. 1971. Technical Note TN 62 (in Dutch), Mathematical centre, Amsterdam.

[16] J. F. Bonnans and J. Laurent-Varin. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. *Numer. Math.*, 103(1):1–10, 2006.

[17] C.-E. Bréhier and G. Vilmart. High Order Integrator for Sampling the Invariant Distribution of a Class of Parabolic Stochastic PDEs with Additive Space-Time Noise. *SIAM J. Sci. Comput.*, 38(4):A2283–A2306, 2016.

[18] E. Buckwar and C. Kelly. Towards a systematic linear stability analysis of numerical methods for systems of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(1):298–321, 2010.

[19] K. Burrage, P. Burrage, and T. Tian. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2041):373–402, 2004.

[20] J. C. Butcher. The effective order of Runge-Kutta methods. In J. L. Morris, editor, *Proceedings of Conference on the Numerical Solution of Differential Equations*, volume 109 of *Lecture Notes in Math.*, pages 133–139, 1969.

[21] C. Carle, M. Hochbruck, and A. Sturm. On Leapfrog-Chebyshev Schemes. *SIAM J. Numer. Anal.*, 58(4):2404–2433, 2020.

[22] Y. Chong and J. B. Walsh. The roughness and smoothness of numerical solutions to the stochastic heat equation. *Potential Anal.*, 37(4):303–332, 2012.

[23] A. M. Davie and J. G. Gaines. Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations. *MATH. COMP*, 70:121–134, 2000.

[24] A. Eftekhari, B. Vandereycken, G. Vilmart, and K. C. Zygalakis. Explicit stabilised gradient descent for faster strongly convex optimisation. *To appear in BIT Numerical Mathematics (2020).*

[25] T. Gard. *Introduction to stochastic differential equations.* Marcel Dekker, New York, 1988.

[26] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[27] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numer. Math.*, 87(2):247–282, 2000.

[28] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006. Structure-preserving algorithms for ordinary differential equations.

[29] E. Hairer and G. Wanner. *Solving ordinary differential equations II. Stiff and differential-algebraic problems*. Springer-Verlag, Berlin and Heidelberg, 1996.

[30] M. Herty, L. Pareschi, and S. Steffensen. Implicit-explicit Runge-Kutta schemes for numerical discretization of optimal control problems. *SIAM J. Numer. Anal.*, 51(4):1875–1899, 2013.

[31] M. Herty and V. Schleper. Time discretizations for numerical optimisation of hyperbolic problems. *Appl. Math. Comput.*, 218(1):183–194, 2011.

[32] D. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.

[33] D. J. Higham. Mean-square and asymptotic stability of the stochastic theta method. *SIAM J. Numer. Anal.*, 38(3):753–769, 2000.

[34] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003.

[35] C. Y. Kaya. Inexact restoration for Runge-Kutta discretization of optimal control problems. *SIAM J. Numer. Anal.*, 48(4):1492–1517, 2010.

[36] P. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer-Verlag, Berlin and New York, 1992.

[37] J. Lang and J. G. Verwer. W-methods in optimal control. *Numer. Math.*, 124(2):337–360, 2013.

[38] A. Laurent and G. Vilmart. Exotic aromatic B-series for the study of long time integrators for a class of ergodic SDEs. *Math. Comp.*, 89(321):169–202, 2020.

[39] B. Leimkuhler and C. Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math. Res. Express.*, 2013(1):34–56, 2013.

[40] B. Leimkuhler, C. Matthews, and M. V. Tretyakov. On the long-time integration of stochastic gradient systems. *Proc. R. Soc. A*, 470(2170), 2014.

[41] Q. Li, L. Chen, C. Tai, and W. E. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18(165):1–29, 2018.

[42] X. Liu and J. Frank. Symplectic runge–kutta discretization of a regularized forward–backward sweep iteration for optimal control problems. *Journal of Computational and Applied Mathematics*, 383:113133, 2021.

[43] Y. Maday, M.-K. Riahi, and J. Salomon. Parareal in time intermediate targets methods for optimal control problems. In *Control and optimization with PDE constraints*, volume 164 of *Internat. Ser. Numer. Math.*, pages 79–92. Birkhäuser/Springer Basel AG, Basel, 2013.

[44] G. N. Mil′shteĭn. A theorem on the order of convergence of mean-square approximations of solutions of systems of stochastic differential equations. *Teor. Veroyatnost. i Primenen.*, 32(4):809–811, 1987.

[45] G. Milstein. Weak approximation of solutions of systems of stochastic differential equations. *Theory Probab. Appl.*, 30(4):750–766, 1986.

[46] G. Milstein and M. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computing. Springer-Verlag, Berlin and New York, 2004.

[47] M. Pereyra, L. Vargas Mieles, and K. C. Zygalakis. Accelerating proximal Markov chain Monte Carlo by using an explicit stabilized method. *SIAM J. Imaging Sci.*, 13(2):905–935, 2020.

[48] G. Rosilho De Souza. Numerical methods for deterministic and stochastic differential equations with multiple scales and high contrasts. page 216, 2020.

[49] Y. Saito and T. Mitsui. Stability analysis of numerical schemes for stochastic differential equations. *SIAM J. Numer. Anal.*, 33:2254–2267, 1996.

[50] J. M. Sanz-Serna. Symplectic Runge-Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more. *SIAM Rev.*, 58(1):3–33, 2016.

[51] B. Sommeijer, L. Shampine, and J. Verwer. RKC: an explicit solver for parabolic PDEs. *J. Comput. Appl. Math.*, 88:316–326, 1998.

[52] B. P. Sommeijer and J. G. Verwer. *A performance evaluation of a class of Runge-Kutta-Chebyshev methods for solving semidiscrete parabolic differential equations*. Afdeling Numerieke Wiskunde [Department of Numerical Mathematics], 91. Mathematisch Centrum, Amsterdam, 1980.

[53] A. Tocino. Mean-square stability of second-order Runge-Kutta methods for stochastic differential equations. *J. Comput. Appl. Math.*, 175(2):355–367, 2005.

[54] 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(10):479–485, 1980.

[55] J. Verwer. Explicit Runge-Kutta methods for parabolic partial differential equations. *Special issue of Appl. Num. Math.*, 22:359–379, 1996.

[56] G. Vilmart. Postprocessed integrators for the high order integration of ergodic SDEs. *SIAM J. Sci. Comput.*, 37(1):A201–A220, 2015.

[57] A. Walther. Automatic differentiation of explicit Runge-Kutta methods for optimal control. *Comput. Optim. Appl.*, 36(1):83–108, 2007.

[58] C. J. Zbinden. Partitioned Runge-Kutta-Chebyshev methods for diffusion-advection-reaction problems. *SIAM J. Sci. Comput.*, 33(4):1707–1725, 2011.

List of Figures