Karlsruhe Institute of Technology

On leapfrog-Chebyshev schemes for second-order differential equations

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie
genehmigte

DISSERTATION

von

Constantin Carle

Tag der mündlichen Prüfung: 15. Dezember 2021

Referentin: Prof. Dr. Marlis Hochbruck
Korreferent: Prof. Dr. Tobias Jahnke

Acknowledgement

I gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 258734477 – CRC 1173.

In German.

An dieser Stelle möchte ich mich bei allen bedanken, die mich, sei es in wissenschaftlicher, beruflicher oder persönlicher Hinsicht, im Laufe der Promotion unterstützt und auf diesen Weg begleitet haben.

Mein allergrößter Dank geht hierbei an meine Betreuerin Prof. Dr. Marlis Hochbruck. Ohne sie wäre diese Arbeit nicht möglich gewesen. Bereits während des Studiums ermöglichte sie es mir, als wissenschaftliche Hilfskraft in ihrer Arbeitsgruppe mitzuarbeiten und sie war es auch, die mich darin bekräftigt hat, zu promovieren. Während der letzten vier Jahren konnte ich mich immer auf ihre Unterstützung und Hilfe verlassen, insbesondere auch bei der Vereinbarkeit von Familie und Beruf. Für ihre wertvollen Korrekturvorschläge, die diese Arbeit wesentlich verbessert haben, bin ich besonders dankbar.

Einen großen Dank gebührt auch meinem Zweitbetreuer Prof. Dr. Tobias Jahnke, der sich ebenfalls vorab die Zeit genommen hat, diese Arbeit mit unzähligen hilfreichen Anmerkungen aufzuwerten.

Der gesamten Arbeitsgruppe Numerik, in all ihrer wechselnden Zusammensetzung, kann ich gar nicht genug danken für die vielen tollen, oft auch sehr unterhaltsamen Gespräche, Mittagsrunden und gemeinsame Abende. Neben dem wissenschaftlichem Alltag kam das soziale Miteinander nie zu kurz. Bernhard und Benjamin bin ich insbesondere dankbar für die Mühen und die Zeit, die sie investiert haben, um sich durch erste Versionen dieser Arbeit zu lesen. Ein besonderer Dank geht auch an Laurette, Mathias und Christian, dir mir bei administrativen beziehungsweise technischen Fragen immer weiterhelfen konnten.

Weiter möchte ich mich bei meinen Eltern bedanken, die mich darin bekräftigt haben, meinen eigenen Weg zu gehen. Auf eure Unterstützung kann ich mich immer verlassen.

Mein letzter Dank gebührt meiner Familie. Ihr wart in den letzten Jahren der beste Rückhalt, den man sich wünschen kann. Obwohl ihr in den Wochen vor der Abgabe etwas zurückstecken musstet und weniger von mir hattet, als ich mir selbst gewünscht hätte, wart ihr immer für mich da und habt für das richtige Maß an Abwechslung und Ablenkung gesorgt. Anja, Theresa, diese Arbeit ist für euch.

Contents

Chapter 1

Introduction

Many phenomena occurring in physical and engineering sciences can be modeled by ordinary or partial differential equations. For instance, the traveling of sound waves through media can be described by a partial differential equation, the acoustic wave equation. The derivation of such equations from a given model and their study is a task which occupies engineers, physicists, and mathematicians for centuries. However, until today, exact solutions of such equations are only known for very few, often simple and non-realistic, cases. Nowadays, the tremendous increase of computing power enables one to simulate realistic problems with sufficient accuracy in a reasonable time. Nevertheless, to understand many phenomena occurring in nature even better, the size of such problems has to be increased even more and, hence, efficient numerical methods for simulating such problems are still inevitable.

In this thesis we focus on the efficient time integration of a specific class of ordinary differential equations, namely *semilinear second-order differential equations* in $\mathbb{R}^m$ of the form

$$\ddot{\mathbf{q}}(t) = -\mathbf{L}\mathbf{q}(t) + \mathbf{g}(t, \mathbf{q}(t)), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0, \tag{1.1}$$

with initial values $\mathbf{q}_0, \dot{\mathbf{q}}_0$. For the matrix $\mathbf{L} \in \mathbb{R}^{m\times m}$ we assume that it is symmetric and positive semidefinite, and $\mathbf{g}$ is a sufficiently smooth function. A particular class of problems leading to such an equation (possibly after a transformation of variables) are spatially discretized wave-type equations such as the acoustic wave equation. Other fields of application in which such equations arises, are, for instance, Hamiltonian equations of motions occurring in astronomy or in molecular dynamics.

The most popular method for approximating the solution of (1.1) is the famous leapfrog scheme, which is also known – depending on its precise formulation and the field of application – under various other names, e.g., Störmer scheme or Verlet scheme. Its widespread use relies on its favorable properties: explicit, second-order, easy to implement. In addition, it is very efficient, because it requires only one evaluation of $\mathbf{g}$ and one matrix-vector multiplication with $\mathbf{L}$ per time step. Moreover, it features nice geometric properties such as symplecticity and symmetry (time-reversibility); cf. [HLW03] or the monograph [HLW06] for these properties and many more details.

The price we have to pay for the explicitness of the leapfrog scheme is a step-size restriction required for stability. To be more precise, for $\mathbf{g} \equiv 0$ we need step sizes satisfying at least $\tau^2\|\mathbf{L}\| \leq 4$ to obtain stable approximations. Although this step-size restriction can be severe,

the leapfrog scheme is in general – without further knowledge of the problem – probably the most efficient scheme for equations of the form (1.1).

However, there are situations in which the step-size restriction causes a significant loss of efficiency. In this thesis we pay special attention to two such situations.

1. The linear part $\mathbf{Lq}$ induces the main stiffness of (1.1) but is cheap to compute compared to the expensive evaluation of $\mathbf{g}$, which is a "nice" function with small Lipschitz constant. Here, the linear part causes a severe step-size restriction in the leapfrog scheme which in turn requires a large number of expensive evaluations of $\mathbf{g}$.
2. The second situation handles cases in which only a few components of the solution of (1.1) are responsible of the severe stiffness. More precisely, we are interested in cases, where this stiffness is induced by a small principle submatrix of $\mathbf{L}$. Such situations occur, for instance, for spatially discretized acoustic wave equations if the triangulation of the domain contains only a few very small elements, or if the material parameter is only in a small part of the domain large compared to the remaining part.

For the first situation so-called *multiple time-stepping* schemes, also known as *impulse methods*, were formulated [GHW+91, TBM92] (for the more general case that $-\mathbf{Lq}$ is replaced by a function $\mathbf{f}(\mathbf{q})$), which are in general explicit schemes. They share the property that the stiff part is evaluated with a smaller step size and thus more often than the non-stiff part. In [BS93] it is shown that these scheme are prone to numerical resonances and instabilities at certain step sizes; see also [GSS99]. In [GSS99] they further suggest a modification of the original method, the so-called *mollified impulse method*; see also [HLW06, Section XIII.1.4]. These schemes are very popular in molecular dynamics and many variants have been proposed. However, most of them are designed and motivated for particular applications such as multiple time scales in molecular dynamics; cf. for example, the monographs [HLW06, Chapters VIII.4, XIII.1], [LM15, Chapter 4], and [LR04, Chapter 10] and the references therein. We emphasize that they are structurally very similar to splitting and composition schemes and some of them can be even interpreted as splitting/composition schemes; see, e.g., [BS93] or [HLW06, Algorithm VIII.4.1]. Up to our best knowledge no rigorous stability and error analysis of these schemes exists in the case of (1.1).

Another possibility to numerically integrate such equations are trigonometric integrators, which are especially designed for the class of equations (1.1). These integrators rely on the variation-of-constants formula, where the stiff linear part is integrated exactly and only the integral term containing the semilinearity term is approximated; see for instance [HLW06, Chapter XIII] and references therein. For these integrators rigorous stability and error analyses exist; see again [HLW06, Chapter XIII]. More recent results for trigonometric integrators can be found, for instance, in [BGG+18] and, in the context of partial differential equations, in [Gau15, BDH21]. A main advantage of these integrators is that the rigorous stability and error analysis provided so far in the literature also covers the case of highly oscillatory solutions (or of low regularity in the context of partial differential equations). However, for an efficient implementation of such schemes one has to (approximately) compute products of matrix functions with vectors which is in general an elaborate task and where usually special structures of the underlying problem have to be exploited. In contrast to this, we (mainly) focus on explicit schemes which are much more easy to implement.

For the second situation *multirate* schemes were formulated, first proposed in [Ric60] (who called it *split Runge–Kutta* schemes). In these schemes, the stiff part is numerically integrated with a smaller step size compared to the non-stiff part or even with a completely different scheme

including implicit ones. For first-order differential equations there is a vast amount of papers dealing with such a situation; see, e.g., [Hof76, And79, GW84, SA89, GR93, EL97, Kvæ00, GKR01, SHV07, SM10] or, more recently, [CS13, GS16, SRS19, RLS+21]. These schemes rely on different time integrators, for instance, explicit as well as implicit Runge–Kutta or multistep methods, and the different parts of the equations are often coupled via inter- or extrapolation. A rigorous stability analysis in the stiff case is missing for almost all proposed schemes. Of special interest for our work are the multirate schemes constructed recently in [AGS21] which are based on Runge–Kutta–Chebyshev schemes [HS80, Ver82, VHS90], since Runge–Kutta–Chebyshev methods can be seen as the first-order counterpart (for equations of parabolic type) to the schemes we consider here. However, we are not aware of a multirate scheme for ordinary differential equations exploiting directly the second-order structure of the underlying equation.

Aims and main results

The main goals of this thesis are the construction and analysis of two classes of (mainly explicit) two-step schemes of second-order for second-order differential equations of the form (1.1). They allow for a more efficient implementation compared to the leapfrog scheme in the specific cases described above. We do not aim for higher order, since in applications second-order convergence is often sufficient to obtain reasonable results and in many cases the step-size restriction of the leapfrog scheme is the limiting factor. The initial motivation of constructing and analyzing these schemes goes back to the aim of a deeper understanding of the local time-stepping schemes proposed in [DG09], for which we discovered that there is a close relation to the methods considered in this thesis.

The first class of schemes, the *leapfrog-Chebyshev* (LFC) *schemes*, utilizes the fact that an evaluation of the semilinearity or inhomogeneity $\mathbf{g}$ is at least as costly as a matrix-vector multiplication with $\mathbf{L}$, whereas the stiffness is induced only by the matrix $\mathbf{L}$. In order to analyze these schemes, we introduce a rather general class of two-step methods, which comprises not only explicit and implicit time integration schemes but also trigonometric integrators. In comparison to [CHS20] we consider a slightly larger and modified general class of schemes which allows for better stability bounds. For this general class of schemes we provide a comprehensive stability and error analysis. The key tool for the analysis is a representation formula of the numerical approximations which we prove via two different techniques. In [CHS20] this is shown via *generating functions*, the other new proof relies on an one-step formulation of our general class of schemes. This formula additionally enables us to analyze the sensitive influence of the starting value to the overall stability which was not considered so far in the literature. Moreover, for the LFC schemes we are able to state explicit values of all constants occurring in the stability and error bounds. A summary of these results or variants thereof are already published in [CHS20].

For a modification of θ-schemes which also belongs to the general class, we further provide improved stability bounds compared to the general ones. In the linear case, the original θ-schemes, see, e.g., [Kar12, Section 3.2] for a derivation, are analyzed, for instance, in [Kar11]. Just recently, an error analysis of modified θ-schemes for linear problems was given in [HHW21]. In [HL21] a so-called *IMEX* (implicit-explicit) scheme for wave equations with damping and forcing terms is considered and rigorous error bounds are provided. Without damping and for $\theta = \frac{1}{4}$ both schemes are (almost) identical. Our improved stability results coincide with the one in [HHW21, HL21] if one restricts to the same situation.

A related class of two-step schemes is analyzed in [CI17] for the homogeneous case $\mathbf{g} \equiv 0$. Although the authors prove an error bound (via an extension of standard energy techniques) they leave many questions open, for instance, they do not specify a starting value for their general class.

The construction of the LFC schemes is motivated by [GJ08, JR10]. For $\mathbf{g} \equiv 0$ they proposed a special case (the unstabilized variant) of the LFC scheme, again without specifying a starting value. Moreover, the same unstabilized LFC scheme occurs as a special case of the local time-stepping scheme in [DG09]. Based on the stability analysis of our general class of two-step methods we could show that this unstabilized variant is prone to instabilities. To overcome this instabilities we introduce a stabilization parameter in the LFC schemes inspired by damped/stabilized Runge–Kutta–Chebyshev (RKC) methods; see, e.g., [HS80, Ver82, VHS90] or [HV03, Chapter V].

The multirate leapfrog-type two-step schemes are designed for situations where the main stiffness of (1.1) is induced only by a (very) small part of the matrix $\mathbf{L}$. To overcome the crippling effect of this small part, we use a clever combination of the leapfrog scheme and the general class of two-step schemes presented before with particular interest on the LFC schemes. Based on the results for the general class, especially the representation formula, we are able to provide error bounds preserving the second-order convergence of the leapfrog scheme. Moreover, for special cases of the general class of two-step schemes, for instance, the stabilized LFC schemes, the step-size restriction of the multirate scheme is independent of the stiff part and approximately as large as for the leapfrog scheme applied to the non-stiff part. A preprint containing most of these results is submitted for publication [CH21].

The construction of these multirate schemes is based on the local time-stepping schemes in [DG09] which represents a special case of our multirate scheme. More precisely, if one equips the multirate scheme with the unstabilized LFC scheme, the multirate scheme and the local time-stepping scheme in [DG09] coincide for $\mathbf{g} \equiv 0$. Moreover, our multirate scheme is also influenced by the locally implicit schemes proposed in [Ver11] and analyzed in [HS16].

In [GMS18] the authors provide the first stability and error bound for their local time-stepping scheme (in [DG09] this is completely missing). Unfortunately, their analysis relies on a step-size restriction which is the same as for the standard leapfrog scheme. Just recently, in [GMS21] the authors adapted their originally proposed local time-stepping scheme by integrating the stabilization introduced in [CHS20] for the LFC schemes. For this scheme they provide a rigorous stability and error analysis in the case of $\mathbf{g} \equiv 0$. In contrast to their results, our theory applies to general semilinear problems (1.1), includes positive semidefinite matrices $\mathbf{L}$, and requires a weaker step-size restriction as well as less regularity in time. Furthermore, our analysis holds for a whole class of multirate schemes.

Outline

The thesis is structured as follows. In Chapter 2 we first present the framework we work in and give an overview about analytic properties of the solutions of (2.2). Furthermore, we present two prominent problems admitting differential equations of the form (2.2). In addition, we recall the leapfrog scheme and its most important properties.

In Chapter 3 we introduce a general class of two-step schemes which comprises among others the leapfrog scheme as well as the LFC schemes. For this general class we provide a comprehensive stability and error analysis. Moreover, we show the influence of the starting

value to the stability of these schemes. Concluding, we show that the general results can be improved in the special case of modified θ-scheme.

Chapter 4 is devoted to LFC schemes. We show that these schemes fit into the framework of the general class presented before and state explicit values for all occurring constants. In addition, we show a close relation between the leapfrog scheme and a special case of the LFC schemes. We further present an efficient implementation of the LFC schemes and discuss their efficiency in comparison to the leapfrog scheme as well as to the modified θ-schemes. We conclude with some numerical examples confirming our theoretical results.

The multirate leapfrog-type two-steps schemes are subject of Chapter 5. After a short motivation on these schemes, we present their construction and show some basic (geometric) properties. Afterwards we analyze their stability behavior and prove error bounds. Moreover, for an special case based on the modified θ-schemes we show that a less strict step-size restriction is required to obtain stability compared to the general case. Subsequently, we consider the efficient implementation of two special cases, based on the LFC schemes and the θ-schemes, and compare their efficiency to the leapfrog scheme. Finally, we validate the theoretical results with numerical examples.

In Appendix A we collect postponed calculations concerning the underlying polynomials of the LFC schemes. Appendix B contains auxiliary results which are used frequently throughout this thesis at various passages.

Chapter 2

The leapfrog scheme for semilinear second-order differential equations

This chapter lays the foundations of the thesis. We first recall the general class of differential equations we work on and state some important analytic properties of their solutions. Afterwards we present in Section 2.2 two important examples fitting into this setting. Finally, we review the famous leapfrog scheme in Section 2.3, where we focus on geometric properties and numerical stability.

Notation In the following $(\cdot,\cdot)$ always denotes the standard Euclidean inner product in $\mathbb{R}^m$, $\|\cdot\|$ the corresponding Euclidean norm as well as the induced matrix norm (*spectral norm*). For a symmetric and positive (semi)definite matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ we further abbreviate with $\|\cdot\|_{\mathbf{A}}^2 = (\cdot,\cdot)_{\mathbf{A}} = (\cdot, \mathbf{A}\,\cdot)$ the (semi)norm induced by $\mathbf{A}$ and the corresponding bilinear form and inner product, respectively.

Moreover, for a function $U \in C^2(\mathbb{R}^m,\mathbb{R})$, $\mathbf{q} \mapsto U(\mathbf{q})$, we denote by $\nabla U(\mathbf{q})$ the gradient of U at $\mathbf{q}$ and by $\nabla^2 U(\mathbf{q})$ its Hessian matrix. For functions $H \in C^1(\mathbb{R}^m \times \mathbb{R}^m,\mathbb{R})$ we denote by $\nabla_{\mathbf{q}} H(\mathbf{p},\mathbf{q})$ the (column) vector containing the partial derivatives of H with respect to $\mathbf{q} = (q_1,\dots,q_m)^T$, i.e.,

$$\nabla_{\mathbf{q}} H(\mathbf{p},\mathbf{q}) = \left(\partial_{q_1} H(\mathbf{p},\mathbf{q}),\dots,\partial_{q_m} H(\mathbf{p},\mathbf{q})\right)^T,$$

and analogously for $\mathbf{p}$.

2.1. Semilinear second-order differential equations

Recall that throughout this thesis we focus on the following general class of semilinear second-order differential equations in $\mathbb{R}^m$

$$\ddot{\mathbf{q}}(t) = -\mathbf{L}\mathbf{q}(t) + \mathbf{g}(t,\mathbf{q}(t)), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0, \tag{2.1}$$

with initial values $\mathbf{q}_0, \dot{\mathbf{q}}_0 \in \mathbb{R}^m$ and $t \geq 0$. We assume that the matrix $\mathbf{L} \in \mathbb{R}^{m\times m}$ is symmetric and positive semidefinite, and $\mathbf{g}\colon [0,\infty) \times \mathbb{R}^m \to \mathbb{R}^m$ is a sufficiently smooth function.

We start by showing that this general class also comprises the even more general class of equations

$$\mathbf{M}\ddot{\mathbf{q}}(t) = -\mathbf{L}\mathbf{q}(t) + \mathbf{M}\mathbf{g}(t, \mathbf{q}(t)), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0, \tag{2.2}$$

where the *mass matrix* $\mathbf{M} \in \mathbb{R}^{m\times m}$ is assumed to be symmetric and positive definite.

Lemma 2.1. *Let* $\mathbf{M} = \mathbf{C}_\mathbf{M}\mathbf{C}_\mathbf{M}^T$ *be the Cholesky decomposition of* $\mathbf{M}$ *and define* $\widehat{\mathbf{q}} = \mathbf{C}_\mathbf{M}^T\mathbf{q}$. *Then the second-order differential equation* (2.2) *is equivalent to*

$$\ddot{\widehat{\mathbf{q}}}(t) = -\widehat{\mathbf{L}}\widehat{\mathbf{q}}(t) + \widehat{\mathbf{g}}(t, \widehat{\mathbf{q}}(t)), \qquad \widehat{\mathbf{q}}(0) = \mathbf{C}_\mathbf{M}^T\mathbf{q}_0, \quad \dot{\widehat{\mathbf{q}}}(0) = \mathbf{C}_\mathbf{M}^T\dot{\mathbf{q}}_0, \tag{2.3}$$

where $\widehat{\mathbf{L}} = \mathbf{C}_\mathbf{M}^{-1}\mathbf{L}\mathbf{C}_\mathbf{M}^{-T}$ *and* $\widehat{\mathbf{g}}(\cdot, \widehat{\mathbf{q}}) = \mathbf{C}_\mathbf{M}^T\,\mathbf{g}(\cdot, \mathbf{C}_\mathbf{M}^{-T}\widehat{\mathbf{q}})$.

Proof. Using the definition of $\widehat{\mathbf{q}}$ and multiplying (2.2) with $\mathbf{C}_\mathbf{M}^{-1}$ from left yields (2.3). □

Obviously, $\widehat{\mathbf{L}}$ inherits the symmetry and positive semi-definiteness from $\mathbf{L}$. Consequently, it is sufficient to consider only differential equations of the form (2.1) instead of (2.2). The main results in this and the following chapters hold true for the more general situation by replacing the Euclidean norm $\|\cdot\|$ with $\|\cdot\|_\mathbf{M}$ everywhere it appears. Further, all time integration schemes, which are presented in the following, can be adapted to the general equation (2.2) without explicitly making use of the Cholesky factors $\mathbf{C}_\mathbf{M}$. We comment on changes emerging in the time integration schemes.

Remark 2.2. Another possibility to handle the mass matrix in (2.2) is to consider the equivalent problem

$$\ddot{\mathbf{q}}(t) = -\mathbf{L}^*\mathbf{q}(t) + \mathbf{g}(t, \mathbf{q}(t)), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0,$$

with $\mathbf{L}^* = \mathbf{M}^{-1}\mathbf{L}$. Although $\mathbf{L}^*$ is in general not a symmetric matrix anymore, it is symmetric with respect to the inner product $(\cdot,\cdot)_\mathbf{M}$. By using this inner product and its induced vector and matrix norms the analysis could be performed in a similar way as it is done in the following but with much more technical effort. ◇

Next, we recall some analytic properties of the problem (2.1) and its solutions, which can be found in many monographs about ordinary differential equations; see, e.g., [PW10, Tes12]. Since the results are mostly stated for first-order differential equations, we rewrite (2.1) in such a form by defining $\mathbf{p} = \dot{\mathbf{q}}$. This then yields

$$\dot{\mathbf{q}}(t) = \mathbf{p}(t), \qquad \mathbf{q}(0) = \mathbf{q}_0, \tag{2.4a}$$

$$\dot{\mathbf{p}}(t) = -\mathbf{L}\mathbf{q}(t) + \mathbf{g}(t, \mathbf{q}(t)), \qquad \mathbf{p}(0) = \dot{\mathbf{q}}_0. \tag{2.4b}$$

Note that the transformation for $\mathbf{q}$ in Lemma 2.1 for the general problem (2.2) does not yield any information about the transformation for $\mathbf{p}$. Depending on the considered problem, different transformations for $\mathbf{p}$ have to be used; see Remark 2.6 and Section 2.2 for two important cases.

For differential equations of the form (2.4) it is well-known that a solution exists if the right side and, thus, the semilinearity $\mathbf{g}$ is continuous. For the existence of a unique solution the following assumption is sufficient.

Assumption 2.3. *The function* $\mathbf{g}\colon [0,\infty) \times \mathbb{R}^m \to \mathbb{R}^m$ *is continuous, and locally Lipschitz continuous in the second argument (uniformly with respect to* t*), i.e., for every* $\mathbf{q} \in \mathbb{R}^m$ *and* $T > 0$ *there exist* $\rho = \rho(\mathbf{q}, T) > 0$ *and* $\mathcal{L}_\mathbf{g} = \mathcal{L}_\mathbf{g}(\rho) > 0$ *such that*

$$\|\mathbf{g}(t, \widetilde{\mathbf{q}}) - \mathbf{g}(t, \widehat{\mathbf{q}})\| \leq \mathcal{L}_\mathbf{g}\|\widetilde{\mathbf{q}} - \widehat{\mathbf{q}}\| \tag{2.5}$$

for all $t \in [0, T]$ *and* $\widetilde{\mathbf{q}}, \widehat{\mathbf{q}} \in \mathbb{R}^m$ *with* $\|\widetilde{\mathbf{q}} - \mathbf{q}\|, \|\widehat{\mathbf{q}} - \mathbf{q}\| \leq \rho$.

We remark that we could also allow for a function $\mathbf{g}\colon [0, t_g) \times G \to \mathbb{R}^m$, which lives in an open and connected subset $G \subseteq \mathbb{R}^m$ up to a finite time $t_g > 0$ instead of the whole space. The subsequent results and analysis could be carried out in the same way but with additional technical effort. Clearly, in this case the initial values $\mathbf{q}_0$, $\dot{\mathbf{q}}_0$ have to be chosen in this subset as well.

Theorem 2.4. *Let Assumption 2.3 hold. Then the differential equation (2.1) has a unique solution $\mathbf{q}\colon [0, t_*) \to \mathbb{R}^m$ with $t_* > 0$, where either $t_* = \infty$ or $\lim_{t \to t_*} \|\mathbf{q}(t)\| = \infty$.*

The statements of the theorem follow by applying the theorem of Picard–Lindelöf to (2.4) and using the definition of the maximal existence interval; see, e.g., [PW10, Tes12]. If $\mathbf{g}$ is even globally Lipschitz continuous, the solution always exists for all times $t \geq 0$.

Further, via the *variation-of-constants formula*, we can express the solution of (2.1) for $t \in [0, t_*)$ as

$$\mathbf{q}(t) = \cos(t\boldsymbol{\Lambda})\mathbf{q}_0 + t\,\operatorname{sinc}(t\boldsymbol{\Lambda})\dot{\mathbf{q}}_0 + \int_0^t (t-s)\operatorname{sinc}\big((t-s)\boldsymbol{\Lambda}\big)\,\mathbf{g}\big(s, \mathbf{q}(s)\big)\,\mathrm{d}s, \tag{2.6a}$$

or in the case of a positive definite $\mathbf{L}$ as

$$\mathbf{q}(t) = \cos(t\boldsymbol{\Lambda})\mathbf{q}_0 + \boldsymbol{\Lambda}^{-1}\sin(t\boldsymbol{\Lambda})\dot{\mathbf{q}}_0 + \int_0^t \boldsymbol{\Lambda}^{-1}\sin\big((t-s)\boldsymbol{\Lambda}\big)\,\mathbf{g}\big(s, \mathbf{q}(s)\big)\,\mathrm{d}s, \tag{2.6b}$$

where $\boldsymbol{\Lambda} = \mathbf{L}^{\frac{1}{2}}$ and

$$\operatorname{sinc}\colon \mathbb{R} \to \mathbb{R}, \quad \operatorname{sinc}(z) = \begin{cases} \sin(z)/z, & z \neq 0, \\ 1, & z = 0. \end{cases} \tag{2.7}$$

In the next chapter we derive a discrete analogue to this formula for the general time integration scheme we consider there.

Remark 2.5. In theory we could restrict ourselves to $\mathbf{L}$ positive definite, since we can insert a zero term in the differential equation (2.1) such that we always obtain a positive definite matrix. More precisely, we can rewrite problem (2.1) as

$$\ddot{\mathbf{q}}(t) = -\widetilde{\mathbf{L}}\mathbf{q}(t) + \widetilde{\mathbf{g}}(t, \mathbf{q}(t)), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0,$$

with $\widetilde{\mathbf{g}}(t, \mathbf{q}) = \mathbf{g}(t, \mathbf{q}) + \mathbf{q}$ and $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{I}$. Obviously, $\widetilde{\mathbf{L}}$ is a positive definite matrix with $\|\mathbf{L}^{-1}\| \geq 1$, even if $\mathbf{L}$ is positive semidefinite but not positive definite. Moreover, if $\mathbf{g}$ is locally Lipschitz continuous, so is $\widetilde{\mathbf{g}}$ (with possibly a slightly larger Lipschitz constant). However, if $\mathbf{g}$ only depends on time t, this reformulation would destroy the structure of the linear problem because of the dependency of $\widetilde{\mathbf{g}}$ on $\mathbf{q}$. Hence, we allow $\mathbf{L}$ to be positive semidefinite. ◇

We now consider two special cases, which rely on different properties for $\mathbf{g}$.

Hamiltonian problems

Let the function $\mathbf{g}$ depend only on the solution $\mathbf{q}$, i.e., $\mathbf{g}(t, \mathbf{q}(t)) = \mathbf{g}(\mathbf{q}(t))$ for all $t \geq 0$, and assume there exists a function $U\colon \mathbb{R}^m \to \mathbb{R}$ such that

$$\nabla U(\mathbf{q}) = -\mathbf{g}(\mathbf{q}). \tag{2.8}$$

Then, problem (2.1) or, equivalently, (2.4) can be written as a *Hamiltonian problem*

$$\begin{aligned} \dot{\mathbf{q}} &= \nabla_{\mathbf{p}} \mathcal{H}(\mathbf{p}, \mathbf{q}), & \mathbf{q}(0) &= \mathbf{q}_0, \\ \dot{\mathbf{p}} &= -\nabla_{\mathbf{q}} \mathcal{H}(\mathbf{p}, \mathbf{q}), & \mathbf{p}(0) &= \dot{\mathbf{q}}_0, \end{aligned} \tag{2.9a}$$

with *Hamiltonian* $\mathcal{H} \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ given by

$$(\mathbf{p}, \mathbf{q}) \mapsto \mathcal{H}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2}(\mathbf{p}, \mathbf{p}) + \tfrac{1}{2}(\mathbf{q}, \mathbf{L}\mathbf{q}) + U(\mathbf{q}); \tag{2.9b}$$

see, e.g., [HLW06, Section VI.1] for a short introduction into the topic or [LR04, MO17] for further insights.

It is well known that the solution $(\mathbf{p}, \mathbf{q})$ of (2.9) preserves the Hamiltonian, i.e., we have $\mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) = \mathcal{H}(\dot{\mathbf{q}}_0, \mathbf{q}_0)$ for $t \geq 0$. In particular, the solution exists for all times $t \geq 0$. The conservation of the Hamiltonian can easily be seen by differentiating the Hamiltonian with respect to t and exploiting (2.9a). In our special case this can also be shown with the differential equation (2.1) and $\mathbf{p} = \dot{\mathbf{q}}$.

Another important property of Hamiltonian systems is the symplecticity of its flow if $\mathcal{H}$ is twice continuously differentiable. Recall that the flow φ_t of a differential equation of the form (2.4) is defined in such a way that $\varphi_t(\dot{\mathbf{q}}_0, \mathbf{q}_0) = (\mathbf{p}(t), \mathbf{q}(t))$ for all $t \geq 0$ if $(\mathbf{p}(0), \mathbf{q}(0)) = (\dot{\mathbf{q}}_0, \mathbf{q}_0)$. The symplecticity of this flow is defined via

$$(\varphi_t'(\dot{\mathbf{q}}_0, \mathbf{q}_0))^T \mathbf{J} \varphi_t'(\dot{\mathbf{q}}_0, \mathbf{q}_0) = \mathbf{J}, \qquad \mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}, \tag{2.10}$$

where

$$\varphi_t'(\dot{\mathbf{q}}_0, \mathbf{q}_0) = \frac{\partial \varphi_t(\dot{\mathbf{q}}_0, \mathbf{q}_0)}{\partial(\dot{\mathbf{q}}_0, \mathbf{q}_0)}$$

denotes the Jacobian of the flow of Hamiltonian's equations (2.9a). A proof of this result can be found, e.g., in [HLW06, Theorem VI.2.4]. For $m = 1$ symplecticity can be interpreted as the area preservation of the flow φ_t, i.e., the area of sets of initial values in the $(\mathbf{p}, \mathbf{q})$-plane is preserved over time; see [HLW06, Section VI.2] and [HLW03], where this is nicely illustrated. For more information to symplecticity in general we refer to, e.g., [LR04].

Remark 2.6 (General mass matrix $\mathbf{M}$). For problems (2.2) the Hamiltonian to (2.9a) is given by

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2}(\mathbf{p}, \mathbf{M}^{-1}\mathbf{p}) + \tfrac{1}{2}(\mathbf{q}, \mathbf{L}\mathbf{q}) + U(\mathbf{q}), \tag{2.11}$$

where $\nabla U(\mathbf{q}) = -\mathbf{M}\mathbf{g}(\mathbf{q})$. Thus, we have $\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p}$ and, in particular, $\mathbf{p}(0) = \mathbf{M}\dot{\mathbf{q}}_0$. Moreover, to transform it into a system with $\mathbf{M} = \mathbf{I}$ we require additionally to the definitions in Lemma 2.1 that $\widehat{\mathbf{p}} = \mathbf{C}_{\mathbf{M}}^{-1}\mathbf{p}$. ◇

Linear, inhomogeneous problems

We consider the case that $\mathbf{g}$ only depends on time t, i.e., $\mathbf{g}(t, \mathbf{q}(t)) = \mathbf{g}(t)$ for all $t \geq 0$. Problem (2.1) then reads

$$\ddot{\mathbf{q}}(t) = -\mathbf{L}\mathbf{q}(t) + \mathbf{g}(t), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0. \tag{2.12}$$

Moreover, the solution of (2.12) exists for all times $t \geq 0$.

Since stability is a crucial point of the time integration schemes, which we consider in the next chapters, we state some stability bounds for the exact solution of (2.12). For bounds in the standard norm we require the following definition.

Definition 2.7. *For a symmetric, positive definite matrix* $\mathbf{L}$ *we denote*

$$c_{\mathrm{inv}} = \|\mathbf{L}^{-1/2}\|. \tag{2.13}$$

If $\mathbf{L}$ *is singular, we formally set* $c_{\mathrm{inv}} = \infty$.

Although we mainly focus on bounds for $\mathbf{q}$ in the Euclidean norm $\|\cdot\|$ (or in $\|\cdot\|_{\mathbf{M}}$), we also consider the so-called *energy norm*, which is defined for once differentiable functions $\mathbf{q}\colon [0,T] \to \mathbb{R}^m$ by

$$|||\mathbf{q}(t)|||^2 = \|\dot{\mathbf{q}}(t)\|^2 + \|\mathbf{q}(t)\|_{\mathbf{L}}^2, \qquad t \in [0,T]. \tag{2.14}$$

Note that, if $\mathbf{L}$ is positive semidefinite with at least one zero eigenvalue, the energy norm is not a true norm but only a seminorm, since $|||\mathbf{q}(t)||| = 0$ for every constant functions consisting of the corresponding eigenvectors, i.e., which lie in the kernel of $\mathbf{L}$.

Lemma 2.8. *The solution of the linear problem* (2.12) *satisfies for all* $t \geq 0$

$$\|\mathbf{q}(t)\| \leq \|\mathbf{q}_0\| + \min\{t, c_{\mathrm{inv}}\}\|\dot{\mathbf{q}}_0\| + \min\{t, c_{\mathrm{inv}}\} \int_0^t \|\mathbf{g}(s)\| \,\mathrm{d}s, \tag{2.15a}$$

$$|||\mathbf{q}(t)||| \leq |||\mathbf{q}(0)||| + \int_0^t \|\mathbf{g}(s)\| \,\mathrm{d}s. \tag{2.15b}$$

Proof. The first bound is a direct consequence of formulae (2.6). To prove the second bound we differentiate the energy norm with respect to t. This yields on the one hand

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||\mathbf{q}(t)|||^2 = (\dot{\mathbf{q}}(t), \ddot{\mathbf{q}}(t)) + (\dot{\mathbf{q}}(t), \mathbf{L}\mathbf{q}(t)) = (\dot{\mathbf{q}}(t), \mathbf{g}(t)) \leq |||\mathbf{q}(t)||| \, \|\mathbf{g}(t)\|,$$

where we used (2.12), Cauchy-Schwarz inequality, and the definition (2.14) of $|||\cdot|||$. On the other hand we also have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|||\mathbf{q}(t)|||^2 = |||\mathbf{q}(t)||| \frac{\mathrm{d}}{\mathrm{d}t}|||\mathbf{q}(t)|||,$$

where we assumed without loss of generality that $|||\mathbf{q}(t)||| > 0$ for all $t \geq 0$. Combining both leads to the desired bound. □

From (2.15a) we observe that, if $\mathbf{L}$ is singular, the minimum is always attained for $t \geq 0$, since $c_{\mathrm{inv}} = \infty$. Hence, for $g \equiv 0$ the bound grows linearly in time $t \geq 0$, whereas for a positive definite matrix it stays uniformly bounded.

For $g \equiv 0$ we additionally have that the energy norm is preserved, i.e.,

$$|||\mathbf{q}(t)||| = |||\mathbf{q}(0)||| \qquad \text{for all } t \geq 0, \tag{2.16}$$

which follows with almost the same arguments as the bound for (2.15b). Alternatively, one can show this result by using that $|||\mathbf{q}(t)|||^2 = 2\mathcal{H}(\dot{\mathbf{q}}(t), \mathbf{q}(t)) + c$ with a fixed constant $c \in \mathbb{R}$.

2.2. Examples for semilinear second-order differential equations

Next, we present two important problems admitting differential equations of the form (2.2), which we use later for our numerical simulations.

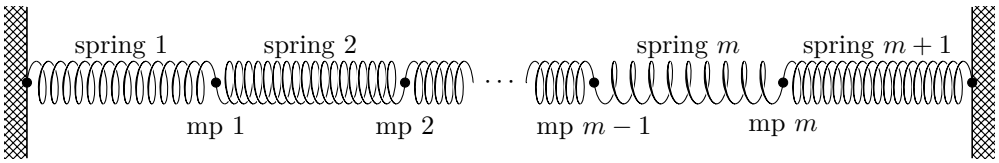


Figure 2.1.: Illustration of the chain of mass points (mp) from the FPUT β-problem.

2.2.1. A modified Fermi–Pasta–Ulam–Tsingou problem

We start with a modification of the famous Fermi–Pasta–Ulam–Tsingou (FPUT) β-problem [FPU+55]. The problem describes the motion of a chain consisting in total of $m+2$ mass points which are connected via nonlinear springs. The mass points at the end of the chain are fixed. An illustration of this model is given in Figure 2.1. In contrast to the original problem, we consider a slightly more general setting: the material constants of the springs and the masses of the points can differ for each spring and mass point, respectively.

In the following, q_i denotes the displacement of the ith mass point from its equilibrium, $i = 0, \dots, m+1$. Since the endpoints are fixed, we set $q_0 = q_{m+1} = 0$. Further, we denote with $\mu_i > 0$ the mass of the ith point and with k_i, $\beta_i^* \geq 0$ the spring constants of the ith spring related to linear and nonlinear material laws, respectively.

Newton's second law, Hooke's law, and a nonlinear (cubic) extension thereof lead for the ith (inner) mass point to the differential equation

$$\mu_i \ddot{q}_i = k_{i+1}(q_{i+1} - q_i) - k_i(q_i - q_{i-1}) + \beta_{i+1}^*(q_{i+1} - q_i)^3 - \beta_i^*(q_i - q_{i-1})^3, \quad i = 1, \dots, m.$$

By setting $\mathbf{q} = (q_1, \dots, q_m)^T$ we obtain a system of ordinary differential equations of the form (2.2) with a diagonal mass matrix $\mathbf{M} = \operatorname{diag}(\mu_1, \dots, \mu_m) \in \mathbb{R}^{m \times m}$, a tridiagonal matrix

$$\mathbf{L} = \begin{pmatrix} k_{1,2} & -k_2 & 0 & \cdots & 0 \\ -k_2 & k_{2,3} & -k_3 & \ddots & \vdots \\ 0 & -k_3 & k_{3,4} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -k_m \\ 0 & \dots & 0 & -k_m & k_{m,m+1} \end{pmatrix}, \quad k_{i,i+1} = k_i + k_{i+1},\ i = 1, \dots, m,$$

and $\mathbf{g} = (g_1, \dots, g_m)^T$ given by

$$\mu_i g_i = \check{g}_i, \quad \check{g}_i(\mathbf{q}) = \beta_{i+1}^*(q_{i+1} - q_i)^3 - \beta_i^*(q_i - q_{i-1})^3, \qquad i = 1, \dots, m.$$

We point out that for the nonlinear term we have to artificially insert the mass matrix, because it does not appear naturally in the formulation. Further, the Hamiltonian, which describes the total energy of the system, is given by (2.11) with

$$U(\mathbf{q}) = \frac{1}{4} \sum_{i=0}^{m} \beta_{i+1}^*(q_{i+1} - q_i)^4, \tag{2.17}$$

where $\nabla_{\mathbf{q}} U(\mathbf{q}) = -\mathbf{M}\mathbf{g}(\mathbf{q}) = -\check{\mathbf{g}}(\mathbf{q})$; cf. Remark 2.6.

2.2.2. Spatially discretized acoustic wave equation

As second problem, we consider space discretizations of partial differential equations which admit differential equations of the form (2.2). Here, we focus on the semilinear acoustic wave equation in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. Equipped with homogeneous Dirichlet boundary conditions and initial values, the problem reads

$$\begin{aligned} \ddot{q}(t,x) &= \nabla \cdot (c(x)\nabla q(t,x)) + g(t,x,q(t,x)), && t \in [0,T], x \in \Omega, \\ q(t,x) &= 0, && t \in [0,T], x \in \partial\Omega, \\ q(0,x) &= q_0(x), \quad \dot{q}(0,x) = \dot{q}_0(x), && x \in \Omega, \end{aligned} \tag{2.18}$$

where the functions $c, q_0, \dot{q}_0 \colon \Omega \to \mathbb{R}$ and $g \colon [0,T] \times \Omega \times \mathbb{R} \to \mathbb{R}$ are given. This equation models, for instance, the traveling of sound waves through the domain Ω. In this case, the function $q \colon [0,T] \times \Omega \to \mathbb{R}$ describes the pressure of these sound waves and the function c the speed of sound of the underlying material in the domain.

To discretize this partial differential equation in space, we use a *discontinuous Galerkin finite element method* (dG-FEM); see, e.g., the monographs [DPE12, HW08]. More precisely, we employ for the discretization of the differential operators a *symmetric interior penalty* dG-FEM, which was originally proposed in [Arn82] for $c \equiv 1$. Since we consider for the numerical example in Chapter 5 also piecewise constant functions c, we employ the variant proposed in [GSS06]. Another variant for c piecewise constant is derived in [Dry03]; see also [DPE12, Chapter 4].

By interpolating the semilinearity g or, in the case of an inhomogeneity, projecting it the space discretization of (2.18) results in the differential equation (2.2) with symmetric positive definite matrices $\mathbf{M}$ and $\mathbf{L}$. The boundary condition in (2.18) is (weakly) enforced through $\mathbf{L}$. Since in the dG-FEM each *degree of freedom* (dof) only belongs to one element of the used mesh, the mass matrix $\mathbf{M}$ is block diagonal, where the size of each block is equal to the number of dofs in the corresponding mesh element. Thus, solving with $\mathbf{M}$ can be done at low cost if there are not too many dofs in the mesh elements.

We point out that other space discretization methods, such as continuous finite element methods or finite difference methods, could be used as well. In particular, these methods lead in most cases also to (2.2) with symmetric, positive definite matrices $\mathbf{L}$ and $\mathbf{M}$.

We conclude this example with some comments about the well-posedness of the continuous as well as the discretized problem and analytic properties of their solutions. By interpreting the partial differential equation (2.18) as an evolution equation one can show via semigroup theory [Paz83, Sho97] that under suitable assumptions on the initial values q, $\dot{q}$, the functions c and g, as well as the domain Ω the problem is well-posed and the unique solution q exists for at least a finite time $T > 0$. For precise assumptions on the data we refer to, e.g., [BDH21, Example 3.1].

For the discretized problem the existence of a unique solution via the Picard-Lindelöf theorem as stated in the previous section has to be taken with care. The reason for this is that t_* given in (2.4) tends to 0 if $h \to 0$ because of $\|\mathbf{M}^{-1}\mathbf{L}\| \sim h^{-2}$, where h denotes the maximum diameter of all mesh elements. This can be fixed by proceeding as for the well-posedness of the continuous problem.

A further difficulty occurs in Assumption 2.3 on the Lipschitz continuity of $\mathbf{g}$. For almost all functions g of the continuous problem, the discretized versions $\mathbf{g}$ only yield Lipschitz constants $\mathcal{L}_{\mathbf{g}}$ which behave as h^{-1}; see, e.g., again [BDH21, Example 3.1 and Table 1]. Hence, to allow for a larger class of functions g in this example, we assume a further, yet weaker Lipschitz condition on $\mathbf{g}$ instead of Assumption 2.3. For this, we require without loss of generality that $\mathbf{L}$ is positive definite; cf. Remark 2.5.

Assumption 2.9. *The function* $\mathbf{g}\colon [0,\infty)\times\mathbb{R}^m\to\mathbb{R}^m$ *is continuous, and locally Lipschitz continuous in the second argument (uniformly with respect to* t*), i.e., for every* $\mathbf{q}\in\mathbb{R}^m$ *and* $T>0$ *there exist* $\widehat{\rho}=\widehat{\rho}(\mathbf{q},T)>0$ *and* $\widehat{\mathcal{L}}_{\mathbf{g}}=\widehat{\mathcal{L}}_{\mathbf{g}}(\widehat{\rho})>0$ *such that*

$$\|\mathbf{g}(t,\widetilde{\mathbf{q}})-\mathbf{g}(t,\widehat{\mathbf{q}})\|_{\mathbf{L}^{-1}}\leq\widehat{\mathcal{L}}_{\mathbf{g}}\|\widetilde{\mathbf{q}}-\widehat{\mathbf{q}}\| \tag{2.19}$$

for all $t\in[0,T]$ *and* $\widetilde{\mathbf{q}},\widehat{\mathbf{q}}\in\mathbb{R}^m$ *with* $\|\widetilde{\mathbf{q}}-\mathbf{q}\|,\|\widehat{\mathbf{q}}-\mathbf{q}\|\leq\widehat{\rho}$.

2.3. Leapfrog scheme

We conclude this chapter by recalling the leapfrog scheme and some of its most important properties; see, e.g., [HLW03] or [HLW06]. As mentioned in the introduction, the leapfrog scheme is an explicit time integration scheme and probably the by far most widely used method to solve differential equations of type (2.1) numerically, or, more generally, differential equations of the form $\ddot{\mathbf{q}}(t)=f(t,\mathbf{q}(t))$ with a (sufficiently smooth) function $f\colon[0,\infty)\times\mathbb{R}^m\to\mathbb{R}^m$. Depending on the application (molecular dynamics, spatially discretized partial differential equations, etc.) and the precisely used formulation the scheme is also known as *Verlet* method, *Störmer* method, or combinations of it (*Störmer–Verlet*). In [HLW03, Subsection 1.3] this is explained nicely together with a brief overview about the historical development of the scheme.

In the following we denote by $\tau>0$ the step size in time and by $\mathbf{q}_n$ the approximations to the exact solution $\mathbf{q}$ of (2.1) at time $t_n=n\tau$, in short $\mathbf{q}_n\approx\mathbf{q}(t_n)$. Moreover, we abbreviate $\mathbf{g}_n=\mathbf{g}(t_n,\mathbf{q}_n)$.

2.3.1. Derivation and equivalent formulations

There exist several ways to derive the leapfrog scheme in one of its various variants; see, e.g., [HLW03]. Probably the easiest possibility is to replace the second derivative of $\mathbf{q}$ by a central second-order difference quotient. Applied to the semilinear problem (2.1) this yields the two-step formulation of the leapfrog scheme

$$\mathbf{q}_{n+1}-2\mathbf{q}_n+\mathbf{q}_{n-1}=\tau^2(-\mathbf{L}\mathbf{q}_n+\mathbf{g}_n),\qquad n=1,2,\dots. \tag{2.20a}$$

The scheme is usually completed with the second-order Taylor approximation to $\mathbf{q}(\tau)$ as starting value

$$\mathbf{q}_1=\mathbf{q}_0+\tau\dot{\mathbf{q}}_0+\tfrac{1}{2}\tau^2(-\mathbf{L}\mathbf{q}_0+\mathbf{g}_0). \tag{2.20b}$$

In applications often an equivalent one-step formulation for the first-order system (2.4) is used instead of the above two-step scheme, especially if one is interested in approximations of the derivative $\mathbf{p}=\dot{\mathbf{q}}$. Moreover, many theoretical aspects are, at least initially, formulated only for one-step schemes, e.g., symplecticity. By defining the approximations $\mathbf{p}_{n+1/2}=\frac{1}{\tau}(\mathbf{q}_{n+1}-\mathbf{q}_n)\approx\dot{\mathbf{q}}(t_{n+1/2})$ the two-step scheme (2.20a) can be reformulated in a first step as a scheme on a staggered time grid

$$\begin{aligned}\mathbf{p}_{n+1/2}&=\mathbf{p}_{n-1/2}+\tau(-\mathbf{L}\mathbf{q}_n+\mathbf{g}_n),\\ \mathbf{q}_{n+1}&=\mathbf{q}_n+\tau\mathbf{p}_{n+1/2},\end{aligned}\qquad n=1,2,\dots.$$

The name leapfrog scheme originates from this formulation since the approximations $\mathbf{q}_n$ and $\mathbf{p}_{n+1/2}$ "leapfrog" over each other along the staggered grid.

If we additionally define the average $\mathbf{p}_n = \frac{1}{2}(\mathbf{p}_{n+1/2} + \mathbf{p}_{n-1/2})$ for $n \geq 1$, we get the following equivalent one-step formulation of the scheme (2.20)

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \tfrac{1}{2}\tau(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \tag{2.21a}$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\mathbf{p}_{n+1/2}, \qquad n = 0, 1, \ldots, \tag{2.21b}$$

$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \tfrac{1}{2}\tau(-\mathbf{L}\mathbf{q}_{n+1} + \mathbf{g}_{n+1}), \tag{2.21c}$$

where we set $\mathbf{p}_0 = \dot{\mathbf{q}}_0$. Note that the scheme naturally incorporates the starting value (2.20b), which is probably the main reason to choose it in this way.

The computation of the one-step scheme (2.21) requires as main cost – as for the two-step scheme (2.20a) – only one matrix-vector multiplication with $\mathbf{L}$ and one evaluation of $\mathbf{g}$ per time step except of the first one. This can be achieved by either reusing the computations from the previous time step, taking the variant on the staggered grid, or employing $\mathbf{p}_{n+1/2} = 2\mathbf{p}_n - \mathbf{p}_{n-1/2}$.

Remark 2.10 (General mass matrix $\mathbf{M}$). If we apply the leapfrog scheme (2.20) to the general problem (2.2), it is obviously not fully explicit anymore, since one has to solve a linear system with the mass matrix $\mathbf{M}$ in each step. This is in many situations, however, not a true restriction. On the one hand, $\mathbf{M}$ is often simple to invert because it is diagonal or block diagonal. On the other hand, the condition number of the mass matrix is mostly small such that iterative methods, e.g., Krylov methods, already yield very good results after a few iterations. For the examples considered in the previous section, for instance, we have that $\mathbf{M}$ is diagonal and block diagonal, respectively. ◇

2.3.2. Properties

We first show the symmetry and symplecticity of the leapfrog scheme before we turn towards stability. We refer again to [HLW03] for more insight into geometric properties of the leapfrog scheme and to [HLW06] for geometric properties in general.

We start with the symmetry, which is defined for one-step methods as follows.

Definition 2.11. *A numerical one-step method applied to (2.4) is called* symmetric *if the numerical flow* $\Phi_\tau\colon (\mathbf{p}_n, \mathbf{q}_n) \mapsto (\mathbf{p}_{n+1}, \mathbf{q}_{n+1})$ *satisfies* $\Phi_\tau = \Phi_{-\tau}^{-1}$.

In other words, symmetry means that by applying the scheme with the negative step size $-\tau$ to the last approximation one gets back the previous ones. For this reason, symmetry is also called *time-reversibility*. For the two-step variant of the leapfrog scheme we can employ the definition of symmetry for linear multistep methods for second-order differential equations; see, e.g., [HNW93, Section III.10]. For later use we generalize this definition to multistep methods for second-order differential equations $\ddot{\mathbf{q}}(t) = f(t, \mathbf{q}(t))$ of the form

$$\sum_{i=0}^{k} \alpha_i \mathbf{q}_{n+i} = \tau^2 \sum_{i=0}^{k} \beta_i(\tau^2\mathbf{A}) f(t_{n+i}, \mathbf{q}_{n+i}) \qquad n = 0, 1, \ldots, \tag{2.22}$$

where the coefficients $\beta_j\colon G \to \mathbb{R}$, $z \mapsto \beta_j(z)$, are sufficiently smooth on $G \subseteq \mathbb{R}$ such that $\beta_j(\tau^2\mathbf{A})$ is well defined for $\mathbf{A} \in \mathbb{R}^{m\times m}$, $j = 0, \ldots, k$; cf. Section B.3 for a definition of matrix functions.

Definition 2.12. *The multistep method (2.22) is called* symmetric, *if*

$$\alpha_j = \alpha_{k-j}, \qquad \beta_j(z) = \beta_{k-j}(z) \quad \text{for all } z \in G, \qquad j = 0, 1, \ldots, k. \tag{2.23}$$

For a two-step method, i.e., $k = 2$, this means that $\alpha_0 = \alpha_2$ and $\beta_0(\cdot) = \beta_2(\cdot)$. Hence, similarly as for one-step methods the definition says that an exchange of $(\mathbf{q}_{n+1}, t_{n+1})$ with $(\mathbf{q}_{n-1}, t_{n-1})$ and τ with $-\tau$ yields again the same method.

Lemma 2.13. *The two-step variant (2.20a) and the one-step variant (2.21) of the leapfrog scheme are symmetric.*

Proof. The proof for the two-step scheme immediately follows from Definition 2.12 (with $k = 2$). For the symmetry of the one-step scheme we refer to, e.g., [HLW06, Theorem V.2.5]. □

For symplecticity we first recall the definition; see, e.g., [HLW06, Definition VI.3.1]. As symplecticity is a property of Hamiltonian problems, it is initially only defined for one-step methods. Moreover, the two-step variant of the leapfrog scheme does not directly admit an approximation for $\mathbf{p}_n$.

Definition 2.14. *A numerical one-step method is called* symplectic*, if, whenever the method is applied to a sufficiently smooth Hamiltonian system (2.9), the numerical flow $\Phi_\tau\colon (\mathbf{p}_n, \mathbf{q}_n) \mapsto (\mathbf{p}_{n+1}, \mathbf{q}_{n+1})$ is symplectic, i.e., the Jacobian*

$$\Phi'_\tau(\mathbf{p}_n, \mathbf{q}_n) = \frac{\partial \Phi_t(\mathbf{p}_n, \mathbf{q}_n)}{\partial(\mathbf{p}_n, \mathbf{q}_n)}$$

of the numerical flow satisfies for all step sizes τ and $(\mathbf{p}_n, \mathbf{q}_n) \in \mathbb{R}^m \times \mathbb{R}^m$

$$\Phi'_\tau(\mathbf{p}_n, \mathbf{q}_n)^T \, \mathbf{J} \, \Phi'_\tau(\mathbf{p}_n, \mathbf{q}_n) = \mathbf{J}, \tag{2.24}$$

where the matrix $\mathbf{J} \in \mathbb{R}^{2m \times 2m}$ is defined as in (2.10).

For a symplectic one-step method it is well-known that it provides approximations which are the exact solution of a perturbed Hamiltonian problem. Thus, they preserve the (perturbed) Hamiltonian for arbitrary long times if we make some additional assumptions; see, e.g., [HLW06, Theorem IX.3.1, Theorem IX.8.1]. For more details to Hamiltonian problems and symplectic schemes we refer to [HLW06, Chapter VI, IX] and for the special case of the leapfrog scheme also to [HLW03, Section 5].

Lemma 2.15. *The leapfrog scheme (2.21) is symplectic.*

Proof. We refer to [HLW03] containing five different proofs of the symplecticity of the leapfrog scheme. The first proof goes back to [DV56], see also [SC20]. □

Besides the nice geometric properties it is desirable for a numerical scheme to exhibit a similar stability behavior as the exact solution. In particular, for linear problems (2.12) with $g \equiv 0$ we would like to have approximations which are bounded or grow at most linearly in time; cf. (2.15) and (2.16). To investigate the stability behavior of the leapfrog scheme we consider, similarly as for first-order equations, a scalar linear test problem

$$\ddot{q}(t) = -\omega^2 q(t), \qquad q(0) = q_0, \quad \dot{q}(0) = \dot{q}_0, \tag{2.25}$$

with $\omega \geq 0$. From (2.15a) we then know $|q(t)| \leq |q_0| + \min\{t, \omega^{-1}\}|\dot{q}_0|$ for $t \geq 0$. Note that this problem describes the motion of an (undamped) harmonic oscillator. Alternatively, it can also be interpreted as special case of the modified FPUT β-problem in Section 2.2.1 by setting $m = 1$, $\mu_1 = 1$, $\beta_1^* = \beta_2^* = 0$, and $\omega^2 = k_{1,2} = k_1 + k_2$.

Lemma 2.16. *Let $\vartheta \in (0,1]$. If $\tau\omega \le 2\vartheta$, the leapfrog scheme (2.20) applied to the linear test equation (2.25) satisfies*

$$|q_n| \le |q_0| + \min\{t_n, \omega^{-1}(1-\vartheta^2)^{-1/2}\}|\dot q_0|.$$

For $\tau\omega > 2$, we have $q_n \sim c\,\xi^n$ for $n \to \infty$ with $|\xi| > 1$ and in general $c \neq 0$.

Obviously, for $\vartheta < 1$ the leapfrog scheme shows a very similar stability behavior as the exact solution. For $\vartheta = 1$ and $\dot q_0 \neq 0$ the numerical solution grows linearly in time, which, for $\omega \neq 0$, is already different from the behavior of the exact solution. In the next chapter we show a possibility how one can avoid the factor $(1-\vartheta^2)^{-1/2}$ for linear problems; see Section 3.5.2.

Proof. Let $\tau > 0$. The two-step recursion (2.20a) of the leapfrog scheme applied to (2.25) leads to the linear recurrence relation

$$q_{n+1} - 2q_n + q_{n-1} = -\tau^2\omega^2 q_n.$$

The roots of the corresponding characteristic polynomial are given by

$$\xi_\pm = 1 - \tfrac{1}{2}\zeta^2 \pm \tfrac{1}{2}\Big(\zeta^2(\zeta^2-4)\Big)^{1/2}, \qquad \zeta = \tau\omega.$$

We now distinguish four cases.

(i) Let $\zeta > 2$. This yields $\xi_\pm \in \mathbb{R}$, where $\xi_+ \in (-1,1)$ and $\xi_- < -1$. Moreover,

$$q_n = c_+\xi_+^n + c_-\xi_-^n, \qquad n = 0,1,\dots,$$

with $c_\pm \in \mathbb{R}$ depending on q_0 and $\dot q_0$ by employing the starting value (2.20b). Thus, in general q_n grows exponentially in n.

(ii) Let $\zeta = 2$, i.e., $\vartheta = 1$. We then have $\xi_\pm = -1$, which implies $q_n = \check c_+(-1)^n + \check c_- n(-1)^n$ with $\check c_\pm \in \mathbb{R}$. With the starting value (2.20b) we obtain

$$q_n = q_0(-1)^n - \dot q_0\,\tau n(-1)^n, \qquad n = 0,1,\dots,$$

leading to $|q_n| \le |q_0| + t_n|\dot q_0|$.

(iii) Let $\vartheta \in (0,1)$ and $0 < \zeta < 2\vartheta$. Thus, we have $\xi_\pm \in \mathbb{C}$ with $|\xi_\pm| = 1$, which yields

$$q_n = \hat c_+\xi_+^n + \hat c_-\xi_-^n, \qquad n = 0,1,\dots,$$

with $\hat c_\pm \in \mathbb{C}$. Employing the starting value (2.20b) leads to

$$q_n = q_0\,\tfrac{1}{2}(\xi_+^n + \xi_-^n) + \tau\dot q_0\,\frac{\xi_+^n - \xi_-^n}{\xi_+ - \xi_-}, \qquad n = 0,1,\dots,$$

(note that $\xi_+^n + \xi_-^n \in \mathbb{R}$ and $\xi_+^n - \xi_-^n \in \mathrm{i}\mathbb{R}$ for all $n \in \mathbb{N}$). Since $|\xi_+ - \xi_-| \ge 2\tau\omega(1-\vartheta^2)^{1/2}$, we obtain on the one hand $|q_n| \le |q_0| + \omega^{-1}(1-\vartheta^2)^{-1/2}|\dot q_0|$. On the other hand we have

$$\frac{|\xi_+^n - \xi_-^n|}{|\xi_+ - \xi_-|} \le \sum_{k=0}^{n-1} |\xi_+^k||\xi_+^{n-1-k}| = n,$$

which yields $|q_n| \le |q_0| + t_n|\dot q_0|$.

(iv) For $\zeta = 0$, i.e., $\omega = 0$, we have $\xi_\pm = 1$, which yields the same bound as in part (ii). □

Finally, we note that one can easily check that the leapfrog scheme is of classical order two, for instance, by checking the order conditions for linear multistep methods for second-order ordinary differential equations; see, e.g., [HNW93, Theorem III.10.3]. In the context of (spatially discretized) partial differential equations it is shown for various equations and space discretizations that the second-order convergence still holds; see, for instance, [CDW96, Jol03, GS09, BV09]. Moreover, the error analysis of a general class of two-step schemes in the next chapter covers the case of the leapfrog scheme.

Remark 2.17. Throughout this thesis we focus – for several reasons – on fixed step sizes $\tau > 0$ for all time steps. The most important reason is that variable step sizes destroy the favorable stability behavior and the symplecticity of the leapfrog scheme in general; see [Ske93] and [HLW06, Chapter VIII], respectively. The latter reference contains an overview of the problematic nature of adaptivity for symplectic schemes and presents some workarounds to retain symplecticity with adaptive step sizes. Moreover, our stability proofs in the following chapters heavily rely on the fact that the step size is constant for all time steps. ◇

Chapter 3

A general class of two-step schemes

In this chapter we introduce and analyze a class of two-step schemes for the numerical time integration of semilinear second-order differential equations (2.1) considered in the last chapter. The aim of constructing and analyzing these schemes is to obtain time integration methods which are stable for larger step sizes than the leapfrog scheme while retaining symmetry, symplecticity, and second-order convergence. We are mainly interested in explicit schemes, although the general scheme also includes implicit ones and even a trigonometric integrator.

The situation we have in mind for applying the schemes are cases where the linear part of the differential equation (2.1) is responsible for the stiffness of the differential equation. In contrast, the semilinearity $\mathbf{g}$ is a "nice" function with small Lipschitz constant but expensive to evaluate. More precisely, we are interested in situations in which the evaluation of $\mathbf{g}$ is approximately as costly as or more costly than the computation of the matrix-vector product $\mathbf{Lq}$. Instead of a semilinearity one can imagine also an inhomogeneity which is expensive to compute compared to the linear part. The efficiency of a specific case of this general scheme in such situations is shown in the next chapter.

In the following, we first introduce the general two-step scheme and state some general properties of it. Afterwards we derive a representation formula for the numerical solution in Section 3.2 which is the basic analytic tool for our stability and error analysis in Sections 3.3 to 3.5. We conclude this chapter with a specific case of the general two-step scheme for which we show some improved stability bounds.

The results in this chapter contain variants and extensions of results published in [CHS20]. In comparison to this paper we consider and analyze a slightly different – and more general – class of two-step schemes. Moreover, by a refined analysis we improve some of the bounds stated there. In particular, Sections 3.2.2, 3.5.2, and 3.6 are not published so far in any variation.

Remark 3.1 (History of generating functions). As fundamental tool for deriving the representation formula of the numerical approximations we make use of the technique of *generating functions*. As (ordinary) generating function one denotes the formal power series $\sum_{n=0}^{\infty} a_n \zeta^n$, $\zeta \in \mathbb{C}$, of a given sequence $(a_n)_{n\in\mathbb{N}_0}$. This technique was first introduced by Abraham de Moivre (1667-1754) in *The Doctrine of Chances: Or, a Method of Calculating the Probabilities of Events in Play* (1718) to solve linear recurrence relations [Sch05]. Later, Leonhard Euler (1707-1783) extended this technique in several papers and to several problems. Pierre-Simon Laplace (1749-1827) further developed it in *Théorie analytique des Probabilités* (1812). He

was the first who called the formal power series generating function [Sti05]. In the context of numerical methods for differential equations, early uses are, for instance, the computation of the coefficients of specific linear multistep methods; see, e.g., [Hen62]. In [Lub83] the approach is used to state a representation formula for the approximations obtained by linear multistep methods applied to Volterra equations of second kind. ◇

3.1. Generalizations of the leapfrog scheme

We start with the construction of a general class of two-step schemes. To achieve our goal of ensuring stability for larger step sizes than the leapfrog scheme, we *modify the right-hand side* of the leapfrog scheme by inserting a suitable matrix function $\widehat{\Psi}(\tau^2\mathbf{L})$; see Section B.3 for the definition of such functions. This modification is motivated by the so-called *modified equation approach*; see, e.g., [SB87, GJ08, JR10]. Similar approaches are used for enlarging the stability region of explicit Runge–Kutta methods by adjusting the polynomial stability function; see, e.g., [HV03, Chapter V]. For larger generality we restrict ourselves not only to polynomials but to a larger class of functions.

Assumption 3.2. $\widehat{\Psi}\colon [0,\infty) \to \mathbb{R}$ *is sufficiently smooth and satisfies* $\widehat{\Psi}(0) = 1$.

As we will see later, the condition $\widehat{\Psi}(0) = 1$ is necessary to obtain at least second-order consistency. In principle, we could restrict ourselves to analytic functions, because all functions we consider are of this type, but, since our analysis does not require this, we allow for more general functions.

Since the semilinearity (or inhomogeneity) $\mathbf{g}$ is assumed to be expensive to compute, we retain the single explicit evaluation of $\mathbf{g}$ per time step as in the leapfrog scheme. Together, this yields the general two-step scheme

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2\widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \qquad n = 1, 2, \dots, \tag{3.1a}$$

which we equip with the starting value

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau\widehat{\Psi}(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2\widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_0 + \mathbf{g}_0). \tag{3.1b}$$

Clearly, for $\widehat{\Psi} \equiv 1$ the general scheme (3.1) reduces to the leapfrog method (2.20). The additional factor of $\widehat{\Psi}$ in front of $\dot{\mathbf{q}}_0$ in the starting value yields better stability bounds, as we will see in the next sections. In general, the choice of the starting value is a delicate matter, rather in terms of stability than of convergence order; see Section 3.5.1. In particular, we see that in specific situations other options may perform better.

For a simpler notation in the following we introduce another function related to $\widehat{\Psi}$ instead of the function $\widehat{\Psi}$ itself. In addition, it turns out that it is more convenient to work with this related function in the subsequent stability and error analysis.

Definition 3.3. *For* $\widehat{\Psi}\colon [0,\infty) \to \mathbb{R}$ *we define the function* $\Psi\colon [0,\infty) \to \mathbb{R}$ *by*

$$\Psi(z) = z\widehat{\Psi}(z) \qquad \text{for } z \geq 0. \tag{3.2}$$

Since we mainly aim for explicit schemes, we are particularly interested in polynomials for $\widehat{\Psi}$ and thus Ψ. More precisely, we are interested in a specific class of polynomials for Ψ, which is based on the Chebyshev polynomial of the first kind T_p; see Section B.2 for a definition and

some properties of these polynomials. These polynomials, which we focus on in this thesis, are given by

$$\Psi(z) = P_p(z) = 2 - \frac{2}{T_p(\nu)} T_p\Big(\nu - \frac{z}{\alpha_p}\Big), \qquad \alpha_p = 2\frac{T_p'(\nu)}{T_p(\nu)}, \tag{3.3}$$

where we call $\nu \geq 1$ *stabilization parameter*. In the next chapter we show that the polynomials (3.3) satisfy Assumption 3.2 as well as all other assumptions made in this chapter and state explicit values for all occurring constants. Further, we demonstrate the efficiency of these schemes in the above mentioned situation. We call the combination of the two-step scheme (3.1a) with the polynomials (3.3) *leapfrog-Chebyshev* (LFC) *schemes* for obvious reasons, and accordingly the polynomials (3.3) *leapfrog-Chebyshev polynomials*.

Other important functions for Ψ are, for instance, the rational functions

$$\Psi(z) = \Psi_\theta(z) = z\widehat{\Psi}_\theta(z) = \frac{z}{1+\theta z}, \qquad \theta \geq 0, \tag{3.4}$$

which are motivated by *θ-schemes*; see Section 3.6 at the end of this chapter. Clearly, $\widehat{\Psi}_\theta$ satisfies Assumption 3.2 for every $\theta \geq 0$. We emphasize that for $\theta > 0$ the resulting scheme (3.1) is implicit, and for $\theta = 0$ it reduces to the leapfrog scheme (2.20).

It is worth mentioning that the general two-step scheme (3.1a) also comprises the trigonometric integrator introduced by Gautschi [Gau61, HLW06, Section XIII.1], given by

$$\mathbf{q}_{n+1} - 2\cos(\tau\mathbf{L}^{1/2})\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2 \operatorname{sinc}\big(\tfrac{1}{2}\tau\mathbf{L}^{1/2}\big)^2 \mathbf{g}_n. \tag{3.5}$$

This can be seen by defining

$$\Psi(z) = \Psi_{\text{trig}}(z) = 2 - 2\cos(z^{1/2}), \tag{3.6}$$

from which we obtain $\widehat{\Psi}(z) = \operatorname{sinc}\big(\frac{1}{2}z^{1/2}\big)^2$ because of (B.2b) (note that $\operatorname{sinc}(0) = 1$). However, since we are mainly interested in explicit schemes, the subsequent (error) analysis is especially designed for polynomials as Ψ and does not employ the additional prerequisites which hold for trigonometric integrators. In addition, the term containing $\dot{\mathbf{q}}_0$ in the starting value $\mathbf{q}_1$ does not correspond to the one for trigonometric integrators.

Remark 3.4 (General mass matrix $\mathbf{M}$). As stated at the beginning of Section 2.1, we can extend the schemes to the general differential equation (2.2). For these cases we simply have to replace $\mathbf{L}$ by $\mathbf{M}^{-1}\mathbf{L}$ in the schemes (3.1) in accordance with Remark 2.2. We emphasize that for the actual implementation of these schemes the inverse of $\mathbf{M}$ does not have to be calculated explicitly; see Section 3.6.3 and Section 4.4, in which implementations of (3.1) for $\Psi = \Psi_\theta$ and $\Psi = P_p$ are given. ◇

Finally, we note that instead of modifying the whole right-hand side of the leapfrog scheme we could also only *modify the linear part*, yielding the variant

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = -\Psi(\tau^2\mathbf{L})\mathbf{q}_n + \tau^2\mathbf{g}_n, \qquad n = 1, 2, \ldots, \tag{3.7}$$

of the above stated two-step scheme. Obviously, for $\mathbf{g} \equiv 0$ the two-step schemes (3.1a) and (3.7) coincide. This variant of the scheme corresponds to the one proposed and analyzed in [CHS20]. In contrast to the paper, we focus on the scheme (3.1a) in this thesis, since one obtains better error bounds. Moreover, numerical experiments indicate a better stability behavior; see the second numerical experiment in Section 4.5.3.

3.1.1. One-step formulations and geometric properties

Before we start with the stability and error analysis of this class of schemes, we have a closer look at one-step formulations of the two-step scheme (3.1a). Additionally, we investigate geometric properties of this scheme.

Lemma 3.5. *Let* $\mathbf{p}_0 = \dot{\mathbf{q}}_0$. *If* $\widehat{\Psi}(\tau^2\mathbf{L})$ *is nonsingular, then the scheme* (3.1) *is equivalent to the one-step scheme*

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \tfrac{1}{2}\tau(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \tag{3.8a}$$
$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\widehat{\Psi}(\tau^2\mathbf{L})\mathbf{p}_{n+1/2}, \tag{3.8b}$$
$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \tfrac{1}{2}\tau(-\mathbf{L}\mathbf{q}_{n+1} + \mathbf{g}_{n+1}), \tag{3.8c}$$

for all $n \in \mathbb{N}_0$.

Proof. We proceed in the same way as for the derivation of the one-step scheme (2.21) of the leapfrog method. In particular, defining

$$\mathbf{p}_{n+1/2} = \tfrac{1}{\tau}\widehat{\Psi}(\tau^2\mathbf{L})^{-1}(\mathbf{q}_{n+1} - \mathbf{q}_n) \qquad \text{and} \qquad \mathbf{p}_{n+1} = \tfrac{1}{2}(\mathbf{p}_{n+3/2} + \mathbf{p}_{n+1/2})$$

for $n \geq 0$ yields the claim. □

Obviously, the scheme (3.1) can be deduced from the one-step scheme (3.8) regardless of whether $\widehat{\Psi}(\tau^2\mathbf{L})$ is singular or not. However, the reverse implication is in general only true if $\widehat{\Psi}(\tau^2\mathbf{L})$ is nonsingular. Otherwise, we could add every vector lying in the kernel of $\widehat{\Psi}(\tau^2\mathbf{L})$ to $\mathbf{p}_{n+1/2}$ or $\mathbf{p}_n$ without changing the approximations $\mathbf{q}_n$. Clearly, the nonsingularity only occurs if $\widehat{\Psi}(z) = 0$ for some $z \geq 0$ and, thus, depends on the step size.

We further emphasize that for starting values $\mathbf{q}_1$, which are different from (3.1b), we can still consider (3.8) for $n \geq 1$ as the one-step scheme belonging to the two-step scheme (3.1a), provided we have an approximation for $\mathbf{p}_1$. Nevertheless, there exist also different one-step formulations. For instance, together with the modified starting value

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2\widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_0 + \mathbf{g}_0),$$

the two-step scheme (3.1a) yields – by proceeding as for the leapfrog scheme in Section 2.3.1 – the *equivalent* one-step scheme

$$\widehat{\mathbf{p}}_{n+1/2} = \widehat{\mathbf{p}}_n + \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \tag{3.9a}$$
$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\widehat{\mathbf{p}}_{n+1/2}, \qquad n = 0, 1, 2, \ldots, \tag{3.9b}$$
$$\widehat{\mathbf{p}}_{n+1} = \widehat{\mathbf{p}}_{n+1/2} + \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_{n+1} + \mathbf{g}_{n+1}), \tag{3.9c}$$

if we set $\widehat{\mathbf{p}}_0 = \dot{\mathbf{q}}_0$. Obviously, we have the relation $\widehat{\mathbf{p}}_n = \widehat{\Psi}(\tau^2\mathbf{L})\mathbf{p}_n$ and the same holds for the half-step approximations.

We now turn towards the geometric properties symmetry and symplecticity. We start with the symmetry of the two-step scheme.

Lemma 3.6. *The two-step scheme* (3.1a) *is symmetric.*

Proof. The claim directly follows from Definition 2.12. □

Next, we investigate the symmetry and (non-)symplecticity of the one-step method (3.8) and its variant (3.9). Recall that we have defined symplecticity only for one-step methods; cf. Section 2.3.2 for the leapfrog scheme. Moreover, to analyze symplectic schemes we require that (2.1) admits a Hamiltonian structure. Thus, we assume that $\mathbf{g}(\cdot, \mathbf{q}) = \mathbf{g}(\mathbf{q})$ and $\mathbf{g}$ satisfies (2.8).

Lemma 3.7. *The one-step scheme (3.8) is symmetric and symplectic.*

Proof. We consider the perturbed second-order differential equation

$$\ddot{\mathbf{q}} = \widehat{\Psi}(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q} + \mathbf{g}(\mathbf{q})) \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \widehat{\Psi}(\tau^2\mathbf{L})\dot{\mathbf{q}}_0,$$

for a fixed, but arbitrary $\tau > 0$. Since $\widehat{\Psi}(\tau^2\mathbf{L})$ is symmetric, it can be written as a Hamiltonian problem (2.9a) with Hamiltonian

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2}(\mathbf{p}, \widehat{\Psi}(\tau^2\mathbf{L})\mathbf{p}) + \tfrac{1}{2}(\mathbf{q}, \mathbf{L}\mathbf{q}) + U(\mathbf{q}).$$

Moreover, application of the leapfrog scheme (2.21) to this modified Hamiltonian problem (with step size τ as in the Hamiltonian) is equivalent to the scheme (3.8). Hence, it inherits the symmetry and symplecticity of the leapfrog method. □

Alternatively, one can prove the symplecticity of the one-step scheme (3.8) by directly verifying condition (2.24) in Definition 2.14. In contrast to the one-step scheme (3.8), we show in the next lemma that the variant (3.9) is in general not symplectic. Nevertheless, we expect that the variant (3.9) shows a similar long-time behavior as the one-step method (3.8) due to the relation $\widehat{\mathbf{p}}_n = \widehat{\Psi}(\tau^2\mathbf{L})\mathbf{p}_n$.

Lemma 3.8. *The one-step scheme (3.9) is symmetric but not symplectic in general.*

Proof. The symmetry follows with the same arguments as in the proof of (3.7).

In order to show that the one-step scheme (3.9) is not symplectic, we calculate the left-hand side of condition (2.24). To shorten the notation we define the function $\mathcal{U}\colon \mathbb{R}^m \to \mathbb{R}$ in such a way that

$$\nabla\mathcal{U}(\mathbf{q}) = \mathbf{L}\mathbf{q} + \nabla U(\mathbf{q}) = \mathbf{L}\mathbf{q} - \mathbf{g}(\mathbf{q}).$$

Using this notation in the one-step scheme (3.9) yields by inserting (3.9a) into (3.9b) and (3.9c)

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\widehat{\mathbf{p}}_n - \tfrac{1}{2}\tau^2\widehat{\boldsymbol{\Psi}}\nabla\boldsymbol{\mathcal{U}}_n \quad \text{and} \quad \widehat{\mathbf{p}}_{n+1} = \widehat{\mathbf{p}}_n - \tfrac{1}{2}\tau\widehat{\boldsymbol{\Psi}}(\nabla\boldsymbol{\mathcal{U}}_{n+1} + \nabla\boldsymbol{\mathcal{U}}_n),$$

where we abbreviate $\boldsymbol{\mathcal{U}}_n = \mathcal{U}(\mathbf{q}_n)$ and $\widehat{\boldsymbol{\Psi}} = \widehat{\Psi}(\tau^2\mathbf{L})$. From this we obtain that the Jacobian of the numerical flow $\Phi_\tau\colon (\widehat{\mathbf{p}}_n, \mathbf{q}_n) \mapsto (\widehat{\mathbf{p}}_{n+1}, \mathbf{q}_{n+1})$ is given by

$$\begin{pmatrix} \mathbf{I} & \frac{1}{2}\tau\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_{n+1} \\ 0 & \mathbf{I} \end{pmatrix} \Phi'_\tau(\widehat{\mathbf{p}}_n, \mathbf{q}_n) = \begin{pmatrix} \mathbf{I} & -\frac{1}{2}\tau\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_n \\ \tau & \mathbf{I} - \frac{1}{2}\tau^2\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_n \end{pmatrix}, \quad \Phi'_\tau(\widehat{\mathbf{p}}_n, \mathbf{q}_n) = \begin{pmatrix} \frac{\partial\widehat{\mathbf{p}}_{n+1}}{\partial\widehat{\mathbf{p}}_n} & \frac{\partial\widehat{\mathbf{p}}_{n+1}}{\partial\mathbf{q}_n} \\ \frac{\partial\mathbf{q}_{n+1}}{\partial\widehat{\mathbf{p}}_n} & \frac{\partial\mathbf{q}_{n+1}}{\partial\mathbf{q}_n} \end{pmatrix},$$

which is equivalent to

$$\Phi'_\tau(\widehat{\mathbf{p}}_n, \mathbf{q}_n) = \begin{pmatrix} \mathbf{I} - \frac{1}{2}\tau^2\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_{n+1}\widehat{\boldsymbol{\Psi}} & -\frac{1}{2}\tau\widehat{\boldsymbol{\Psi}}(\nabla^2\boldsymbol{\mathcal{U}}_n + \nabla^2\boldsymbol{\mathcal{U}}_{n+1} - \frac{1}{2}\tau^2\nabla^2\boldsymbol{\mathcal{U}}_{n+1}\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_n) \\ \tau\mathbf{I} & \mathbf{I} - \frac{1}{2}\tau^2\widehat{\boldsymbol{\Psi}}\nabla^2\boldsymbol{\mathcal{U}}_n \end{pmatrix}.$$

By a simple calculation one then sees that (2.24) is satisfied (note that the Hessian $\nabla^2\boldsymbol{\mathcal{U}}_n$ is a symmetric matrix) if and only if $\widehat{\Psi}(\tau^2\mathbf{L})$ and $\nabla^2 U(\mathbf{q})$ commute. Thus, the one-step scheme (3.9) is in general not symplectic. □

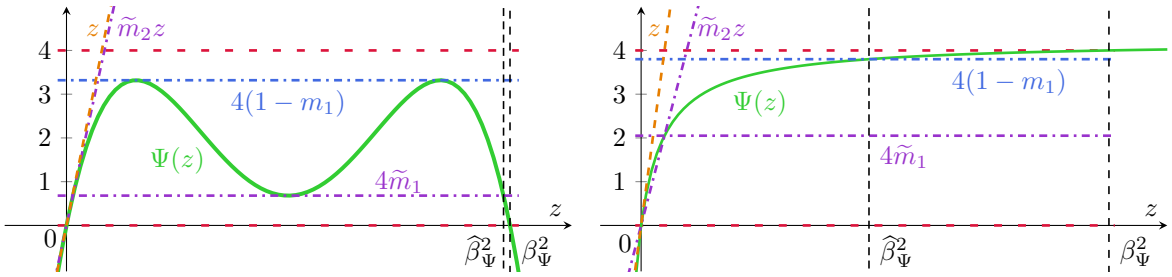


Figure 3.1.: Illustration of Definition 3.9 for the polynomial P_4 with $\nu = 1.03$ (left) given in (3.3) and Ψ_θ with $\theta = 0.2375$ (right) in (3.4).

3.1.2. Properties and assumptions on Ψ

We now state assumptions and definitions on the function $\widehat{\Psi}$, which are necessary for the subsequent stability and error analysis not only in this chapter but also in Chapter 5. Because it is more convenient to work with the function Ψ instead of $\widehat{\Psi}$, most of the subsequent definitions rely on Ψ. We point out that for $\widehat{\Psi}$ sufficiently smooth the condition $\widehat{\Psi}(0) = 1$ in Assumption 3.2 is equivalent to the *consistency conditions*

$$\Psi(0) = 0, \qquad \Psi'(0) = 1. \tag{3.10}$$

We start with two definitions bounding Ψ from above and below. These definitions are used later on to define step-size restrictions required for stability of the scheme (3.1).

Definition 3.9. *Let $\widehat{\Psi}$ satisfy Assumption 3.2.*

(a) *We define $\beta_\Psi > 0$ as the maximum value such that*

$$0 \le \Psi(z) \le 4 \qquad \text{for all } z \in [0, \beta_\Psi^2], \tag{3.11}$$

and $\beta_\Psi = \infty$ if (3.11) holds for all $z \ge 0$.

(b) *For given $m_1, \widetilde{m}_1, \widetilde{m}_2 \in (0,1)$ with $\widetilde{m}_1 \le 1 - m_1$ we define $\widehat{\beta}_\Psi = \widehat{\beta}_\Psi(m_1, \widetilde{m}_1, \widetilde{m}_2) > 0$ as the maximum value such that*

$$\min\{4\widetilde{m}_1, \widetilde{m}_2 z\} \le \Psi(z) \le 4(1-m_1) \qquad \text{for all } z \in [0, \widehat{\beta}_\Psi^2], \tag{3.12}$$

and $\widehat{\beta}_\Psi = \infty$ if (3.12) holds for all $z \ge 0$.

The consistency conditions (3.10) guarantee the existence of $\beta_\Psi, \widehat{\beta}_\Psi > 0$ for every $m_1, \widetilde{m}_1, \widetilde{m}_2$ because $\Psi(z) = z + \mathcal{O}(z^2)$ for small z. Obviously, the definitions imply that $\widehat{\beta}_\Psi \le \beta_\Psi$, where equality only occurs for $\beta_\Psi = \widehat{\beta}_\Psi = \infty$. More precisely, the ratio between $\widehat{\beta}_\Psi$ and β_Ψ strongly depends on the function Ψ and the choice of $m_1, \widetilde{m}_1$, and $\widetilde{m}_2$. In Figure 3.1, the definitions are illustrated for two different functions Ψ.

In the next section we see that it is desirable to have $\widehat{\beta}_\Psi \approx \beta_\Psi$. This can be achieved by choosing Ψ and the constants m_1, $\widetilde{m}_1$, and $\widetilde{m}_2$ appropriately. For the LFC polynomials (3.3) we give explicit values for these constants in the next chapter.

Remark 3.10. In the case of a finite $\widehat{\beta}_\Psi$, the lower bound in (3.12) implies

$$\Psi(z) \ge m_2 z \qquad \text{for all } z \in [0, \widehat{\beta}_\Psi^2] \tag{3.13}$$

with $m_2 = \min\{\widetilde{m}_2, 4\widetilde{m}_1/\widehat{\beta}_\Psi^2\} > 0$. In principle, if $\widehat{\beta}_\Psi < \infty$, we could perform the following stability and error analysis also with (3.13) with $m_2 \in (0,1)$. In fact, in [CHS20] all calculations are done with this condition instead of the condition for the lower bound in (3.12). However, since m_2 tends to zero for $\widehat{\beta}_\Psi \to \infty$, this approach is unsuitable for $\widehat{\beta}_\Psi = \infty$. For the polynomials (3.3) another drawback of condition (3.13) is pointed out in Remark 4.10. ◇

Example 3.11 (LF). For the leapfrog scheme (2.20) we have $\Psi(z) = z$. Obviously, the lower bound in (3.12) holds for every $\widetilde{m}_1, \widetilde{m}_2 \in (0,1)$. We even could choose $\widetilde{m}_1 = \widetilde{m}_2 = 1$ and $m_2 = 1$ to satisfy the first bound in (3.12) and (3.13), respectively. For the upper bound in (3.12) let $m_1 = 1 - \vartheta^2$ for some $\vartheta \in (0,1)$. This choice then yields $\widehat{\beta}_{\mathrm{LF}}^2 = 4\vartheta^2 < 4 = \beta_{\mathrm{LF}}^2$. ◇

Further, for our (error) analysis we need bounds of the form $|\widehat{\Psi}(z) - 1| \le cz$ for $z \in [0, \beta_\Psi^2] \cap \mathbb{R}$ and a constant $c > 0$. For this, we first define another function based on $\widehat{\Psi}$.

Definition 3.12. *Let $\widehat{\Psi}$ satisfy Assumption 3.2. We define $\Upsilon\colon [0,\infty) \to \mathbb{R}$ as the continuous extension of*

$$\Upsilon\colon (0,\infty) \to \mathbb{R}, \quad \Upsilon(z) = \frac{\widehat{\Psi}(z) - 1}{z} = \frac{\Psi(z) - z}{z^2}. \tag{3.14}$$

It is easy to see that Υ is again a smooth function, where $\Upsilon(0) = \widehat{\Psi}'(0) = \frac{1}{2}\Psi''(0)$. Moreover, if $\widehat{\Psi}$ is a polynomial of degree $p \ge 1$, then Υ and Ψ are polynomials of degree $p-1$ and $p+1$, respectively, due to Assumption 3.2.

Definition 3.13. *Let $\widehat{\Psi}$ satisfy Assumption 3.2. We define $m_3 \ge 0$ as the smallest constant such that*

$$|\Upsilon(z)| \le \tfrac{1}{2} m_3 \qquad \text{for all } z \in [0, \beta_\Psi^2] \cap \mathbb{R}. \tag{3.15}$$

The factor $\frac{1}{2}$ is only for convenience and is motivated by the fact that $\Upsilon(0) = \frac{1}{2}\Psi''(0)$. The existence of m_3 follows again from Assumption 3.2 and the definition of β_Ψ.

Additionally to the estimate (3.15) for Υ the subsequent error analysis of the two-step scheme (3.1a) requires another estimate involving the function Υ. To state this estimate we first point out that from the lower bound in (3.12) we obtain

$$\widehat{\Psi}(z) \ge \min\{\widetilde{m}_2, 4\widetilde{m}_1/z\} > 0 \qquad \text{for all } z \in [0, \widehat{\beta}_\Psi^2] \cap \mathbb{R}. \tag{3.16}$$

Thus, $1/\widehat{\Psi}(z)$ exists for all $z \in [0, \widehat{\beta}_\Psi^2] \cap \mathbb{R}$ and the following bound holds.

Lemma 3.14. *Let $\widehat{\Psi}$ satisfy Assumption 3.2. There exists a constant $\widetilde{m}_3 \ge 0$ such that*

$$|\widehat{\Psi}(z)^{-1}\Upsilon(z)| \le \widetilde{m}_3 \qquad \text{for all } z \in [0, \widehat{\beta}_\Psi^2] \cap \mathbb{R}. \tag{3.17}$$

Proof. Definition (3.14) of Υ and Definition 3.3 for Ψ yield for $z > 0$

$$\widehat{\Psi}(z)^{-1}\Upsilon(z) = \frac{1}{z} - \frac{1}{\Psi(z)}.$$

Thus, for every $\varepsilon > 0$ we have that $|\widehat{\Psi}(z)^{-1}\Upsilon(z)|$ is bounded for all $z \in (\varepsilon, \widehat{\beta}_\Psi^2] \cap \mathbb{R}$, since $\Psi(z) \ge \delta$ for $z \in (\varepsilon, \widehat{\beta}_\Psi^2] \cap \mathbb{R}$ and some $\delta > 0$. Using $\widehat{\Psi}(0)^{-1}\Upsilon(0) = \frac{1}{2}\Psi''(0)$ and the continuity of the functions completes the proof. □

Example 3.15 (LF continued). For the leapfrog scheme (2.20) we have $\Upsilon \equiv 0$ and, thus, $m_3 = 0$. Moreover, Lemma 3.14 holds for every $\widetilde{m}_3 \ge 0$. ◇

Finally, to simplify the presentation we assume that additionally to (3.10) we have the following condition for the function $\widehat{\Psi}$ or, equivalently, for Ψ.

Assumption 3.16. *The function Ψ satisfies*

$$\Psi(z) \leq z \qquad \text{for all } z \in \left[0, \beta_\Psi^2\right] \cap \mathbb{R}. \tag{3.18}$$

We emphasize that (3.18) is fulfilled for the functions P_p and Ψ_θ defined in (3.3) and (3.4), respectively; see Lemma 4.3 and Lemma 3.61. Clearly, from the assumption we immediately obtain with Definitions 3.3, 3.12, and 3.13 for Ψ and Υ, respectively, that

$$0 \leq \widehat{\Psi}(z) \leq 1, \quad -\tfrac{1}{2} m_3 \leq \Upsilon(z) \leq 0 \qquad \text{for all } z \in \left[0, \beta_\Psi^2\right] \cap \mathbb{R}. \tag{3.19}$$

We further note that this assumption could theoretically be dropped, since Assumption 3.2 and Definition 3.9 yield the existence of a constant $c_* \geq 1$ such that $\Psi(z) \leq c_* z$ for all $z \in \left[0, \beta_\Psi^2\right] \cap \mathbb{R}$. The stability and error bounds in the following would still be valid, however, with possibly additional or larger constants.

In the remaining part of this chapter let Assumptions 3.2 and 3.16 hold without mentioning it explicitly everywhere. Moreover, we abbreviate

$$\boldsymbol{\Psi} = \Psi(\tau^2 \mathbf{L}), \quad \widehat{\boldsymbol{\Psi}} = \widehat{\Psi}(\tau^2 \mathbf{L}), \quad \boldsymbol{\Upsilon} = \Upsilon(\tau^2 \mathbf{L}),$$

and so forth.

3.2. Representation formulae for numerical approximations

For analyzing the stability of the two-step scheme (3.1a), we derive a formula for the numerical approximations $\mathbf{q}_n$ in dependence of the starting values $\mathbf{q}_0$, $\mathbf{q}_1$, and the right-hand side $\mathbf{g}_1, \dots, \mathbf{g}_{n-1}$. Exploiting the starting value (3.1b) then enables us to write $\mathbf{q}_n$ in terms of the data $\mathbf{q}_0, \dot{\mathbf{q}}_0$, and $\mathbf{g}$, if $\mathbf{g}$ only depends on time.

Before we start with the derivation of such a representation formula we first introduce a discrete analogue to the energy norm (2.14), since the two-step scheme (3.1a) does not directly permit an approximation to $\dot{\mathbf{q}}$. By using the centered difference quotient we define

$$\dot{\mathbf{q}}(t_n) \approx \partial_\tau \mathbf{q}_n = \begin{cases} \frac{1}{2\tau}(\mathbf{q}_{n+1} - \mathbf{q}_{n-1}), & n \geq 1, \\ \widehat{\boldsymbol{\Psi}} \dot{\mathbf{q}}_0, & n = 0. \end{cases} \tag{3.20a}$$

With this definition at hand we define the *discrete energy norm* via

$$|||\mathbf{q}_n|||_\tau^2 = \|\partial_\tau \mathbf{q}_n\|^2 + \|\mathbf{q}_n\|_{\mathbf{L}}^2. \tag{3.20b}$$

Like its continuous version, the discrete energy norm is in general only a seminorm, as long as $\mathbf{L}$ is not positive definite. It is worth mentioning that $\partial_\tau \mathbf{q}_n$ is closely related to the approximation $\mathbf{p}_n$ of the one-step scheme (3.8) by

$$\partial_\tau \mathbf{q}_n = \frac{1}{2\tau}(\mathbf{q}_{n+1} - \mathbf{q}_{n-1}) = \widehat{\boldsymbol{\Psi}} \mathbf{p}_n, \tag{3.21}$$

which follows from (3.8b) and $\mathbf{p}_{n+1/2} + \mathbf{p}_{n-1/2} = 2\mathbf{p}_n$.

Further, for the derivation of the representation formula a restriction for the step size τ is required such that all terms are well-defined.

Definition 3.17. *For β_Ψ defined in Definition 3.9(a) we define $\tau_{\mathrm{SSR}} > 0$ via*

$$\tau_{\mathrm{SSR}}^2 = \begin{cases} \beta_\Psi^2/\|\mathbf{L}\| & \text{if } \beta_\Psi < \infty, \\ \infty & \text{if } \beta_\Psi = \infty. \end{cases} \tag{3.22}$$

Clearly, if $\beta_\Psi = \infty$ the condition $\tau \le \tau_{\mathrm{SSR}}$ is not a restriction at all. The necessity of the condition $\tau \le \tau_{\mathrm{SSR}}$ for ensuring stability of the scheme (3.1a) for linear problems (2.12) has already been shown in [GJ08, JR10] for polynomials Ψ and in [CI17] for rational functions Ψ. In the first two papers the stability is proven via the characteristic equation of linear difference equations and the root criterion. In fact, if we apply (3.1) to the test equation (2.25), it can be shown similarly to Lemma 2.16 that the step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ is necessary for stability. In [CI17] modified energy techniques are used to prove stability in the standard norm $\|\cdot\|$.

3.2.1. Proof via generating functions

With the step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ at hand we now derive a representation formula of the scheme (3.1). We start with the two-step scheme (3.1a); cf. [CHS20, Theorem 3.3].

Theorem 3.18. *Let $\tau \le \tau_{\mathrm{SSR}}$. For the approximations of the two-step scheme (3.1a) we have*

$$\mathbf{q}_n = \cos(n\mathbf{\Phi})\mathbf{q}_0 + \boldsymbol{\mathcal{S}}_n(\mathbf{q}_1 - \cos\mathbf{\Phi}\,\mathbf{q}_0) + \tau^2 \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell}\widehat{\mathbf{\Psi}}\mathbf{g}_\ell, \quad \boldsymbol{\mathcal{S}}_k = \frac{\sin(k\mathbf{\Phi})}{\sin\mathbf{\Phi}}\ (k \in \mathbb{N}_0), \tag{3.23a}$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times m}$ is a symmetric matrix with spectrum in $[0,\pi]$ which is uniquely defined by

$$\cos\mathbf{\Phi} = \mathbf{I} - \tfrac{1}{2}\mathbf{\Psi} \qquad \text{and satisfies} \qquad \sin\mathbf{\Phi} = \big(\mathbf{\Psi}(\mathbf{I} - \tfrac{1}{4}\mathbf{\Psi})\big)^{1/2}. \tag{3.23b}$$

It is worth mentioning that with the definition of $\mathbf{\Phi}$ we can rewrite the two-step scheme (3.1a) in a way similar to a trigonometric integrator

$$\mathbf{q}_{n+1} - 2\cos\mathbf{\Phi}\,\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2\widehat{\mathbf{\Psi}}\mathbf{g}_n = \tau^2(\tfrac{1}{2}\tau\mathbf{L}^{1/2})^{-2}\sin\big(\tfrac{1}{2}\mathbf{\Phi}\big)^2\mathbf{g}_n. \tag{3.24}$$

The second identity obviously only holds for $\mathbf{L}$ positive definite, where we used $\mathbf{\Psi} = 4\sin\big(\frac{1}{2}\mathbf{\Phi}\big)^2$, following from (B.2b). In fact, for the special case of $\Psi = \Psi_{\mathrm{trig}}$ defined in (3.6) we actually have with (B.2b)

$$\cos\mathbf{\Phi} = \mathbf{I} - \tfrac{1}{2}\big(2\mathbf{I} - 2\cos(\tau\mathbf{L}^{1/2})\big) = \cos(\tau\mathbf{L}^{1/2}) \qquad \text{and} \qquad \sin\big(\tfrac{1}{2}\mathbf{\Phi}\big)^2 = \sin\big(\tfrac{1}{2}\tau\mathbf{L}^{1/2}\big)^2,$$

which yields Gautschi's trigonometric integrator (3.5).

Remark 3.19. In [HLW06, Section XIII.8] a similar notation is used to analyze geometric properties of the leapfrog method (2.20) applied to Hamiltonian problems. By considering the leapfrog scheme as a perturbed trigonometric integrator, the authors transfer their previously obtained results for trigonometric integrators to the leapfrog scheme. ◇

Proof of Theorem 3.18. In order to prove this result we apply the *generating functions* technique; see Remark 3.1. Hence, we multiply the recursion (3.1a) by ζ^{n+1}, $\zeta \in \mathbb{C}$, and sum over $n \ge 1$

$$\sum_{n=1}^{\infty}(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1})\zeta^{n+1} = -\mathbf{\Psi}\sum_{n=1}^{\infty}\mathbf{q}_n\zeta^{n+1} + \tau^2\widehat{\mathbf{\Psi}}\sum_{n=1}^{\infty}\mathbf{g}_n\zeta^{n+1}.$$

Defining the formal power series (generating functions)

$$\mathbf{q}(\zeta) = \sum_{n=0}^{\infty} \mathbf{q}_n \zeta^n \qquad \text{and} \qquad \mathbf{g}(\zeta) = \sum_{n=0}^{\infty} \mathbf{g}_n \zeta^n$$

yields

$$(1 - 2\zeta + \zeta^2)\,\mathbf{q}(\zeta) - \zeta\mathbf{q}_1 - \mathbf{q}_0 + 2\zeta\mathbf{q}_0 = -\zeta\boldsymbol{\Psi}\,\mathbf{q}(\zeta) + \zeta\boldsymbol{\Psi}\,\mathbf{q}_0 + \tau^2\widehat{\boldsymbol{\Psi}}\zeta(\mathbf{g}(\zeta) - \mathbf{g}_0).$$

This is equivalent to

$$\boldsymbol{\varrho}(\zeta)\mathbf{q}(\zeta) = \mathbf{q}_0 + \zeta\mathbf{q}_1 - \zeta(2\mathbf{I} - \boldsymbol{\Psi})\mathbf{q}_0 + \tau^2\widehat{\boldsymbol{\Psi}}\zeta(\mathbf{g}(\zeta) - \mathbf{g}_0) \tag{3.25a}$$

with

$$\boldsymbol{\varrho}(\zeta) = \zeta^2\mathbf{I} - \zeta(2\mathbf{I} - \boldsymbol{\Psi}) + \mathbf{I}. \tag{3.25b}$$

Using the symmetry of $\boldsymbol{\Psi}$ and the step-size restriction $\tau \leq \tau_{\mathrm{SSR}}$ we have by (3.11) that the matrix-valued roots $\boldsymbol{\zeta}_\pm$ of $\boldsymbol{\varrho}$ are given by

$$\boldsymbol{\zeta}_\pm = \mathbf{I} - \tfrac{1}{2}\boldsymbol{\Psi} \pm \mathrm{i}\big(\boldsymbol{\Psi}(\mathbf{I} - \tfrac{1}{4}\boldsymbol{\Psi})\big)^{1/2}.$$

Moreover, since the spectrum of $\mathbf{I} - \frac{1}{2}\boldsymbol{\Psi}$ is contained in $[-1, 1]$ due to $\tau \leq \tau_{\mathrm{SSR}}$, there exists a uniquely defined symmetric matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m\times m}$ satisfying (3.23b) whose spectrum is a subset of $[0, \pi]$. From this we obtain with $\sin x = (1 - (\cos x)^2)^{1/2}$ for $x \in [0, \pi]$ that

$$\sin\boldsymbol{\Phi} = \big(\mathbf{I} - (\mathbf{I} - \tfrac{1}{2}\boldsymbol{\Psi})^2\big)^{1/2} = \big(\boldsymbol{\Psi}(\mathbf{I} - \tfrac{1}{4}\boldsymbol{\Psi})\big)^{1/2},$$

which shows (3.23b). Together this leads to

$$\boldsymbol{\zeta}_\pm = \cos\boldsymbol{\Phi} \pm \mathrm{i}\sin\boldsymbol{\Phi} = \mathrm{e}^{\pm\mathrm{i}\boldsymbol{\Phi}},$$

and $\boldsymbol{\zeta}_+ = \boldsymbol{\zeta}_-^{-1}$.

By exploiting these relations we have

$$\boldsymbol{\varrho}(\zeta) = (\zeta\mathbf{I} - \boldsymbol{\zeta}_+)(\zeta\mathbf{I} - \boldsymbol{\zeta}_-) = (\zeta\boldsymbol{\zeta}_- - \mathbf{I})(\zeta\boldsymbol{\zeta}_+ - \mathbf{I}) = (\mathbf{I} - \zeta\mathrm{e}^{-\mathrm{i}\boldsymbol{\Phi}})(\mathbf{I} - \zeta\mathrm{e}^{\mathrm{i}\boldsymbol{\Phi}}).$$

In particular, $\boldsymbol{\varrho}$ is nonsingular for every $|\zeta| < 1$, $\zeta \in \mathbb{C}$. Employing the Neumann series for $|\zeta| < 1$ and the Cauchy product yields

$$\begin{aligned}\boldsymbol{\varrho}(\zeta)^{-1} &= \left(\sum_{n=0}^{\infty} \mathrm{e}^{-\mathrm{i}n\boldsymbol{\Phi}}\zeta^n\right)\left(\sum_{n=0}^{\infty} \mathrm{e}^{\mathrm{i}n\boldsymbol{\Phi}}\zeta^n\right) = \sum_{n=0}^{\infty}\sum_{\ell=0}^{n}\Big(\mathrm{e}^{-\mathrm{i}(n-\ell)\boldsymbol{\Phi}}\zeta^{n-\ell}\Big)\Big(\mathrm{e}^{\mathrm{i}\ell\boldsymbol{\Phi}}\zeta^{\ell}\Big)\\ &= \sum_{n=0}^{\infty} \mathrm{e}^{-\mathrm{i}n\boldsymbol{\Phi}}\zeta^n \sum_{\ell=0}^{n} \mathrm{e}^{2\mathrm{i}\ell\boldsymbol{\Phi}} = \sum_{n=0}^{\infty} \boldsymbol{\mathcal{S}}_{n+1}\zeta^n,\end{aligned}$$

where the last equality follows from

$$\mathrm{e}^{-\mathrm{i}n\boldsymbol{\Phi}}\sum_{\ell=0}^{n} \mathrm{e}^{2\mathrm{i}\ell\boldsymbol{\Phi}} = \mathrm{e}^{-\mathrm{i}n\boldsymbol{\Phi}}\,\frac{\mathbf{I} - \mathrm{e}^{\mathrm{i}2(n+1)\boldsymbol{\Phi}}}{\mathbf{I} - \mathrm{e}^{\mathrm{i}2\boldsymbol{\Phi}}} = \frac{\mathrm{e}^{-\mathrm{i}(n+1)\boldsymbol{\Phi}} - \mathrm{e}^{\mathrm{i}(n+1)\boldsymbol{\Phi}}}{\mathrm{e}^{-\mathrm{i}\boldsymbol{\Phi}} - \mathrm{e}^{\mathrm{i}\boldsymbol{\Phi}}} = \frac{\sin\big((n+1)\boldsymbol{\Phi}\big)}{\sin\boldsymbol{\Phi}} = \boldsymbol{\mathcal{S}}_{n+1}.$$

Hence, by using the formula for $\boldsymbol{\varrho}^{-1}(\zeta)$ in (3.25) we obtain again with the Cauchy product

$$\begin{aligned}\mathbf{q}(\zeta) &= \sum_{n=0}^{\infty} \boldsymbol{\mathcal{S}}_{n+1}\Big(\mathbf{q}_0\,\zeta^n + \mathbf{q}_1\,\zeta^{n+1} - (2\mathbf{I}-\boldsymbol{\Psi})\mathbf{q}_0\,\zeta^{n+1} + \tau^2\widehat{\boldsymbol{\Psi}}(\mathbf{g}(\zeta)-\mathbf{g}_0)\zeta^{n+1}\Big)\\ &= \sum_{n=0}^{\infty} \boldsymbol{\mathcal{S}}_{n+1}\Big(\mathbf{q}_0\,\zeta^n + (\mathbf{q}_1 - 2\cos\boldsymbol{\Phi}\,\mathbf{q}_0)\,\zeta^{n+1}\Big) + \sum_{n=0}^{\infty}\left(\tau^2\sum_{\ell=1}^{n}\boldsymbol{\mathcal{S}}_{n+1-\ell}\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell\right)\zeta^{n+1}.\end{aligned}$$

From this we deduce by comparing the coefficients of ζ^n

$$\mathbf{q}_n = \boldsymbol{\mathcal{S}}_{n+1}\mathbf{q}_0 + \boldsymbol{\mathcal{S}}_n(\mathbf{q}_1 - 2\cos\boldsymbol{\Phi}\,\mathbf{q}_0) + \tau^2\sum_{\ell=1}^{n-1}\boldsymbol{\mathcal{S}}_{n-\ell}\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell.$$

The angle sum identity (B.1a) applied to $\sin((n+1)\boldsymbol{\Phi})$ yields $\boldsymbol{\mathcal{S}}_{n+1} = \cos(n\boldsymbol{\Phi}) + \boldsymbol{\mathcal{S}}_n\cos\boldsymbol{\Phi}$, which completes the proof. □

So far, the starting value (3.1b) is not incorporated in the representation formula (3.23a). In order to state such a formula we define $\xi_{\ell,n}$, $\ell = 0,\dots,n$, via

$$\xi_{0,0} = 0 \qquad \text{and} \qquad \xi_{\ell,n} = \begin{cases}\frac{1}{2}, & \ell \in \{0,n\},\\ 1, & \ell \in \{1,\dots,n-1\},\end{cases} \qquad n \geq 1. \tag{3.26}$$

which we also use later on.

Corollary 3.20. *Let $\tau \leq \tau_{\mathrm{SSR}}$. For the approximations of the scheme (3.1) we have*

$$\mathbf{q}_n = \cos(n\boldsymbol{\Phi})\mathbf{q}_0 + \tau\boldsymbol{\mathcal{S}}_n\widehat{\boldsymbol{\Psi}}\dot{\mathbf{q}}_0 + \tau^2\sum_{\ell=0}^{n-1}\xi_{\ell,n}\,\boldsymbol{\mathcal{S}}_{n-\ell}\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell. \tag{3.27}$$

Proof. Inserting (3.1b) in (3.23a) and using (3.23b) yields the result. □

For the stability and error analysis we further need a representation formula for the quantity $\partial_\tau\mathbf{q}_n$ arising in the discrete energy norm (3.20b). The formula can be concluded by exploiting the previous results.

Corollary 3.21. *Let $\tau \leq \tau_{\mathrm{SSR}}$. For the approximations $\mathbf{q}_n$ of the two-step scheme (3.1a) we have for $n \geq 1$*

$$\tau\,\partial_\tau\mathbf{q}_n = -\sin\boldsymbol{\Phi}\sin(n\boldsymbol{\Phi})\mathbf{q}_0 + \cos(n\boldsymbol{\Phi})(\mathbf{q}_1 - \cos\boldsymbol{\Phi}\,\mathbf{q}_0) + \tau^2\sum_{\ell=1}^{n}\xi_{\ell,n}\cos((n-\ell)\boldsymbol{\Phi})\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell, \tag{3.28}$$

and for the approximations of the full scheme (3.1)

$$\tau\,\partial_\tau\mathbf{q}_n = -\sin\boldsymbol{\Phi}\sin(n\boldsymbol{\Phi})\mathbf{q}_0 + \tau\cos(n\boldsymbol{\Phi})\widehat{\boldsymbol{\Psi}}\dot{\mathbf{q}}_0 + \tau^2\sum_{\ell=0}^{n}\xi_{\ell,n}\cos((n-\ell)\boldsymbol{\Phi})\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell, \tag{3.29}$$

where $\partial_\tau\mathbf{q}_n$ is defined in (3.20a) and the coefficients $\xi_{\ell,n}$, $0 \leq \ell \leq n$, in (3.26).

Proof. The formulae follows from the definition of $\partial_\tau\mathbf{q}_n$, Theorem 3.18, Corollary 3.20, and the sum-to-product formulae for sine (B.4a) and cosine (B.4c). □

3.2.2. Alternative proof via one-step formulation

We next show an alternative proof of Corollary 3.20, which is based on the one-step formulation (3.8) of the scheme (3.1). In contrast to the proof of Theorem 3.18, we do not use the technique of generating functions, but instead employ a block representation of the one-step scheme. As a byproduct we also obtain a representation formula for $\mathbf{p}_n$.

In the proof we make use of the relations

$$\sin(\tfrac{1}{2}\mathbf{\Phi}) = \tfrac{1}{2}\mathbf{\Psi}^{1/2}, \qquad \cos(\tfrac{1}{2}\mathbf{\Phi}) = (\mathbf{I} - \tfrac{1}{4}\mathbf{\Psi})^{1/2}, \tag{3.30}$$

which follow from (3.23b) and the half-angle formulae (B.3), since the spectrum of $\mathbf{\Phi}$ is in $[0, \pi]$.

Alternative proof of Corollary 3.20. We first rewrite the one-step formulation (3.8) such that the approximations $\mathbf{p}_{n+1/2}$ at half steps are eliminated; cf., for example, [Ver11, HS16, HL21], where the same approach is used. Subtraction and addition of (3.8a) and (3.8c) yields

$$\begin{aligned}\mathbf{p}_{n+1/2} &= \tfrac{1}{2}(\mathbf{p}_n + \mathbf{p}_{n+1}) + \tfrac{1}{4}\tau(\mathbf{L}\mathbf{q}_{n+1} - \mathbf{L}\mathbf{q}_n - \mathbf{g}_{n+1} + \mathbf{g}_n),\\ \mathbf{p}_{n+1} &= \mathbf{p}_n + \tfrac{1}{2}\tau(-\mathbf{L}\mathbf{q}_{n+1} - \mathbf{L}\mathbf{q}_n + \mathbf{g}_{n+1} + \mathbf{g}_n).\end{aligned}$$

By inserting the first equation into (3.8b) we then obtain in a block notation

$$\boldsymbol{\mathcal{R}}_- \mathbf{u}_{n+1} = \boldsymbol{\mathcal{R}}_+ \mathbf{u}_n + \tfrac{1}{2}\tau \begin{pmatrix} \frac{1}{2}\tau\widehat{\mathbf{\Psi}}(\mathbf{g}_n - \mathbf{g}_{n+1}) \\ \mathbf{g}_n + \mathbf{g}_{n+1} \end{pmatrix}, \tag{3.31a}$$

where we abbreviate

$$\mathbf{u}_n = \begin{pmatrix} \mathbf{q}_n \\ \mathbf{p}_n \end{pmatrix}, \qquad \boldsymbol{\mathcal{R}}_\pm = \begin{pmatrix} \mathbf{I} - \frac{1}{4}\mathbf{\Psi} & \pm\frac{1}{2}\tau\widehat{\mathbf{\Psi}} \\ \mp\frac{1}{2}\tau\mathbf{L} & \mathbf{I} \end{pmatrix}. \tag{3.31b}$$

By exploiting the structure of $\boldsymbol{\mathcal{R}}_\pm$ we can rewrite (3.31a) as

$$\boldsymbol{\mathcal{R}}_- \mathbf{u}_{n+1} = \boldsymbol{\mathcal{R}}_+ \mathbf{u}_n + \tfrac{1}{2}\tau(\boldsymbol{\mathcal{R}}_- \widehat{\mathbf{f}}_{n+1} + \boldsymbol{\mathcal{R}}_+ \widehat{\mathbf{f}}_n), \qquad \widehat{\mathbf{f}}_n = \begin{pmatrix} 0 \\ \mathbf{g}_n \end{pmatrix}.$$

Further, the inverse of $\boldsymbol{\mathcal{R}}_-$ can be computed explicitly for every $\tau \geq 0$, which gives

$$\boldsymbol{\mathcal{R}}_-^{-1} = \begin{pmatrix} \mathbf{I} & \frac{1}{2}\tau\widehat{\mathbf{\Psi}} \\ -\frac{1}{2}\tau\mathbf{L} & \mathbf{I} - \frac{1}{4}\mathbf{\Psi} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{R}}_-^{-1}\boldsymbol{\mathcal{R}}_+ = \begin{pmatrix} \mathbf{I} - \frac{1}{2}\mathbf{\Psi} & \tau\widehat{\mathbf{\Psi}} \\ -\tau\mathbf{L}(\mathbf{I} - \frac{1}{4}\mathbf{\Psi}) & \mathbf{I} - \frac{1}{2}\mathbf{\Psi} \end{pmatrix}.$$

Altogether, this yields for (3.31)

$$\mathbf{u}_{n+1} = \boldsymbol{\mathcal{R}}\mathbf{u}_n + \tfrac{1}{2}\tau(\boldsymbol{\mathcal{R}}\widehat{\mathbf{f}}_n + \widehat{\mathbf{f}}_{n+1}),$$

from which we obtain the discrete variation-of-constants formula

$$\mathbf{u}_n = \boldsymbol{\mathcal{R}}^n\mathbf{u}_0 + \tfrac{1}{2}\tau\sum_{\ell=0}^{n-1}\boldsymbol{\mathcal{R}}^{n-1-\ell}(\boldsymbol{\mathcal{R}}\widehat{\mathbf{f}}_\ell + \widehat{\mathbf{f}}_{\ell+1}) = \boldsymbol{\mathcal{R}}^n\mathbf{u}_0 + \tau\sum_{\ell=0}^{n}\xi_{\ell,n}\boldsymbol{\mathcal{R}}^{n-\ell}\widehat{\mathbf{f}}_\ell \tag{3.32}$$

with $\xi_{\ell,n}$ defined in (3.26). Hence, we have derived a formula for the numerical solution $\mathbf{u}_n$ of the scheme (3.8) depending on the initial values $\mathbf{q}_0, \dot{\mathbf{q}}_0$, and $\mathbf{g}_0, \mathbf{g}_1, \dots, \mathbf{g}_{n-1}$.

We note that up to this point we have not exploited the step-size restriction $\tau \leq \tau_{\mathrm{SSR}}$. By using this together with (3.23b) and (3.30) we are able to write $\boldsymbol{\mathcal{R}}$ as

$$\boldsymbol{\mathcal{R}} = \begin{pmatrix} \cos \boldsymbol{\Phi} & \tau \boldsymbol{\mathcal{S}}_1 \widehat{\boldsymbol{\Psi}} \\ -\tau \mathbf{L}\, \boldsymbol{\mathcal{S}}_1 \cos\left(\frac{1}{2}\boldsymbol{\Phi}\right)^2 & \cos \boldsymbol{\Phi} \end{pmatrix},$$

since $\boldsymbol{\mathcal{S}}_1 = \mathbf{I}$. With an induction argument one then obtains together with (3.30), the angle sum identities (B.1), and the double-angle formula (B.2a)

$$\boldsymbol{\mathcal{R}}^n = \begin{pmatrix} \cos(n\boldsymbol{\Phi}) & \tau \boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}} \\ -\tau \mathbf{L}\, \boldsymbol{\mathcal{S}}_n \cos\left(\frac{1}{2}\boldsymbol{\Phi}\right)^2 & \cos(n\boldsymbol{\Phi}) \end{pmatrix}.$$

Inserting this into (3.32) yields for the first block row the representation formula (3.27) for $\mathbf{q}_n$ (note that $\widehat{\mathbf{f}}_\ell$ is 0 in the first block component). □

By considering the second block row in (3.32) we obtain for the quantity $\mathbf{p}_n$ from the one-step scheme (3.8)

$$\mathbf{p}_n = -\tau \boldsymbol{\mathcal{S}}_n \cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2 \mathbf{L}\mathbf{q}_0 + \cos(n\boldsymbol{\Phi})\dot{\mathbf{q}}_0 + \tau \sum_{\ell=0}^{n} \xi_{\ell,n} \cos\big((n-\ell)\boldsymbol{\Phi}\big)\mathbf{g}_\ell.$$

We point out that, if we multiply this equation by $\tau\widehat{\boldsymbol{\Psi}}$, we again obtain formula (3.29) for $\partial_\tau \mathbf{q}_n$ by using (3.30) and (B.2a). This is in accordance with (3.21), where we have shown $\widehat{\boldsymbol{\Psi}}\mathbf{p}_n = \partial_\tau \mathbf{q}_n$.

We further emphasize that from the block formula (3.31) one also easily observes the symmetry of the one-step scheme (3.8) because an exchange of $(\mathbf{u}_{n+1}, t_{n+1})$ with $(\mathbf{u}_n, t_n)$ and τ with $-\tau$ leads again to the same formula; see Lemma 3.7 for another proof. The symmetry of the one-step scheme (3.9) can be shown analogously by deriving a a similar block formula.

3.3. Stability and long-time behavior

After we have derived representation formulae for the numerical approximation $\mathbf{q}_n$ as well as for some related quantities, we are now able to analyze under which assumptions the scheme (3.1a) yields stable approximations to the solution of (2.1).

We restrict ourselves to linear problems in this section. The reason for this restriction is on the one hand that we can state stability bounds in the semilinear case for locally Lipschitz continuous functions $\mathbf{g}$ only together with the error analysis. On the other hand, even for globally Lipschitz continuous functions we gain no new insight into the stability behavior of these schemes.

3.3.1. Stability for linear problems

We start by considering the stability behavior of the scheme (3.1) if it is applied to linear differential equations of the form (2.12). To do so, we first show some bounds for norms of matrix functions occurring in the representation formulae and, hence, in the stability and error analysis. For some of these bounds we require a stronger step-size restriction than the one given through Definition 3.17.

Definition 3.22. *For $\widehat{\beta}_\Psi$ defined in Definition 3.9(b) we define $\widehat{\tau}_{\mathrm{SSR}} > 0$ via*

$$\widehat{\tau}_{\mathrm{SSR}}^2 = \begin{cases} \widehat{\beta}_\Psi^2 / \|\mathbf{L}\| & \text{if } \widehat{\beta}_\Psi < \infty, \\ \infty & \text{if } \widehat{\beta}_\Psi = \infty. \end{cases} \tag{3.33}$$

As for the definition of τ_{SSR} we have that, if $\widehat{\beta}_\Psi = \infty$, the condition $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ is not a restriction at all and we could theoretically omit it. By exploiting this step-size restriction as well as the one stemming from (3.22) we are able to state bounds for the matrix functions occurring in the representation formulae (3.27) for $\mathbf{q}_n$ and (3.29) for $\partial_\tau \mathbf{q}_n$. Clearly, for $s \in \mathbb{R}$ and $\mathbf{q} \in \mathbb{R}^m$ we have $\|\cos(s\mathbf{\Phi})\mathbf{q}\| \le \|\mathbf{q}\|$ and $\|\sin(s\mathbf{\Phi})\mathbf{q}\| \le \|\mathbf{q}\|$ for $\tau \le \tau_{\mathrm{SSR}}$. Moreover, we have the following bounds involving $\sin \mathbf{\Phi}$.

Lemma 3.23. (a) *Let $\tau \le \tau_{\mathrm{SSR}}$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\|\boldsymbol{\mathcal{S}}_n \mathbf{q}\| \le n \|\mathbf{q}\| \tag{3.34a}$$

as well as

$$\|\sin \mathbf{\Phi}\, \mathbf{q}\| \le \tau \|\mathbf{q}\|_{\mathbf{L}}. \tag{3.34b}$$

(b) *Let $\tau \le \widehat{\tau}_{\mathrm{SSR}}$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \widehat{\mathbf{\Psi}} \mathbf{q} \right\|_{\mathbf{L}} \le m_1^{-1/2} \|\mathbf{q}\| \tag{3.35}$$

where m_1 is defined in (3.12).

(c) *Let $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ and $\mathbf{L}$ be positive definite. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \widehat{\mathbf{\Psi}} \mathbf{q} \right\| \le m_1^{-1/2} \|\mathbf{q}\|_{\mathbf{L}^{-1}} \le c_{\mathrm{inv}}\, m_1^{-1/2} \|\mathbf{q}\|, \tag{3.36a}$$

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \mathbf{q} \right\| \le c_{\mathrm{stb}} \|\mathbf{q}\|, \tag{3.36b}$$

with $c_{\mathrm{stb}} = m_1^{-1/2} \max\{\frac{1}{2}\tau \widetilde{m}_1^{-1/2}, c_{\mathrm{inv}} \widetilde{m}_2^{-1/2}\}$, where c_{inv} and $\widetilde{m}_1$, $\widetilde{m}_2$ are given in (2.13) and (3.12), respectively. Moreover, $\mathbf{\Psi}$ and $\sin \mathbf{\Phi}$ are nonsingular for $\tau > 0$.

We note that the dependency of the constant c_{stb} on the step size τ is not a problem at all, because for good approximations $\mathbf{q}_n \approx \mathbf{q}(t_n)$ one usually chooses step sizes $\tau < 1$. Hence, with such an additional assumption we simply could get rid of the factor τ.

Proof. Since the matrices $\mathbf{L}$ and $\mathbf{\Phi}$ are symmetric and simultaneously unitarily diagonalizable, it is sufficient to show the bounds for the eigenvalues $z \in [0, \beta_\Psi^2] \cap \mathbb{R}$ and $\phi \in [0, \pi]$ of $\tau^2 \mathbf{L}$ and $\mathbf{\Phi}$, respectively, belonging to the same eigenvector.

(a) The bound (3.34a) follows from the fact that for $n \in \mathbb{N}_0$ we have

$$\left| \frac{\sin(n\phi)}{\sin \phi} \right| = \left| \frac{\mathrm{e}^{\mathrm{i}n\phi} - \mathrm{e}^{-\mathrm{i}n\phi}}{\mathrm{e}^{\mathrm{i}\phi} - \mathrm{e}^{-\mathrm{i}\phi}} \right| \le \sum_{\ell=0}^{n-1} \left| \mathrm{e}^{\mathrm{i}(n-1-\ell)\phi} \mathrm{e}^{-\mathrm{i}\ell\phi} \right| \le n,$$

where equality holds for $\phi \in \{0, \pi\}$ or, equivalently, for $z \in [0, \beta_\Psi^2]$ with $\Psi(z) \in \{0, 4\}$. For the second bound (3.34b) we use the definition (3.23b) for $\sin \mathbf{\Phi}$, (3.11), and (3.19) to obtain

$$\sin(\phi) = \Psi(z)^{1/2} (1 - \tfrac{1}{4}\Psi(z))^{1/2} \le \widehat{\Psi}(z)^{1/2} z^{1/2} \le z^{1/2}.$$

(b) Because of the stronger step-size restriction we now have $z \in [0, \widehat{\beta}_\Psi^2] \cap \mathbb{R}$ and $\phi \in [0, \pi)$. The first inequality (3.35) follows from

$$\left| \frac{\sin(n\phi)}{\sin\phi} \widehat{\Psi}(z) z^{1/2} \right| = \left| \frac{\sin(n\phi)}{\cos(\frac{1}{2}\phi)} \widehat{\Psi}(z)^{1/2} \right| \leq |\cos(\tfrac{1}{2}\phi)^{-1}| = |(1 - \tfrac{1}{4}\Psi(z))^{-1/2}| \leq m_1^{-1/2},$$

where we used (3.30), (B.2a) in the first step and then (3.19), again (3.30), and (3.12).

(c) For the inequalities (3.36) we note that $z \geq \tau^2 c_{\mathrm{inv}}^{-2}$ due to the positive definiteness of $\mathbf{L}$. Thus, for $\tau > 0$ this yields $\Psi(z) > 0$ and $\sin(\phi) > 0$ because of $\phi > 0$. In particular, we have that $\Psi(\tau^2\mathbf{L})$ and $\sin\boldsymbol{\Phi}$ are nonsingular. Employing the same equations as before yields

$$\tau \big|\sin(\phi)^{-1} \widehat{\Psi}(z)\big| = \tau \Big|(1 - \tfrac{1}{4}\Psi(z))^{-1/2} z^{-1/2} \widehat{\Psi}(z)^{1/2}\Big| \leq \tau m_1^{-1/2} |z^{-1/2}| \leq m_1^{-1/2} c_{\mathrm{inv}}$$

and

$$\tau \big|\sin(\phi)^{-1}\big| = \tau \Big|(1 - \tfrac{1}{4}\Psi(z))^{-1/2} \Psi(z)^{-1/2}\Big| \leq m_1^{-1/2}\, \tau |\Psi(z)^{-1/2}| \leq c_{\mathrm{stb}},$$

which finishes the proof. □

From the proof we see that by utilizing (3.13) in the case of $\widehat{\beta}_\Psi < \infty$ we could theoretically replace the bound in (3.36b) with $c_{\mathrm{stb}} = c_{\mathrm{inv}}(m_1 m_2)^{-1/2}$; see [CHS20]. However, as mentioned in Remark 3.10 we avoid the usage of m_2. For the same reasons we do not make use of the bound (for $\widehat{\beta}_\Psi < \infty$)

$$\big\| \sin(\boldsymbol{\Phi})^{-1} \mathbf{q} \big\|_{\mathbf{L}} \leq m_1^{-1/2} \max\{\widehat{\beta}_\Psi \widetilde{m}_1^{-1/2}, \widetilde{m}_2^{-1/2}\} \|\mathbf{q}\|, \tag{3.37}$$

which can be shown in a similar way as (3.36b).

Remark 3.24. The constant c_{stb} in the estimate (3.36b) can be improved if one exploits that the function Ψ cannot simultaneously be near four and zero. Together with

$$x - \tfrac{1}{4}x^2 \geq \min\{x_1 - \tfrac{1}{4}x_1^2, x_2 - \tfrac{1}{4}x_2^2\} \qquad \text{for } 0 < x_1 \leq x \leq x_2 < 4,$$

this leads then to a smaller constant. We omit the details for the sake of readability. ◇

Remark 3.25. The bound for $\boldsymbol{\mathcal{S}}_n$ in (3.34a) is sharp even for positive definite $\mathbf{L}$ if $\Psi(z) \in \{0, 4\}$ for some $z \in (\widehat{\beta}_\Psi^2, \beta_\Psi^2)$. This can be seen from the proof of the estimate if one chooses $\mathbf{q}$ as an eigenvector of $\tau^2\mathbf{L}$ corresponding to an eigenvalue z such that $\Psi(z) \in \{0, 4\}$. A close inspection of the second and third part of the proof shows that for the bound (3.36a) the additional factor $\widehat{\boldsymbol{\Psi}}$ ensures that only those z are problematic where $\Psi(z) = 4$. We refer to Section 3.5 for more insight into this behavior. ◇

Example 3.26 (LF continued). With m_1, $\widetilde{m}_1$ and $\widetilde{m}_2$ chosen as in Example 3.11 we obtain for the leapfrog scheme (2.20) that $\tau^2\|\mathbf{L}\| \leq 4\vartheta^2 < 4$, $\vartheta \in (0, 1)$, which is in accordance with Lemma 2.16. Moreover, the bounds in Lemma 3.23 hold with $m_1 = 1 - \vartheta^2$ and $c_{\mathrm{stb}} = c_{\mathrm{inv}}(1 - \vartheta^2)^{-1/2}$. ◇

We are now in a position to study the stability in the standard norm and the discrete energy norm. We start with the first one.

Theorem 3.27. *The approximations* $\mathbf{q}_n$ *obtained by the general scheme* (3.1) *applied to* (2.12) *satisfy for* $n \geq 0$ *and*

(a) $\tau \leq \tau_{\mathrm{SSR}}$

$$\|\mathbf{q}_n\| \leq \|\mathbf{q}_0\| + t_n \|\dot{\mathbf{q}}_0\| + t_n \tau \sum_{\ell=0}^{n-1} \|\mathbf{g}_\ell\|, \tag{3.38a}$$

(b) $\tau \le \widehat{\tau}_{\mathrm{SSR}}$

$$\|\mathbf{q}_n\| \le \|\mathbf{q}_0\| + \min\{t_n, c_{\mathrm{inv}} m_1^{-1/2}\}\|\dot{\mathbf{q}}_0\| + \min\{t_n, c_{\mathrm{inv}} m_1^{-1/2}\}\tau \sum_{\ell=0}^{n-1} \|\mathbf{g}_\ell\|, \tag{3.38b}$$

where τ_{SSR} *and* $\widehat{\tau}_{\mathrm{SSR}}$ *are defined in Definitions* 3.17 *and* 3.22*, respectively.*

From the first stability estimate (3.38a) in the theorem we see that under the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ the approximations $\mathbf{q}_n$ can only be bounded with constants growing linearly in time, even in the case of a positive definite matrix $\mathbf{L}$. This is in contrast to the behavior of the exact solution; cf. (2.15a).

The situation changes if we employ the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ (recall that we formally set $c_{\mathrm{inv}} = \infty$ for $\mathbf{L}$ singular). For a positive definite $\mathbf{L}$ we then obtain uniformly bounded approximations in the homogeneous case. This perfectly corresponds to the behavior of the exact solution. Note that the stronger step-size restriction is even advantageous for positive semidefinite $\mathbf{L}$, since only eigenvector(s) of $\mathbf{L}$ belonging to zero eigenvalue(s) are responsible for the factor t_n.

Proof of Theorem 3.27. Let $\tau \le \tau_{\mathrm{SSR}}$. From Corollary 3.20 we have by taking norms and the triangle inequality

$$\|\mathbf{q}_n\| \le \|\cos(n\mathbf{\Phi})\mathbf{q}_0\| + \tau\|\boldsymbol{\mathcal{S}}_n\widehat{\mathbf{\Psi}}\dot{\mathbf{q}}_0\| + \tau^2 \sum_{\ell=0}^{n-1} \|\boldsymbol{\mathcal{S}}_{n-\ell}\widehat{\mathbf{\Psi}}\mathbf{g}_\ell\|,$$

which implies the first part with (3.34a), (3.19), and $n\tau = t_n$.

For the second part, we obtain the bounds with t_n as before because of $\widehat{\tau}_{\mathrm{SSR}} \le \tau_{\mathrm{SSR}}$. If $\mathbf{L}$ is positive definite, we can employ (3.36a) to obtain the bound with $c_{\mathrm{inv}} m_1^{-1/2}$. □

We now turn towards stability estimates in the discrete energy norm (3.20b), where we consider the single terms $\|\partial_\tau \mathbf{q}_n\|$ and $\|\mathbf{q}_n\|_{\mathbf{L}}$ separately.

Theorem 3.28. *The approximations* $\mathbf{q}_n$ *obtained by the general scheme* (3.1) *applied to* (2.12) *satisfy for* $n \ge 0$ *and*

(a) $\tau \le \tau_{\mathrm{SSR}}$

$$\|\partial_\tau \mathbf{q}_n\| \le \|\mathbf{q}_0\|_{\mathbf{L}} + \|\dot{\mathbf{q}}_0\| + \tau \sum_{\ell=0}^{n} \|\mathbf{g}_\ell\|, \tag{3.39a}$$

$$\|\mathbf{q}_n\|_{\mathbf{L}} \le \|\mathbf{q}_0\|_{\mathbf{L}} + t_n\|\dot{\mathbf{q}}_0\|_{\mathbf{L}} + t_n\tau \sum_{\ell=0}^{n-1} \|\mathbf{g}_\ell\|_{\mathbf{L}}, \tag{3.39b}$$

(b) $\tau \le \widehat{\tau}_{\mathrm{SSR}}$

$$\|\mathbf{q}_n\|_{\mathbf{L}} \le \|\mathbf{q}_0\|_{\mathbf{L}} + m_1^{-1/2}\|\dot{\mathbf{q}}_0\| + m_1^{-1/2}\tau \sum_{\ell=0}^{n-1} \|\mathbf{g}_\ell\|. \tag{3.39c}$$

where τ_{SSR} *and* $\widehat{\tau}_{\mathrm{SSR}}$ *are defined in Definitions* 3.17 *and* 3.22*, respectively.*

We observe that $\|\partial_\tau \mathbf{q}_n\|$ can be bounded uniformly in time for $\mathbf{g} \equiv 0$ with the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$. In contrast to this, we get for $\|\mathbf{q}_n\|_{\mathbf{L}}$ in general only a bound with grows linearly in time in the homogeneous case. Additionally, we have to measure $\dot{\mathbf{q}}_0$ and

the inhomogeneity $\mathbf{g}$ in a "stronger" norm compared to the stability bound (2.15b) of the exact solution. Hence, only the part $\|\mathbf{q}_n\|_{\mathbf{L}}$ of the discrete energy norm $|||\mathbf{q}_n|||_\tau$ is responsible for the bad stability behavior under the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$.

As before, if we use the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, we also get for $\|\mathbf{q}_n\|_{\mathbf{L}}$ a bound with constants uniformly bounded in time. A combination of (3.39a) and (3.39c) then yields for the discrete energy norm

$$|||\mathbf{q}_n|||_\tau = \big(\|\partial_\tau \mathbf{q}_n\|^2 + \|\mathbf{q}_n\|_{\mathbf{L}}^2\big)^{1/2} \le 2^{1/2}\bigg(\|\mathbf{q}_0\|_{\mathbf{L}} + m_1^{-1/2}\|\dot{\mathbf{q}}_0\| + m_1^{-1/2}\tau\sum_{\ell=0}^{n}\|\mathbf{g}_\ell\|\bigg),$$

since $m_1 < 1$.

Proof of Theorem 3.28. Let $\tau \le \tau_{\mathrm{SSR}}$. Taking norms in (3.27) and (3.29) yields

$$\|\mathbf{q}_n\|_{\mathbf{L}} \le \|\mathbf{q}_0\|_{\mathbf{L}} + \tau\big\|\boldsymbol{\mathcal{S}}_n\widehat{\boldsymbol{\Psi}}\dot{\mathbf{q}}_0\big\|_{\mathbf{L}} + \tau^2\sum_{\ell=0}^{n-1}\big\|\boldsymbol{\mathcal{S}}_{n-\ell}\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell\big\|_{\mathbf{L}},$$

$$\|\partial_\tau \mathbf{q}_n\| \le \tau^{-1}\|\sin\boldsymbol{\Phi}\mathbf{q}_0\| + \|\widehat{\boldsymbol{\Psi}}\dot{\mathbf{q}}_0\| + \tau\sum_{\ell=0}^{n}\|\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell\|.$$

The second inequality leads to the bound (3.39a) by using (3.34b) and (3.19).

For the first inequality we employ on the one hand the bound for $\boldsymbol{\mathcal{S}}_n$ in (3.34a) and again (3.19), which then yields (3.39b). On the other hand, we obtain for $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ with (3.35) the bound (3.39c). □

In Section 3.5 we show two possibilities how one can achieve uniform bounds for homogeneous problems even under the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$. One option consists in modifying the starting value, the other option is to look at averaged approximations instead of $\mathbf{q}_n$.

3.3.2. Preservation of a discrete energy

Next, we prove that the scheme (3.1) applied to the linear problem (2.12) with $\mathbf{g} \equiv 0$ nearly preserves the discrete energy norm $|||\cdot|||_\tau$ defined in (3.20b). To do so, we show existence of a conserved quantity of the scheme (3.1), which is a second-order perturbation of the square of the discrete energy norm. In combination, this reflects the energy conserving behavior of the exact solution of the linear, homogeneous problem as stated in (2.16).

Definition 3.29. *For the approximations $\mathbf{q}_n$ obtained by the scheme (3.1) we define for $n \ge 0$*

$$\tau^2\mathcal{M}_{\mathbf{q},n} = \tau^2\|\partial_\tau \mathbf{q}_n\|^2 + (\boldsymbol{\Psi}\mathbf{q}_n, \mathbf{q}_n) - \tfrac{1}{4}\|\boldsymbol{\Psi}\mathbf{q}_n\|^2, \tag{3.40}$$

where $\partial_\tau \mathbf{q}_n$ is defined in (3.20a).

We note that there exist other quantities for the two-step scheme (3.1a) which are conserved for $\mathbf{g} \equiv 0$; cf. [CHS20] for a variant on half steps. For the quantity (3.40) we now prove that it is nonnegative under the step-size restriction $\tau \le \tau_{\mathrm{SSR}}$. In addition, we show that the approximations of the scheme (3.1a) conserve the quantity for all $n \in \mathbb{N}_0$ in the homogeneous case.

Lemma 3.30. *Let $\tau \le \tau_{\mathrm{SSR}}$. The approximations obtained by (3.1) satisfy*

$$\mathcal{M}_{\mathbf{q},n} \ge 0 \qquad \text{for all } n = 0, 1, \dots\ . \tag{3.41}$$

Proof. Rewriting of $\mathcal{M}_{\mathbf{q},n}$ yields for all $n \geq 0$

$$\tau^2 \mathcal{M}_{\mathbf{q},n} = \tau^2 \|\partial_\tau \mathbf{q}_n\|^2 + (\mathbf{\Psi}\mathbf{q}_n, (\mathbf{I} - \tfrac{1}{4}\mathbf{\Psi})\mathbf{q}_n).$$

Since the eigenvalues of $\mathbf{\Psi}$ are contained in $[0,4]$ due to Definition 3.17 of τ_{SSR}, we obtain the nonnegativity of the second term. □

Lemma 3.31. *Let* $\mathbf{g} \equiv 0$. *The approximations obtained by* (3.1) *satisfy*

$$\mathcal{M}_{\mathbf{q},n} = \mathcal{M}_{\mathbf{q},0} \qquad \text{for all } n = 0, 1, \dots \,. \tag{3.42}$$

Proof. We begin with $\mathcal{M}_{\mathbf{q},n+1} = \mathcal{M}_{\mathbf{q},n}$ for $n \geq 1$. For this, we first observe that the sum and difference of two consecutive steps of the two-step scheme (3.1a) for $\mathbf{g} \equiv 0$ are given by

$$\begin{aligned} \mathbf{q}_{n+2} - \mathbf{q}_{n+1} - \mathbf{q}_n + \mathbf{q}_{n-1} &= -\mathbf{\Psi}(\mathbf{q}_{n+1} + \mathbf{q}_n), \\ \mathbf{q}_{n+2} - 3\mathbf{q}_{n+1} + 3\mathbf{q}_n - \mathbf{q}_{n-1} &= -\mathbf{\Psi}(\mathbf{q}_{n+1} - \mathbf{q}_n). \end{aligned}$$

Using this, we obtain

$$\begin{aligned} \tau^2\|\partial_\tau \mathbf{q}_{n+1}\|^2 - \tau^2\|\partial_\tau \mathbf{q}_n\|^2 &= \tfrac{1}{4}((\mathbf{q}_{n+2} - \mathbf{q}_n) - (\mathbf{q}_{n+1} - \mathbf{q}_{n-1}), (\mathbf{q}_{n+2} - \mathbf{q}_n) + (\mathbf{q}_{n+1} - \mathbf{q}_{n-1})) \\ &= \tfrac{1}{4}(-\mathbf{\Psi}(\mathbf{q}_{n+1} + \mathbf{q}_n), 4\mathbf{q}_{n+1} - 4\mathbf{q}_n - \mathbf{\Psi}(\mathbf{q}_{n+1} - \mathbf{q}_n)) \\ &= (-\mathbf{\Psi}\mathbf{q}_{n+1}, \mathbf{q}_{n+1}) + (\mathbf{\Psi}\mathbf{q}_n, \mathbf{q}_n) + \tfrac{1}{4}\|\mathbf{\Psi}\mathbf{q}_{n+1}\| - \tfrac{1}{4}\|\mathbf{\Psi}\mathbf{q}_n\|. \end{aligned}$$

Rearranging this formula yields with Definition 3.29 $\mathcal{M}_{\mathbf{q},n+1} = \mathcal{M}_{\mathbf{q},n}$ for $n \geq 1$.

For proving $\mathcal{M}_{\mathbf{q},1} = \mathcal{M}_{\mathbf{q},0}$ we can proceed in a similar way. Here, in contrast to the previous calculations, we take the sum and difference between the first step of the two-step scheme (3.1a) and twice of the starting value (3.1b). □

As consequence of these two lemmas we obtain that $\mathcal{M}_{\mathbf{q},n}$ is conserved and nonnegative for all $n \in \mathbb{N}$ under the weaker step-size restriction $\tau \leq \tau_{\mathrm{SSR}}$. In contrast to this, we have shown in Theorems 3.27 and 3.28 that $\|\mathbf{q}_n\|$ and $\|\mathbf{q}_n\|_{\mathbf{L}}$ grow linearly in time for $\dot{\mathbf{q}}_0 \neq 0$ under the weaker step-size restriction; see also Remark 3.25. Hence, conservation of $\mathcal{M}_{\mathbf{q},n}$ and correct long-time behavior of the numerical solution are not directly correlated for $\tau \leq \tau_{\mathrm{SSR}}$.

A close inspection of the previous two proofs also shows that independent of the choice of the starting value $\mathbf{q}_1$ we get $\mathcal{M}_{\mathbf{q},n} = \mathcal{M}_{\mathbf{q},1}$ and the nonnegativity of $\mathcal{M}_{\mathbf{q},n}$ for $n \geq 1$. Hence, a modification of the starting value has only influence on the conserved quantity for $n = 0$. Moreover, by an adaption of $\mathcal{M}_{\mathbf{q},0}$ in (3.40) to another "sensible" starting value we could prove similar results as for (3.1b); see Section 3.5.1 and in particular Remark 3.50.

The next theorem shows that the approximations of the scheme (3.1) do not have a drift in the discrete energy for arbitrarily long simulation times. With the above comment it is not surprising that we need the stronger step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ to show this result. Otherwise we would get again a linear drift in time.

Theorem 3.32. *Let* $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ *and* $\mathbf{g} \equiv 0$. *The approximations obtained by* (3.1) *satisfy*

$$\left| |||\mathbf{q}_n|||_\tau^2 - \mathcal{M}_{\mathbf{q},0} \right| \leq C\tau^2, \qquad \text{for all } n = 0, 1, 2, \dots,$$

with a constant $C = C(\|\mathbf{L}\mathbf{q}_0\|, \|\dot{\mathbf{q}}_0\|_{\mathbf{L}})$ *which is independent of* $\mathbf{L}, \tau$, n, *and* t_n.

Proof. Let $\tau \leq \tau_{\mathrm{SSR}}$. From Lemma 3.31 and the definitions (3.20b) of $|||\cdot|||_\tau$ as well as (3.40) of $\mathcal{M}_{\mathbf{q},n}$ we obtain

$$\begin{aligned}|||\mathbf{q}_n|||_\tau^2 - \mathcal{M}_{\mathbf{q},0} = |||\mathbf{q}_n|||_\tau^2 - \mathcal{M}_{\mathbf{q},n} &= \tau^{-2}((\tau^2\mathbf{L} - \boldsymbol{\Psi})\mathbf{q}_n, \mathbf{q}_n) + \tfrac{1}{4}\tau^{-2}\|\boldsymbol{\Psi}\mathbf{q}_n\|^2 \\ &= -\tau^2(\boldsymbol{\Upsilon}\mathbf{L}\mathbf{q}_n, \mathbf{L}\mathbf{q}_n) + \tfrac{1}{4}\tau^2\|\widehat{\boldsymbol{\Psi}}\mathbf{L}\mathbf{q}_n\|^2,\end{aligned}$$

where we made use of Definitions 3.3 and 3.12 in the last equality. With $\boldsymbol{\Upsilon} \leq 0$ by (3.19) and Definition 3.13 we can bound this by

$$0 \leq |||\mathbf{q}_n|||_\tau^2 - \mathcal{M}_{\mathbf{q},0} \leq \tau^2(\tfrac{1}{2}m_3 + \tfrac{1}{4})\|\mathbf{L}\mathbf{q}_n\|^2.$$

Further, multiplying the representation formula (3.27) with $\mathbf{L}$ and taking norms yields with (3.35) under the step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ as in the proof of Theorem 3.28

$$\|\mathbf{L}\mathbf{q}_n\| \leq \|\mathbf{L}\mathbf{q}_0\| + m_1^{-1/2}\|\dot{\mathbf{q}}_0\|_{\mathbf{L}},$$

which concludes the proof. □

3.3.3. Stability for an extended linear problem

We conclude this section about stability by showing that for semilinear problems the stronger step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ is indispensable in general. To do so, we consider a special linear case of (2.1). More precisely, we set $\mathbf{g}(t, \mathbf{q}) = -\mathbf{G}\mathbf{q} + \mathbf{f}(t)$ for all $t \geq 0$, where $\mathbf{G} \in \mathbb{R}^{m\times m}$ is a symmetric and positive semidefinite matrix. The differential equation (2.1) then reduces to the linear system

$$\ddot{\mathbf{q}}(t) = -(\mathbf{L} + \mathbf{G})\mathbf{q}(t) + \mathbf{f}(t), \qquad \mathbf{q}(0) = \mathbf{q}_0, \qquad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0. \tag{3.43}$$

Clearly, the exact solution of (3.43) satisfies the stability bounds (2.15) if we replace $\mathbf{L}$ with $\mathbf{L} + \mathbf{G}$ and $\mathbf{g}$ with $\mathbf{f}$.

Theorem 3.33. *Let $\mathbf{L}$ and $\mathbf{G}$ commute and let the step-size restrictions*

$$\tau^2\|\mathbf{L}\| \leq \widehat{\beta}_\Psi^2, \qquad \tau^2\|\mathbf{G}\| \leq 4\vartheta^2, \qquad \vartheta^2 \in (0, m_1), \tag{3.44}$$

be satisfied. Then the approximations $\mathbf{q}_n$ obtained by the general scheme (3.1) applied to (3.43) satisfy for $n \geq 0$

$$\|\mathbf{q}_n\| \leq \|\mathbf{q}_0\| + \min\{t_n, c_{\mathrm{stb},\mathbf{G}}\}\|\dot{\mathbf{q}}_0\| + \min\{t_n, c_{\mathrm{stb},\mathbf{G}}\}\,\tau\sum_{\ell=0}^{n-1}\|\mathbf{f}(t_\ell)\| \tag{3.45}$$

with $c_{\mathrm{stb},\mathbf{G}} = m_{1,\vartheta}^{-1/2}\max\{\frac{1}{2}\tau\widetilde{m}_1^{-1/2}, c_{\mathrm{inv}}\widetilde{m}_2^{-1/2}\}$, where $m_{1,\vartheta} = m_1 - \vartheta^2$ and c_{inv} and m_1, $\widetilde{m}_1$, $\widetilde{m}_2$ are given in (2.13) and (3.12), respectively.

The theorem states that we do not only require the stronger step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ for obtaining bounded approximations $\mathbf{q}_n$ in the case of $\mathbf{f} \equiv 0$ but also a second step-size restriction involving $\|\mathbf{G}\|$. However, even for small $m_1 > 0$ the second condition in the step-size restriction (3.44) is often less restrictive for the step size τ than the first one (recall that we are interested in functions $\mathbf{g}$ with small to moderate Lipschitz constant, i.e., $\|\mathbf{G}\| \ll \|\mathbf{L}\|$). We further point out that in [CHS20] a similar theorem is proven for the two-step scheme (3.7) where the assumption on the comutativity of $\mathbf{L}$ and $\mathbf{G}$ is not required.

Proof. We first show that with the step-size restriction (3.44) the matrix $\mathbf{\Psi_G} = \tau^2\widehat{\Psi}(\tau^2\mathbf{L})(\mathbf{L}+\mathbf{G})$ is symmetric, positive semidefinite, and the largest eigenvalue bounded by 4. The symmetry of $\mathbf{\Psi_G}$ follows from the commutativity of $\mathbf{L}$ and $\mathbf{G}$. By using again the commutativity together with the first step-size restriction in (3.44) and (3.12) we further obtain for all $\mathbf{q} \in \mathbb{R}^m$

$$(\mathbf{\Psi_G q}, \mathbf{q}) = (\mathbf{\Psi q}, \mathbf{q}) + \tau^2(\widehat{\mathbf{\Psi}}\mathbf{G q}, \mathbf{q}) \geq (\mathbf{\Psi q}, \mathbf{q}) \geq \min\{4\widetilde{m}_1, \widetilde{m}_2\tau^2 c_{\text{inv}}^{-2}\}\|\mathbf{q}\|^2 \geq 0. \tag{3.46}$$

Moreover, we have with (3.2), (3.44), (3.12), and (3.19)

$$\begin{aligned}(\mathbf{\Psi_G q}, \mathbf{q}) &= (\mathbf{\Psi q}, \mathbf{q}) + \tau^2(\widehat{\mathbf{\Psi}}\mathbf{G q}, \mathbf{q}) \\ &\leq (4(1-m_1) + \tau^2\|\mathbf{G}\|)\|\mathbf{q}\|^2 \leq 4(1-m_{1,\vartheta})\|\mathbf{q}\|^2.\end{aligned} \tag{3.47}$$

Hence, under the step-size restriction (3.44) the eigenvalues of $\mathbf{\Psi_G}$ are contained in $[0, 4]$.

We can now proceed as before by showing stability via a representation formula. In particular, we obtain

$$\mathbf{q}_n = \cos(n\mathbf{\Phi_G})\mathbf{q}_0 + \tau\frac{\sin(n\mathbf{\Phi_G})}{\sin\mathbf{\Phi_G}}\widehat{\mathbf{\Psi}}\dot{\mathbf{q}}_0 + \tau^2\sum_{\ell=0}^{n-1}\xi_{\ell,n}\frac{\sin((n-\ell)\mathbf{\Phi_G})}{\sin\mathbf{\Phi_G}}\widehat{\mathbf{\Psi}}\mathbf{f}(t_\ell), \tag{3.48}$$

where $\xi_{\ell,n}$ is given in (3.26) and $\mathbf{\Phi_G}$ with spectrum in $[0, \pi]$ is defined by $\cos\mathbf{\Phi_G} = \mathbf{I} - \frac{1}{2}\mathbf{\Psi_G}$. Note that, if $\mathbf{L}$ is positive definite, we obtain with (3.46) and (3.47) similarly to the proof of the bound (3.36b) that

$$\tau\|(\sin\mathbf{\Phi_G})^{-1}\mathbf{q}\| \leq c_{\text{stb},\mathbf{G}}\|\mathbf{q}\|,$$

Taking the norm in (3.48) concludes the proof. □

Remark 3.34. From (3.47) we see that under the weaker step-size restriction $\tau \leq \tau_{\text{SSR}}$ with τ_{SSR} given in (3.22) the eigenvalues of $\tau^2\mathbf{\Psi_G}$ are contained in $[0, 4]$ in general only if $\mathbf{G} = 0$, since $(\mathbf{\Psi q}, \mathbf{q}) \leq 4\|\mathbf{q}\|^2$ because of (3.11). Hence, in general the weaker step-size restriction $\tau \leq \tau_{\text{SSR}}$ is not sufficient to guarantee stability for the scheme (3.1) in the semilinear case; see also the second numerical example in Section 4.5.3. ◇

3.4. Error analysis

In the previous section we established the stability of the general scheme (3.1) for linear problems. The aim of this section is to provide the error analysis not only for linear but also for semilinear problems. We show convergence results for linear problems in both the standard norm $\|\cdot\|$ and the discrete energy norm $|||\cdot|||_\tau$ as well as convergence results in the standard norm for semilinear problems.

In the following, let always $T \in (0, t_*)$ with t_* defined in Theorem 2.4. For $t_n \leq T$ we denote by

$$\mathbf{e}_n = \widetilde{\mathbf{q}}_n - \mathbf{q}_n, \qquad \widetilde{\mathbf{q}}_n = \mathbf{q}(t_n), \tag{3.49}$$

the error between the exact solution $\mathbf{q}$ of (2.1) at time t_n and the numerical approximation $\mathbf{q}_n$ of the general scheme (3.1). Further, for $\mathbf{q} \in C^k([0, T])$, $k \in \mathbb{N}$, we abbreviate bounds on the kth derivative of $\mathbf{q}$ by

$$B_t^{(k)} = \max_{0\leq s\leq t}\|\mathbf{q}^{(k)}(s)\|, \qquad k = 1, 2, \dots, \tag{3.50}$$

and by

$$\boldsymbol{\delta}_{n,\pm}^{(k)} = \tau^k \int_0^1 \kappa_\pm^{(k-1)}(\sigma)\, \mathbf{q}^{(k)}(t_n \pm \tau\sigma)\, \mathrm{d}\sigma, \qquad \kappa_\pm^{(\ell)}(\sigma) = \frac{(\pm 1)^{\ell+1}(1-\sigma)^\ell}{\ell!}, \tag{3.51}$$

the remainder terms of the $(k-1)$st-order Taylor expansion of $\widetilde{\mathbf{q}}_{n\pm1}$ at t_n.

Lemma 3.35. *For $\mathbf{q} \in C^k([0,T])$ and $t_{n+1} \le T$ the remainder terms (3.51) are bounded by*

$$\|\boldsymbol{\delta}_{n,+}^{(k)}\| \le \tau^k \tfrac{1}{k!} \max_{t_n \le s \le t_{n+1}} \|\mathbf{q}^{(k)}(s)\|, \qquad \|\boldsymbol{\delta}_{n,-}^{(k)}\| \le \tau^k \tfrac{1}{k!} \max_{t_{n-1} \le s \le t_n} \|\mathbf{q}^{(k)}(s)\|. \tag{3.52}$$

Proof. The bound for $\boldsymbol{\delta}_{n,+}$ follows from

$$\|\boldsymbol{\delta}_{n,+}^{(k)}\| \le \tau^k \max_{t_n \le s \le t_{n+1}} \|\mathbf{q}^{(k)}(s)\| \int_0^1 |\kappa_+^{(\ell)}(\sigma)|\, \mathrm{d}\sigma.$$

For $\boldsymbol{\delta}_{n,-}^{(k)}$ the proof is done analogously. □

3.4.1. Representation formula for errors

As for the stability analysis we start with the derivation of a representation formula for the errors. With this representation at hand we are able to prove the error bounds for linear and semilinear problems. To derive such a representation formula we have to prove an error recursion.

Lemma 3.36. *Let $\mathbf{q} \in C^4([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_n$ of the two-step scheme (3.1a) satisfies for $n \ge 1$ the recursion*

$$\mathbf{e}_{n+1} - 2\mathbf{e}_n + \mathbf{e}_{n-1} = \tau^2 \widehat{\boldsymbol{\Psi}}(-\mathbf{L}\mathbf{e}_n + \mathbf{r}_n) + \mathbf{d}_n, \tag{3.53a}$$

where the defect $\mathbf{d}_n$ is given by

$$\mathbf{d}_n = \boldsymbol{\Delta}_n + \widehat{\boldsymbol{\Psi}}\boldsymbol{\delta}_n^{(4)}, \qquad \boldsymbol{\Delta}_n = -\tau^2 \boldsymbol{\Upsilon}\mathbf{L}(\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1}), \quad \boldsymbol{\delta}_n^{(4)} = \boldsymbol{\delta}_{n,+}^{(4)} - \boldsymbol{\delta}_{n,-}^{(4)}, \tag{3.53b}$$

and

$$\mathbf{r}_n = \mathbf{g}(t_n, \widetilde{\mathbf{q}}_n) - \mathbf{g}(t_n, \mathbf{q}_n). \tag{3.53c}$$

Obviously, we have $\mathbf{r}_n = 0$ for linear problems. For the leapfrog scheme we have $\boldsymbol{\Upsilon} \equiv 0$, hence, only the second term in $\mathbf{d}_n$ remains.

Proof. As usually, we first insert the exact solution $\widetilde{\mathbf{q}}_n$ into the scheme (3.1a). This yields

$$\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1} = \tau^2 \widehat{\boldsymbol{\Psi}}(-\mathbf{L}\widetilde{\mathbf{q}}_n + \mathbf{g}(t_n, \widetilde{\mathbf{q}}_n)) + \mathbf{d}_n \tag{3.54}$$

with a defect $\mathbf{d}_n$. Subtracting (3.1a) from this relation leads to (3.53a) with $\mathbf{r}_n$ given in (3.53c).

In order to determine $\mathbf{d}_n$, we first observe that Taylor expansion of the exact solution $\widetilde{\mathbf{q}}_{n\pm1}$ at t_n yields

$$\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1} = \tau^2 \ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_{n,+}^{(4)} + \boldsymbol{\delta}_{n,-}^{(4)} = \tau^2 \ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_n^{(4)}. \tag{3.55}$$

Equating (3.54) and (3.55) then leads together with (2.1) to

$$\mathbf{d}_n = \tau^2(\mathbf{I} - \widehat{\boldsymbol{\Psi}})\ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_n^{(4)} = (\mathbf{I} - \widehat{\boldsymbol{\Psi}})(\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1}) + \widehat{\boldsymbol{\Psi}}\boldsymbol{\delta}_n^{(4)}, \tag{3.56}$$

where we used in the second step again (3.55). Applying (3.14) completes the proof. □

Remark 3.37. The first equality in (3.56) yields with (3.14) the more obvious form of the defect

$$\mathbf{d}_n = -\tau^4 \boldsymbol{\Upsilon} \mathbf{L} \ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_n^{(4)}. \tag{3.57}$$

However, we will see that for an error bound in the discrete energy norm the form (3.53b) of the defect is beneficial. ◇

Using the error recursion (3.53a) we now derive a representation formula for the error. Additionally, we state a formula for differences of errors $\partial_\tau \mathbf{e}_n$, which are defined as $\partial_\tau \mathbf{q}_n$ for $\mathbf{q}_n$; see (3.20a). To show these formulae we proceed as in Theorem 3.18 and Corollary 3.21.

Lemma 3.38. *Let $\tau \leq \tau_{\mathrm{SSR}}$ and let $\mathbf{q} \in C^4([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_n$, $n \geq 1$, of the two-step scheme (3.1a) satisfies*

$$\mathbf{e}_n = \boldsymbol{\mathcal{S}}_n \mathbf{e}_1 + \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell} (\tau^2 \hat{\boldsymbol{\Psi}} \mathbf{r}_\ell + \mathbf{d}_\ell) \tag{3.58}$$

with $\mathbf{d}_\ell, \mathbf{r}_\ell$ given in (3.53b) and (3.53c).

Moreover, the differences $\partial_\tau \mathbf{e}_n$ satisfy

$$\tau\, \partial_\tau \mathbf{e}_n = \cos(n\boldsymbol{\Phi}) \mathbf{e}_1 + \sum_{\ell=1}^{n} \xi_{\ell,n} \cos\big((n-\ell)\boldsymbol{\Phi}\big) (\tau^2 \hat{\boldsymbol{\Psi}} \mathbf{r}_\ell + \mathbf{d}_\ell). \tag{3.59}$$

Proof. From the error recursion in Lemma 3.36 we obtain analogously to the proof of Theorem 3.18 for $\tau \leq \tau_{\mathrm{SSR}}$

$$\mathbf{e}_n = \cos(n\boldsymbol{\Phi}) \mathbf{e}_0 + \boldsymbol{\mathcal{S}}_n (\mathbf{e}_1 - \cos \boldsymbol{\Phi}\, \mathbf{e}_0) + \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell} (\tau^2 \hat{\boldsymbol{\Psi}} \mathbf{r}_\ell + \mathbf{d}_\ell).$$

This yields (3.58), since $\mathbf{e}_0 = 0$ due to $\mathbf{q}(0) = \mathbf{q}_0$. The formula (3.59) for $\partial_\tau \mathbf{e}_n$ follows as in Corollary 3.21. □

Next, we turn our attention to the error $\mathbf{e}_1$ of the starting value (3.1b). For this, we have similarly to Lemma 3.36 the following.

Lemma 3.39. *Let $\mathbf{q} \in C^3([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_1$ of the starting value (3.1b) satisfies*

$$\mathbf{e}_1 = \mathbf{d}_0 = \boldsymbol{\Delta}_{0,*} + \hat{\boldsymbol{\Psi}} \boldsymbol{\delta}_{0,+}^{(3)}, \qquad \boldsymbol{\Delta}_{0,*} = -\tau^2 \boldsymbol{\Upsilon} \mathbf{L} (\tilde{\mathbf{q}}_1 - \tilde{\mathbf{q}}_0), \tag{3.60}$$

with $\boldsymbol{\delta}_{0,+}^{(3)}$ defined in (3.51).

Proof. Inserting the exact solution into the formula for the starting value (3.1b) yields the defect

$$\mathbf{d}_0 = \tilde{\mathbf{q}}_1 - \tilde{\mathbf{q}}_0 - \tau \hat{\boldsymbol{\Psi}} \dot{\mathbf{q}}(0) - \tfrac{1}{2}\tau^2 \hat{\boldsymbol{\Psi}} (-\mathbf{L}\tilde{\mathbf{q}}_0 + \mathbf{g}(0, \tilde{\mathbf{q}}_0)) = (\mathbf{I} - \hat{\boldsymbol{\Psi}})(\tilde{\mathbf{q}}_1 - \tilde{\mathbf{q}}_0) + \hat{\boldsymbol{\Psi}} \boldsymbol{\delta}_{0,+}^{(3)},$$

where we used in the last step again the differential equation (2.1) and Taylor expansion of $\tilde{\mathbf{q}}_1$ at 0. Using (3.14) shows (3.60) for $\mathbf{d}_0$. Further, since the initial values coincide with the exact values of the solution, we have $\mathbf{e}_1 = \mathbf{d}_0$. □

Clearly, for other starting values $\mathbf{q}_1$ satisfying $\|\widetilde{\mathbf{q}}_1 - \mathbf{q}_1\| \leq C\tau^3$ we obtain different defects. We again refer to Section 3.5.1 for more insight into the influence of $\mathbf{q}_1$. Further, as for the defect $\mathbf{d}_n$, $n \geq 1$, for the two-step scheme (3.1a) the defect $\mathbf{d}_0$ can be rewritten as

$$\mathbf{d}_0 = -\tau^3 \boldsymbol{\Upsilon} \mathbf{L}(\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau\ddot{\mathbf{q}}(0)) + \boldsymbol{\delta}_{0,+}^{(3)} \tag{3.61}$$

such that there is no factor $\widehat{\boldsymbol{\Psi}}$ in front of $\boldsymbol{\delta}_{0,+}^{(3)}$; cf. Remark 3.37.

We are now in a position to state error bounds for the scheme (3.1).

3.4.2. Error analysis for linear problems

In the following we show two error bounds for the standard norm $\|\cdot\|$ as well as one for the discrete energy norm $|||\cdot|||_\tau$. As for the stability analysis we distinguish between the two step-size restrictions $\tau \leq \tau_{\mathrm{SSR}}$ and $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$.

Recall that for linear problems (2.12) we have $\mathbf{r}_n = 0$ in the error recursion (3.53a). Moreover, the exact solution of (2.12) exists for all times $t \geq 0$; see Section 2.1.

We start with the standard norm and the weaker step-size restriction.

Theorem 3.40. *Let Assumptions 3.2 and 3.16 hold. Further, assume that the solution* $\mathbf{q}$ *of* (2.12) *satisfies* $\mathbf{q} \in C^4([0,T])$. *Then, for* $\tau \leq \tau_{\mathrm{SSR}}$ *and* $t_n \leq T$ *we have for the approximations* $\mathbf{q}_n$ *of the scheme* (3.1)

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \leq T\Big(C_1 + \tfrac{1}{2}C_T T\Big)\tau^2 \tag{3.62a}$$

with

$$C_1 = \tfrac{1}{2}m_3 \max_{0\leq s\leq\tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\| + \tfrac{1}{6}B_\tau^{(3)}, \qquad C_T = \tfrac{1}{2}m_3 \max_{0\leq s\leq T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\| + \tfrac{1}{12}B_T^{(4)}, \tag{3.62b}$$

where $B_\tau^{(3)}$, $B_T^{(4)}$ *are given by* (3.50) *and* m_3 *in* (3.15).

A similar result has already been proven in [CI17] for the exact starting value $\mathbf{q}_1 = \mathbf{q}(\tau)$. In this paper the authors consider a (modified) version of (3.1a), where $\widehat{\Psi}$ is either a polynomial or a rational function, e.g., the functions (3.3) or (3.4). In contrast to our work they prove their results with an extension of standard *energy techniques*.

Proof. From Lemma 3.38 we obtain with $\mathbf{r}_n = 0$, (3.53b), and (3.60)

$$\|\mathbf{e}_n\| \leq \|\boldsymbol{\mathcal{S}}_n \boldsymbol{\Delta}_{0,*}\| + \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_{0,+}^{(3)}\| + \sum_{\ell=1}^{n-1} \Big(\|\boldsymbol{\mathcal{S}}_{n-\ell} \boldsymbol{\Delta}_\ell\| + \|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_\ell^{(4)}\| \Big). \tag{3.63}$$

We bound the terms separately. The remainder terms of Taylor's theorem are bounded with (3.52) by

$$\|\boldsymbol{\delta}_\ell^{(4)}\| \leq \tfrac{1}{12}\tau^4 B_T^{(4)}, \qquad \|\boldsymbol{\delta}_{0,+}^{(3)}\| \leq \tfrac{1}{6}\tau^3 B_\tau^{(3)}. \tag{3.64}$$

For the defects $\boldsymbol{\Delta}_\ell$ given in (3.53b) we first observe that by Taylor expansion we obtain again with (3.52)

$$\|\mathbf{L}(\widetilde{\mathbf{q}}_{\ell+1} - 2\widetilde{\mathbf{q}}_\ell + \widetilde{\mathbf{q}}_{\ell-1})\| \leq \|\mathbf{L}\boldsymbol{\delta}_{\ell,+}^{(2)} + \mathbf{L}\boldsymbol{\delta}_{\ell,-}^{(2)}\| \leq \tau^2 \max_{0\leq s\leq T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\|. \tag{3.65}$$

and similar for $\boldsymbol{\Delta}_{0,*}$ given in (3.60). Hence, we get with Definition 3.13

$$\|\boldsymbol{\Delta}_\ell\| \le \tfrac{1}{2} m_3 \tau^4 \max_{0\le s\le T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\|, \qquad \|\boldsymbol{\Delta}_{0,*}\| \le \tfrac{1}{2} m_3 \tau^3 \max_{0\le s\le \tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\|. \tag{3.66}$$

Collecting and inserting these bounds in (3.63) yields with (3.34a) and (3.19)

$$\|\mathbf{e}_n\| \le t_n \tau^2 \Big(\tfrac{1}{2} m_3 \max_{0\le s\le \tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\| + \tfrac{1}{6} B_\tau^{(3)} \Big) + \sum_{\ell=1}^{n-1} (n-\ell)\tau^4 \Big(\tfrac{1}{2} m_3 \max_{0\le s\le T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\| + \tfrac{1}{12} B_T^{(4)} \Big),$$

which completes the proof by using $\sum_{\ell=1}^{n-1}(n-\ell) \le \frac{1}{2}n^2$. □

A close inspection of the proof shows that for the starting value $\mathbf{q}_1$ we only need the estimate $\|\mathbf{e}_1\| \le C_1\tau^3$. Hence, for other starting values satisfying the same bound with a different constant C_1 we would obtain similar results; see also Section 3.5.1.

We now turn towards an error bound under the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$.

Theorem 3.41. *Let the assumptions of Theorem 3.40 be satisfied. Then, for $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ and $t_n \le T$ we have for the approximations $\mathbf{q}_n$ of the scheme (3.1)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \le \min\{T, c_{\mathrm{inv}} m_1^{-1/2}\} \Big(\widetilde{C}_1 + \widetilde{C}_T T \Big) \tau^2 \tag{3.67a}$$

with

$$\widetilde{C}_1 = \widetilde{m}_3 \max_{0\le s\le \tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\| + \tfrac{1}{6} B_\tau^{(3)}, \qquad \widetilde{C}_T = \widetilde{m}_3 \max_{0\le s\le T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\| + \tfrac{1}{12} B_T^{(4)}. \tag{3.67b}$$

where $B_\tau^{(3)}$, $B_T^{(4)}$ are given by (3.50) and $\widetilde{m}_3$ in (3.17).

We observe that, in contrast to the error bound in Theorem 3.40, the error bound (3.67) only grows linearly in time if $\mathbf{L}$ is positive definite, since the minimum is bounded by $c_{\mathrm{inv}} m_1^{-1/2}$ for T large enough.

Proof. In contrast to the previous proof we now have that $\widehat{\boldsymbol{\Psi}}$ is nonsingular because of $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ and (3.16). Thus, we rewrite (3.63) as

$$\|\mathbf{e}_n\| \le \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}}\, \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_{0,*}\| + \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_{0,+}^{(3)}\| + \sum_{\ell=1}^{n-1} \Big(\|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}}\, \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_\ell\| + \|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_\ell^{(4)}\| \Big).$$

With Lemma 3.14 and (3.65) we then have on the one hand

$$\|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_\ell\| = \tau^2 \|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Upsilon} \mathbf{L} (\widetilde{\mathbf{q}}_{\ell+1} - 2\widetilde{\mathbf{q}}_\ell + \widetilde{\mathbf{q}}_{\ell-1})\| \le \tau^4 \widetilde{m}_3 \max_{0\le s\le T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\| \tag{3.68a}$$

as well as

$$\|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_{0,*}\| = \tau^2 \|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Upsilon} \mathbf{L} (\widetilde{\mathbf{q}}_1 - \widetilde{\mathbf{q}}_0)\| \le \tau^3 \widetilde{m}_3 \max_{0\le s\le \tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\|. \tag{3.68b}$$

On the other hand we get due to $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, (3.34a), (3.19), and (3.36a)

$$\tau \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}} \mathbf{q}\| \le \min\{t_n, c_{\mathrm{inv}} m_1^{-1/2}\} \|\mathbf{q}\|.$$

Combining these yields together with (3.64) the bound (3.67). □

We note that the estimates of the defects $\boldsymbol{\Delta}_n$ and $\boldsymbol{\Delta}_{0,*}$ could also have been done as in the proof of Theorem 3.40 by employing (3.66) instead of (3.68). Clearly, we then have to replace the bound (3.36a) by (3.36b). However, for the error bound in the discrete energy norm $|||\cdot|||_\tau$ under the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, which we show next, the approach as used in the proof is favorable.

Theorem 3.42. *Let the assumptions of Theorem 3.40 be satisfied. Then, for $\tau \le \widehat{\tau}_{\mathrm{SSR}}$ and $t_{n+1} \le T$ we have for the approximations $\mathbf{q}_n$ of the scheme (3.1)*

$$|||\mathbf{q}(t_n) - \mathbf{q}_n|||_\tau \le (1 + m_1^{-1})^{1/2} \big(\max\{C_1, \widetilde{C}_1\} + \max\{C_T, \widetilde{C}_T\}\, T\big)\tau^2 \tag{3.69}$$

with C_1, C_T and $\widetilde{C}_1, \widetilde{C}_T$ given as in Theorems 3.40 and 3.41, respectively.

As for the error bound in the standard norm under the stronger step-size restriction we obtain in the discrete energy norm bounds which grow only linearly in time. Moreover, the bounds for the defects are (almost) the same as before.

Proof. We consider the two terms in the discrete energy norm separately. For the difference of errors, equation (3.59) in Lemma 3.38 (with $\mathbf{r}_\ell = 0$) yields with (3.19)

$$\|\partial_\tau \mathbf{e}_n\| \le \tfrac{1}{\tau}\|\mathbf{e}_1\| + \sum_{\ell=1}^{n} \tfrac{1}{\tau}\|\mathbf{d}_\ell\| \le \|\boldsymbol{\Delta}_{0,*}\| + \|\boldsymbol{\delta}_{0,+}^{(3)}\| + \sum_{\ell=1}^{n-1} \Big(\|\boldsymbol{\Delta}_\ell\| + \|\boldsymbol{\delta}_\ell^{(4)}\|\Big) \le (C_1 + C_T t_n)\tau^2$$

by estimating the defects as in the proof of Theorem 3.40.

The second term can be estimated similarly as $\|\mathbf{e}_n\|$ in the proof of Theorem 3.41. Using (3.58) we have with (3.35), (3.64), and (3.68)

$$\begin{aligned}
\|\mathbf{e}_n\|_{\mathbf{L}} &\le \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}}\, \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_{0,*}\|_{\mathbf{L}} + \|\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_{0,+}^{(3)}\|_{\mathbf{L}} + \sum_{\ell=1}^{n-1} \Big(\|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}}\, \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_\ell\|_{\mathbf{L}} + \|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}} \boldsymbol{\delta}_\ell^{(4)}\|_{\mathbf{L}}\Big) \\
&\le m_1^{-1/2} \tfrac{1}{\tau} \Big(\|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_{0,*}\| + \|\boldsymbol{\delta}_{0,+}^{(3)}\| + \sum_{\ell=1}^{n-1} \Big(\|\widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Delta}_\ell\| + \|\boldsymbol{\delta}_\ell^{(4)}\|\Big)\Big) \\
&\le m_1^{-1/2} (\widetilde{C}_1 + \widetilde{C}_T t_n)\tau^2.
\end{aligned}$$

Combining both estimates completes the proof. □

A crucial point in the proof of the bound for $\|\mathbf{e}_n\|_{\mathbf{L}}$ is that we employ the estimate (3.35) for every term. If we estimated $\boldsymbol{\Delta}_n$ and $\boldsymbol{\Delta}_{0,*}$ with (3.66), we would have to use (3.34a) to get the bound $\tau\|\boldsymbol{\mathcal{S}}_n \mathbf{q}\|_{\mathbf{L}} \le t_n \|\mathbf{q}\|_{\mathbf{L}}$, since we do not have a bound of the form $\|\boldsymbol{\mathcal{S}}_n \mathbf{q}\|_{\mathbf{L}} \le c\|\mathbf{q}\|$ with a constant $c > 0$ which is independent of $\widehat{\beta}_\Psi$; see estimate (3.37). This, however, would lead to worse error constants (quadratic instead of linear growth in time) and additional factors $\mathbf{L}$ in front of derivatives of the exact solution. The same problem would occur if we used the variant of the defects in (3.57) and (3.61); cf. Remark 3.37.

For the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ it would be possible to derive an error bound, but it suffers from the same problems as described before. From the proof we see that, similarly to the stability results in Theorem 3.28, only $\|\mathbf{e}_n\|_{\mathbf{L}}$ is responsible for the bad behavior. In fact, the first term $\|\partial_\tau \mathbf{e}_n\|$ can bounded with the same constants under the weaker step-size restriction.

Remark 3.43. If we assume $\mathbf{g} \in C^2([0,T])$, we can even avoid the terms $\|\mathbf{L}\ddot{\mathbf{q}}(s)\|$ and $\|\mathbf{L}\dot{\mathbf{q}}(s)\|$ occurring in (3.62b) as well as in (3.67b) due to the terms $\boldsymbol{\Delta}_n$ and $\boldsymbol{\Delta}_{0,*}$, respectively. This can be seen if instead of the estimate (3.65) for $\boldsymbol{\Delta}_\ell$ we use the following estimate by employing the differential equation (2.12)

$$\begin{aligned}\|\mathbf{L}(\widetilde{\mathbf{q}}_{\ell+1} - 2\widetilde{\mathbf{q}}_\ell + \widetilde{\mathbf{q}}_{\ell-1})\| &\le \|\ddot{\mathbf{q}}(t_{\ell+1}) - 2\ddot{\mathbf{q}}(t_\ell) + \ddot{\mathbf{q}}(t_{\ell-1})\| + \|\mathbf{g}(t_{\ell+1}) - 2\mathbf{g}(t_\ell) + \mathbf{g}(t_{\ell-1})\| \\ &\le \tau^2 \Big(B_T^{(4)} + \max_{0\le s\le T} \|\ddot{\mathbf{g}}(s)\| \Big),\end{aligned}$$

and similarly for $\boldsymbol{\Delta}_{0,*}$. In Chapter 5 we show another approach to avoid (at least) the term $\|\mathbf{L}\ddot{\mathbf{q}}(s)\|$ if we measure the error in the standard norm. ◇

3.4.3. Error analysis for semilinear problems

After this extensive error analysis for linear problems we turn towards error bounds in the standard norm $\|\cdot\|$ for the scheme (3.1) applied to semilinear problems (2.1). Recall that we have two conditions on the Lipschitz continuity of $\mathbf{g}$, which differ in the considered norms, see Assumptions 2.3 and 2.9.

We start with the first one.

Theorem 3.44. *Let Assumptions 3.2 and 3.16 as well as Assumption 2.3 on* $\mathbf{g}$ *hold. Further, assume that for* $T \in (0, t_*)$ *the solution* $\mathbf{q}$ *of* (2.1) *satisfies* $\mathbf{q} \in C^4([0,T])$. *Then, there exists a* $\tau_* > 0$ *such that for* $\tau \le \min\{\tau_*, \tau_{\mathrm{SSR}}\}$ *and* $t_n \le T$ *we have for the approximations* $\mathbf{q}_n$ *of the scheme* (3.1)

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \le T\big(C_1 + \tfrac{1}{2} C_T T\big) \mathrm{e}^{\mathcal{L}_{\mathbf{g}}^{1/2} T}\, \tau^2, \tag{3.70}$$

where the constants C_1, C_T *are defined as in* (3.62b). *Moreover, the Lipschitz constant* $\mathcal{L}_{\mathbf{g}}$ *only depends on* T, *the exact solution* $\mathbf{q}$, *and* $\mathbf{g}$.

We emphasize that in numerical experiments the additional step-size restriction $\tau \le \tau_*$ coming from the local Lipschitz continuity is usually not visible, since the step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ is often much more restrictive. For (globally) Lipschitz continuous functions $\mathbf{g}$ the theorem even holds without this additional restriction; see [CHS20], where a similar result for Lipschitz continuous $\mathbf{g}$ is shown.

Proof. First, we observe that Assumption 2.3 yields the existence of a radius $\rho > 0$ depending only on T and $\{\mathbf{q}(t) \mid t \in [0,T]\}$ such that for all $t \in [0,T]$ and $\widehat{\mathbf{q}} \in \mathbb{R}^m$ satisfying $\|\mathbf{q}(t) - \widehat{\mathbf{q}}\| \le \rho$ for some $t \in [0,T]$ it holds

$$\|\mathbf{g}(t, \mathbf{q}(t)) - \mathbf{g}(t, \widehat{\mathbf{q}})\| \le \mathcal{L}_{\mathbf{g}} \|\mathbf{q}(t) - \widehat{\mathbf{q}}\|,$$

where $\mathcal{L}_{\mathbf{g}} = \mathcal{L}_{\mathbf{g}}(\rho) > 0$ only depends on ρ and, thus, on T and the exact solution of (2.1). In particular, there exists $\tau_* > 0$ such that

$$T\big(C_1 + \tfrac{1}{2} C_T T\big) \mathrm{e}^{\mathcal{L}_{\mathbf{g}}^{1/2} T}\, \tau^2 \le \rho \qquad \text{for all } \tau \le \tau_*.$$

(i) Let $\tau \le \min\{\tau_*, \tau_{\mathrm{SSR}}\}$. As we show in the second part of the proof, we then have $\|\mathbf{e}_n\| \le \rho$ for all $n \in \mathbb{N}$ with $t_n \le T$. Thus, we can apply the above Lipschitz condition.

From (3.58) we obtain with (3.34a) and (3.19)

$$\|\mathbf{e}_n\| \leq (C_1 t_n + \tfrac{1}{2} C_T t_n^2)\tau^2 + \tau^2 \sum_{\ell=1}^{n-1} (n-\ell)\|\mathbf{r}_\ell\|,$$

where the defects are estimated as in the proof of Theorem 3.40. Using the definition of $\mathbf{r}_n$ in (3.53c) together with the Lipschitz continuity yields $\|\mathbf{r}_n\| \leq \mathcal{L}_{\mathbf{g}}\|\mathbf{e}_n\|$. Hence, an application of the Gronwall Lemma B.19 shows the error bound (3.70).

(ii) It remains to show that $\|\mathbf{e}_n\| \leq \rho$ for all $n \in \mathbb{N}$ with $t_n \leq T$. To do so, we define

$$m_* = \max\{\mathcal{M}\}, \quad \mathcal{M} = \{m \in \mathbb{N}_0 \mid t_m \leq T \text{ and } \|\mathbf{e}_k\| \leq \rho \text{ for all } k = 0, \dots, m\}.$$

We obviously have $m_* \geq 0$ because of $0 \in \mathcal{M}$. Suppose $t_{m_*+1} \leq T$. Then we can apply the first part, since we only need the local Lipschitz continuity up to t_{m_*} due to the explicit treatment of the function $\mathbf{g}$ in the scheme. Thus, we get by definition of τ_* for $\tau \leq \min\{\tau_*, \tau_{\text{SSR}}\}$

$$\|\mathbf{e}_{m_*+1}\| \leq T(C_1 + \tfrac{1}{2} C_T T) \mathrm{e}^{\mathcal{L}_{\mathbf{g}}^{1/2} T} \tau^2 \leq \rho.$$

This is in contradiction to the definition of m_*, since we supposed $t_{m_*+1} \leq T$. Hence, we have shown $\|\mathbf{e}_n\| \leq \rho$ for all $n \in \mathbb{N}$ with $t_n \leq T$, which finishes the proof. □

For the stronger step-size restriction we can show a similar result by replacing the estimates for the defects with those of Theorem 3.41. We skip the details, since we receive – because of the semilinearity – exponential growth in time anyway.

For the weaker assumption on the local Lipschitz continuity (Assumption 2.9) we require – in contrast to the previous error bound – the stronger step-size restriction $\tau \leq \widehat{\tau}_{\text{SSR}}$ to prove convergence of the scheme (3.1). Additionally, $\mathbf{L}$ has to be positive definite; see Section 2.2.2. As mentioned in Remark 2.5 this poses no restriction for semilinear problems (2.1)

Theorem 3.45. *Let $\mathbf{L}$ be positive definite and let Assumption 3.2, 3.16, as well as Assumption 2.9 on $\mathbf{g}$ hold. Further, assume that for $T \in (0, t_*)$ the solution $\mathbf{q}$ of (2.1) satisfies $\mathbf{q} \in C^4([0,T])$. Then, there exists a $\tau_* > 0$ such that for $\tau \leq \min\{\tau_*, \widehat{\tau}_{\text{SSR}}\}$ and $t_n \leq T$ we have for the approximations $\mathbf{q}_n$ of the scheme (3.1)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \leq c_{\text{inv}} m_1^{-1/2} \Big(\widetilde{C}_1 + \widetilde{C}_T T\Big) \mathrm{e}^{m_1^{-1/2} \widehat{\mathcal{L}}_{\mathbf{g}} T} \tau^2, \tag{3.71}$$

where the constants $\widetilde{C}_1$, $\widetilde{C}_T$ are defined as in (3.67b). Moreover, the Lipschitz constant $\widehat{\mathcal{L}}_{\mathbf{g}}$ only depends on T, the exact solution $\mathbf{q}$, and $\mathbf{g}$.

Proof. We only show the estimate for the error, since the remaining part follows as in the previous proof. Again from (3.58) we get with (3.36a) and (3.19)

$$\|\mathbf{e}_n\| \leq c_{\text{inv}} m_1^{-1/2} \Big(\widetilde{C}_1 + \widetilde{C}_T T\Big) \tau^2 + \tau \sum_{\ell=1}^{n-1} m_1^{-1/2} \|\mathbf{r}_\ell\|_{\mathbf{L}^{-1}},$$

where the defects are estimated as in the proof of Theorem 3.41. Further, from the local Lipschitz continuity in Assumption 2.9 we have $\|\mathbf{r}_\ell\|_{\mathbf{L}^{-1}} \leq \widehat{\mathcal{L}}_{\mathbf{g}} \|\mathbf{e}_n\|$ if we assume that $\|\mathbf{q}_n - \widetilde{\mathbf{q}}_n\| \leq \rho$ for all $n \in \mathbb{N}$. Together with the Gronwall Lemma B.18 this yields the error bound (3.71). □

A close inspection of the proof shows why the use of the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ would fail in this situation. In order to apply the Lipschitz estimate (2.9) we have to bound

$$\tau \|\boldsymbol{\mathcal{S}}_{n-\ell} \widehat{\boldsymbol{\Psi}} \mathbf{r}_\ell\| \le C \, \|\mathbf{r}_\ell\|_{\mathbf{L}^{-1}} \qquad \ell = 1, \dots, n-1,$$

with a constant $C > 0$ which is independent of τ, n and $\mathbf{L}$. For the stronger step-size restriction this is possible with $C = m_1^{-1/2}$ as shown in the proof. However, for $\tau \le \tau_{\mathrm{SSR}}$ such a constant does not exist for general τ and $\mathbf{L}$, as one can see by choosing $\mathbf{r}_\ell$ as an eigenvector of $\tau^2\mathbf{L}$ corresponding to an eigenvalue z such that $\Psi(z) \in \{0, 4\}$; see also Lemma 3.23 and Remark 3.25.

3.5. Modifications for improved stability results for linear problems

In the previous two sections we have seen that in general the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ leads to worse constants in the stability and error analysis if the analysis is possible at all. The reason for this undesired behavior is the existence of points $z \in (0, \beta_\Psi^2]$ with $\Psi(z) \in \{0, 4\}$. With the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, which guarantees $\Psi(z) \in (\varepsilon, 4-\varepsilon)$ for $z \in [\delta, \widehat{\beta}_\Psi^2] \cap \mathbb{R}$ and some $\varepsilon, \delta > 0$, we were able to resolve this problem.

In the following we present two ways to regain the favorable stability behavior of the two-step scheme (3.1a) under the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$. The first option consists in a modification of the starting value $\mathbf{q}_1$ given in (3.1b). However, this option can be favorably utilized only for $\mathbf{g} \equiv 0$. As a second option we consider averaged approximations instead of $\mathbf{q}_n$. This ansatz will be shown to be beneficial for general linear problems (2.12).

3.5.1. Influence of starting value

We start with the influence of the starting value $\mathbf{q}_1$ on the stability of the two-step scheme (3.1a). For this we assume $\mathbf{g} \equiv 0$ and $\beta_\Psi < \infty$. The first assumption is motivated by the fact that we cannot change the influence of $\mathbf{g}$ in the two-step scheme (3.1a) with a modification of the starting value. The second assumption is for the sake of presentation and satisfied, for example, if Ψ is a polynomial, e.g., the LFC polynomial (3.3).

If in addition $\mathbf{L}$ is positive definite, we know from Section 2.1 that the exact solution of (2.12) with $\mathbf{g} \equiv 0$ is uniformly bounded in $\|\cdot\|$ as well as in $|||\cdot|||$. For the two-step scheme (3.1a) with starting value (3.1b) we have seen in Section 3.3 that uniform boundedness of the numerical solution $\mathbf{q}_n$ in $\|\cdot\|$ and $|||\cdot|||$ is only possible for $\tau \le \widehat{\tau}_{\mathrm{SSR}}$.

In order to analyze, if for different starting values uniform bounds are possible under the weaker step-size restriction $\tau \le \tau_{\mathrm{SSR}}$, we consider the general variant

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau a(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 - \tfrac{1}{2}\tau^2 b(\tau^2\mathbf{L})\mathbf{L}\mathbf{q}_0 + \tau^3\boldsymbol{\delta}_0 \tag{3.72}$$

with sufficiently smooth functions $a, b\colon [0, \beta_\Psi^2] \to \mathbb{R}$ satisfying $a(0) = b(0) = 1$ and a bounded perturbation $\boldsymbol{\delta}_0 \in \mathbb{R}^m$ with $\|\boldsymbol{\delta}_0\| \le C_\delta$. As for $\widehat{\Psi}$, the conditions on a and b are necessary for second-order consistency.

Example 3.46. The general starting value comprises among others

(i) the starting value (3.1b) with $a = b = \widehat{\Psi}$ and $\boldsymbol{\delta}_0 = 0$,

(ii) the starting value (2.20b) for the leapfrog scheme (second-order Taylor polynomial) with $a \equiv b \equiv 1$ and $\boldsymbol{\delta}_0 = 0$, and

(iii) the exact starting value $\mathbf{q}_1 = \mathbf{q}(\tau)$ with again $a \equiv b \equiv 1$ and $\tau^3\boldsymbol{\delta}_0 = \boldsymbol{\delta}_{0,+}^{(3)}$. ◇

As a first step we include (3.72) into the representation formula (3.23a) for the two-step scheme (3.1a), similarly to what we have done for (3.1b).

Lemma 3.47. *Let $\tau \le \tau_{\mathrm{SSR}}$ and $\mathbf{g} \equiv 0$. For the approximations of the two-step scheme (3.1a) with the general starting value (3.72) we have*

$$\mathbf{q}_n = \Big(\cos(n\mathbf{\Phi}) + \boldsymbol{\mathcal{S}}_n\big(\mathbf{I} - \tfrac{1}{2}\tau^2 b(\tau^2\mathbf{L})\mathbf{L} - \cos\mathbf{\Phi}\big)\Big)\mathbf{q}_0 + \tau\boldsymbol{\mathcal{S}}_n a(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 + \tau^3\boldsymbol{\mathcal{S}}_n\boldsymbol{\delta}_0 \tag{3.73}$$

where $\mathbf{\Phi}$ and $\boldsymbol{\mathcal{S}}_n$ are given as in Theorem 3.18.

Proof. The formula (3.73) follows from (3.23a) by inserting the general starting value (3.72). □

The problematic terms in (3.73) are those with a factor $\boldsymbol{\mathcal{S}}_n$ because for $\tau \le \tau_{\mathrm{SSR}}$ we only have $\|\boldsymbol{\mathcal{S}}_n\| \le n$; see Lemma 3.23. Thus, to get rid of this undesired linear growth, we have to show that for specific choices of the functions a and b as well as of the perturbation $\boldsymbol{\delta}_0$ these terms are uniformly bounded in time.

For the term $\tau^3\boldsymbol{\mathcal{S}}_n\boldsymbol{\delta}_0$ the simplest and most efficient choice is $\boldsymbol{\delta}_0 = 0$. For the term containing $\mathbf{q}_0$ we recall that $\mathbf{I} - \frac{1}{2}\mathbf{\Psi} = \cos\mathbf{\Phi}$. Thus, the choice $b(\tau^2\mathbf{L}) = \widehat{\mathbf{\Psi}}$ leads to

$$\big(\mathbf{I} - \tfrac{1}{2}\tau^2 b(\tau^2\mathbf{L})\mathbf{L} - \cos\mathbf{\Phi}\big) = 0$$

because of (3.2); see also Theorem 3.18 and Corollary 3.20.

It remains to bound $\tau\boldsymbol{\mathcal{S}}_n a(\tau^2\mathbf{L})\dot{\mathbf{q}}_0$. For this, we recall that for $\tau \le \tau_{\mathrm{SSR}}$ we have

$$\boldsymbol{\mathcal{S}}_n = \frac{\sin(n\mathbf{\Phi})}{\sin\mathbf{\Phi}} \qquad \text{with} \quad \sin\mathbf{\Phi} = \big(\mathbf{\Psi}(\mathbf{I} - \tfrac{1}{4}\mathbf{\Psi})\big)^{1/2};$$

cf. (3.23). Hence, the function a should be chosen in such a way that possible zero eigenvalues of $\sin\mathbf{\Phi}$, which stem from those eigenvalues $z \in (0, \beta_\Psi^2]$ of $\tau^2\mathbf{L}$ with $\Psi(z) \in \{0, 4\}$, are compensated. Note that for $z = 0$ we have $\Psi(z) = 0$, which is required to obtain second-order convergence; see Sections 3.1.2 and 3.4. Hence, we restrict ourselves to positive definite matrices $\mathbf{L}$ in the following considerations.

To eliminate the effect caused by the undesired zero eigenvalues of $\sin\mathbf{\Phi}$ we observe that, if $\Psi(z) \in \{0, 4\}$ for some $z \in (0, \beta_\Psi^2)$, Ψ has a local extremum point in z because of Definition 3.9(a) of β_Ψ^2, i.e., $\Psi'(z) = 0$. Hence, we expect that the choice $a = \Psi'$ compensates these zero eigenvalues of $\sin\mathbf{\Phi}$. The details are given in Lemma 3.48 below. Note that $\Psi'(0) = 1$ due to the consistency conditions (3.10).

We emphasize that there exist other choices for a which compensate the zero eigenvalues of $\sin\mathbf{\Phi}$. One of these choices is $a(\tau^2\mathbf{L}) = \widehat{\mathbf{\Psi}}(\mathbf{I} - \frac{1}{4}\mathbf{\Psi}) = (\tau^2\mathbf{L})^{-1}\sin(\mathbf{\Phi})^2$, which yields

$$\tau\boldsymbol{\mathcal{S}}_n a(\tau^2\mathbf{L}) = \tau(\tau^2\mathbf{L})^{-1}\sin(n\mathbf{\Phi})\sin(\mathbf{\Phi}) = \mathbf{L}^{-1/2}\sin(n\mathbf{\Phi})\widehat{\mathbf{\Psi}}^{1/2}\big(\mathbf{I} - \tfrac{1}{4}\mathbf{\Psi}\big)^{1/2}.$$

Since this option implicitly occurs in the next subsection, we focus here on $a = \Psi'$.

Combining these considerations leads to the following choice for the starting value

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau\Psi'(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 - \tfrac{1}{2}\tau^2\widehat{\Psi}(\tau^2\mathbf{L})\mathbf{L}\mathbf{q}_0 = \mathbf{q}_0 + \tau\mathbf{\Psi}'\dot{\mathbf{q}}_0 - \tfrac{1}{2}\mathbf{\Psi}\mathbf{q}_0. \tag{3.74}$$

Moreover, we have the following estimates for the matrix function $\boldsymbol{\mathcal{S}}_n\mathbf{\Psi}'$.

Lemma 3.48. *Let $\beta_\Psi < \infty$ and $\tau \le \vartheta\tau_{\mathrm{SSR}}$ for an arbitrary $\vartheta \in (0,1)$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \boldsymbol{\Psi}' \mathbf{q} \right\|_{\mathbf{L}} \le C^\star(\vartheta, \Psi) \|\mathbf{q}\|, \tag{3.75a}$$

and, if $\mathbf{L}$ is positive definite,

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \boldsymbol{\Psi}' \mathbf{q} \right\| \le c_{\mathrm{inv}}\, C^\star(\vartheta, \Psi) \|\mathbf{q}\|, \tag{3.75b}$$

where the constant $C^\star(\vartheta, \Psi)$ only depends on ϑ and Ψ but not on $\mathbf{L}$ or τ.

Proof. The main work in this proof consist of showing that the function $\Gamma\colon [0, \vartheta^2\beta_\Psi^2] \to \mathbb{R}$, defined by

$$\Gamma(z) = \frac{z\,(\Psi'(z))^2}{\Psi(z)(1 - \frac{1}{4}\Psi(z))}, \tag{3.76}$$

is well-defined, continuous, and bounded by a constant which only depends on ϑ and Ψ. Using this we then obtain together with the definition (3.23b) of $\sin \boldsymbol{\Phi}$ that

$$\tau^2 \mathbf{L}\, \boldsymbol{\mathcal{S}}_n^2 (\boldsymbol{\Psi}')^2 = \sin(n\boldsymbol{\Phi})^2 \Gamma(\tau^2 \mathbf{L}).$$

Taking the square root yields (3.75), where for the second estimate we additionally use (2.13).

As already indicated in the preliminary considerations we have that if $z_* \in (0, \beta_\Psi^2)$ satisfies $\Psi(z_*) \in \{0, 4\}$, then Ψ has a local extremum point at z_* due to the definition of β_Ψ and, thus, $\Psi'(z_*) = 0$. L'Hôpital's rule applied to Γ then yields

$$\Gamma(z_*) = \lim_{z \to z_*} \frac{z(\Psi'(z))^2}{\Psi(z)(1 - \frac{1}{4}\Psi(z))} = \lim_{z \to z_*} \frac{\Psi'(z) + 2z\Psi''(z)}{1 - \frac{1}{2}\Psi(z)} = \pm 2 z_* \Psi''(z_*),$$

depending on whether $\Psi(z_*) = 0$ $(+)$ or $\Psi(z_*) = 4$ $(-)$. Note that this expression is always positive, since for $\Psi(z_*) = 4$ we have a local extremum, meaning that $\Psi''(z_*) \le 0$, and vice versa for $\Psi(z_*) = 0$. Hence, Γ is well-defined and continuous in $(0, \beta_\Psi^2)$.

Further, we have that $\Gamma(0) = 1$ because of Assumption 3.2 and the consistency conditions (3.10). For $z = \beta_\Psi^2$ we have (unfortunately) $\lim_{z \to \beta_\Psi^2} \Gamma(z) = \infty$, since $\Psi(\beta_\Psi^2) \in \{0, 4\}$ due to the definition of β_Ψ and $\beta_\Psi^2 < \infty$. Altogether, we get that Γ is continuous on $[0, \vartheta^2\beta_\Psi^2]$ for every $\vartheta \in (0,1)$. In particular, there exists a constant $C = C(\vartheta, \Psi)$ such that $\|\Gamma(\tau^2 \mathbf{L})\| \le C$ for $\tau \le \vartheta\tau_{\mathrm{SSR}}$. This finishes the proof. □

With these estimates we are in a position to state the stability of the scheme (3.1a) combined with the starting value (3.74).

Theorem 3.49. *Let $\beta_\Psi < \infty$, $\tau \le \vartheta\tau_{\mathrm{SSR}}$ for some $\vartheta \in (0,1)$, and $\mathbf{g} \equiv 0$. Then the approximations obtained by (3.1a) with the special starting value (3.74) satisfy for $n = 0, 1, \ldots$,*

$$\|\mathbf{q}_n\| \le \|\mathbf{q}_0\| + \min\{\Psi'_{\max} t_n, c_{\mathrm{inv}}\, C^\star(\vartheta, \Psi)\} \|\dot{\mathbf{q}}_0\|, \tag{3.77a}$$

$$|||\mathbf{q}_n|||_\tau \le 2^{1/2} \Big(\|\mathbf{q}_0\|_{\mathbf{L}} + \max\{\Psi'_{\max}, C^\star(\vartheta, \Psi)\} \|\dot{\mathbf{q}}_0\| \Big), \tag{3.77b}$$

where $\Psi'_{\max} = \max_{z \in [0, \beta_\Psi^2]} |\Psi'(z)|$. For $\tau = \tau_{\mathrm{SSR}}$ the bound (3.77a) holds with $\Psi'_{\max} t_n$ instead of the minimum.

We observe that, in contrast to Theorems 3.27 and 3.28, we already obtain under the step-size restriction $\tau \leq \vartheta\tau_{\mathrm{SSR}}$, $\vartheta \in (0,1)$, uniform bounds of $\mathbf{q}_n$ in the standard norm $\|\cdot\|$ for positive definite $\mathbf{L}$ and in the discrete energy norm $|||\cdot|||_\tau$ for positive semidefinite $\mathbf{L}$. However, the bounds deteriorate as $\tau \to \tau_{\mathrm{SSR}}$, since $C^\star(\vartheta,\Psi) \to \infty$ for $\vartheta \to 1$, which can be seen from the proof of Lemma 3.48.

As we have indicated in our preliminary considerations, almost all other choices for the starting value, even the exact solution, fail in giving such uniformly bounded approximations for certain step sizes. This can be easily shown with the aid of Lemma 3.47 (see also Example 3.46) and it is confirmed by some numerical experiments in Section 4.5 for the LFC polynomials (3.3). For these polynomials explicit values of $\Psi'_{\max}$ and $C^\star(\vartheta,\Psi)$ are stated in Theorem 4.7 and Conjecture 4.8, respectively.

Proof of Theorem 3.49. Inserting (3.74) in (3.23a) and using (3.23b) yields for $\tau \leq \tau_{\mathrm{SSR}}$

$$\mathbf{q}_n = \cos(n\mathbf{\Phi})\mathbf{q}_0 + \tau\boldsymbol{\mathcal{S}}_n\mathbf{\Psi}'\dot{\mathbf{q}}_0. \tag{3.78}$$

The first bound in (3.77a) involving t_n follows from (3.34a). If additionally $\tau \leq \vartheta\tau_{\mathrm{SSR}}$ and $\mathbf{L}$ is positive definite, we apply (3.75b) to get the uniform bound in (3.77a).

For the estimate in the discrete energy norm we observe that analogously to the proof of Corollary 3.21 and Theorem 3.28 we have for $\tau \leq \tau_{\mathrm{SSR}}$

$$\|\partial_\tau \mathbf{q}_n\| \leq \|\mathbf{q}_0\|_{\mathbf{L}} + \|\mathbf{\Psi}'\dot{\mathbf{q}}_0\| \leq \|\mathbf{q}_0\|_{\mathbf{L}} + \Psi'_{\max}\|\dot{\mathbf{q}}_0\|.$$

Further, for $\tau \leq \vartheta\tau_{\mathrm{SSR}}$ we have with (3.75a) that $\|\mathbf{q}_n\|_{\mathbf{L}} \leq \|\mathbf{q}_0\|_{\mathbf{L}} + C^\star(\vartheta,\Psi)\|\dot{\mathbf{q}}_0\|$. A combination of both then yields the bound (3.77b). □

Remark 3.50. If we use the special starting value (3.74) instead of (3.1b) and replace the discrete quantity $\mathcal{M}_{\mathbf{q},0}$ in Definition 3.29 by

$$\tau^2\mathcal{M}_{\mathbf{q},0} \to \tau^2\|\Psi'\dot{\mathbf{q}}_0\|^2 + (\mathbf{\Psi}\mathbf{q}_0,\mathbf{q}_0) - \tfrac{1}{4}\|\mathbf{\Psi}\mathbf{q}_0\|^2,$$

one can show that Theorem 3.32 also holds uniformly in time for $\tau \leq \vartheta\tau_{\mathrm{SSR}}$ and $\vartheta \in (0,1)$. ◇

Unfortunately, we are not able to transfer this good stability behavior of the two-step scheme (3.1a) with the special starting value (3.74) to the error analysis. Proceeding as usual one encounters the problem that the defects $\mathbf{d}_n$ in Lemma 3.36 of the two-step scheme (3.1a) do not permit a factor $\mathbf{\Psi}'$. Hence, the above bounds cannot be applied.

Nevertheless, we are still able to show bounds as in Theorems 3.40 and 3.42. In particular, we can directly apply these theorems if we can show that the special starting value (3.74) yields $\|\mathbf{e}_1\| \leq C\tau^3$ with a constant C independent of τ, n and $\|\mathbf{L}\|$. To do so, we require a similar definition for Ψ' as Definition 3.13 for Ψ.

Definition 3.51. *We define $m_3^\star \geq 0$ as the smallest constant such that*

$$|\Psi'(z) - 1| \leq m_3^\star z, \qquad \text{for all } z \in [0,\beta_\Psi^2]. \tag{3.79}$$

Again, the existence of $m_3^\star$ is guaranteed by the consistency conditions (3.10) and by (3.11) due to $\beta_\Psi < \infty$. With this we can show a bound for $\mathbf{e}_1 = \mathbf{q}(\tau) - \mathbf{q}_1$, where we use the notation of the previous section.

Lemma 3.52. *Let $\mathbf{q} \in C^4([0,T])$ be the solution of (2.12) with $\mathbf{g} \equiv 0$. The specific starting value $\mathbf{q}_1$ given in (3.74) satisfies*

$$\|\mathbf{q}(\tau) - \mathbf{q}_1\| \leq \widehat{C}_1 \tau^3, \qquad \widehat{C}_1 = \tfrac{1}{4}\tau\, m_3 B_0^{(4)} + m_3^\star B_0^{(3)} + \tfrac{1}{6} B_\tau^{(3)}. \tag{3.80}$$

where $B_0^{(3)}$, $B_0^{(4)}$, $B_\tau^{(3)}$ are given by (3.50) and m_3, $m_3^\star$ in (3.15) and (3.79), respectively.

Proof. Inserting the exact solution into the starting value (3.74) and Taylor expansion yields for the emerging defect

$$\mathbf{e}_1 = \widehat{\mathbf{d}}_0 = \widetilde{\mathbf{q}}_1 - \widetilde{\mathbf{q}}_0 - \tau \boldsymbol{\Psi}'\dot{\mathbf{q}}(0) + \tfrac{1}{2}\boldsymbol{\Psi}\widetilde{\mathbf{q}}_0 = \tfrac{1}{2}(\boldsymbol{\Psi} - \tau^2\mathbf{L})\widetilde{\mathbf{q}}_0 + \tau(\mathbf{I} - \boldsymbol{\Psi}')\dot{\mathbf{q}}(0) + \boldsymbol{\delta}_{0,+}^{(3)}. \tag{3.81}$$

Thus, we obtain with (3.14), (3.15), (3.79), and (3.52)

$$\|\mathbf{e}_1\| = \|\widehat{\mathbf{d}}_0\| \leq \tfrac{1}{4} m_3 \tau^4 \|\mathbf{L}^2 \widetilde{\mathbf{q}}_0\| + m_3^\star \tau^3 \|\mathbf{L}\dot{\mathbf{q}}(0)\| + \tfrac{1}{6}\tau^3 \max_{0\leq s\leq \tau} \|\mathbf{q}^{(3)}(s)\|.$$

Employing $\ddot{\mathbf{q}} = -\mathbf{L}\mathbf{q}$ and (3.50) completes the proof. □

We conclude this section by showing in a refined analysis that under additional assumptions the scheme (3.1a) with special starting value (3.74) converges with order four for homogeneous problems. For this, we require the following additional definition.

Definition 3.53. *Let $m_3' = -\Psi''(0)$. Then we define $m_4, m_4^\star \geq 0$ as the smallest constants such that*

$$\left|\Psi(z) - z + \tfrac{1}{2}m_3' z^2\right| \leq m_4 z^3, \quad \left|\Psi'(z) - 1 + m_3' z\right| \leq 3 m_4^\star z^2 \qquad \text{for all } z \in [0, \beta_\Psi^2]. \tag{3.82}$$

Note that we have $m_3' \geq 0$ or, equivalently, $\Psi''(0) \leq 0$ because of (3.18) and the consistency conditions (3.10). The factor 3 in the second estimate is motivated by the derivative of z^3. We see in the next chapter that the LFC polynomials (3.3) satisfy these bounds with $m_4 = m_4^\star$.

Theorem 3.54. *Let $\beta_\Psi < \infty$ and let Assumptions 3.2 and 3.16 hold. Further, assume that the solution $\mathbf{q}$ of (2.12) with $\mathbf{g} \equiv 0$ satisfies $\mathbf{q} \in C^6([0,T])$. Then, for $\tau \leq \tau_{\mathrm{SSR}}$ and $t_n \leq T$ we have for the approximations $\mathbf{q}_n$ of the two-step scheme (3.1a) with special starting value (3.74)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \leq T(M_3 C_\diamondsuit \tau^2 + C_\triangle \tau^4), \tag{3.83a}$$

where $M_3 = \frac{1}{24}|1 - 6m_3'|$ and

$$C_\diamondsuit = 4B_0^{(3)} + \tau B_0^{(4)} + T B_n^{(4)}, \quad C_\triangle = (3m_4^\star + \tfrac{1}{120})B_\tau^{(5)} + \tfrac{1}{2}\tau m_4 B_0^{(6)} + \tfrac{1}{2}T(m_4 + \tfrac{1}{360})B_T^{(6)}. \tag{3.83b}$$

The theorem shows that the two-step scheme (3.1a) with special starting value (3.74) applied to a linear problem with $\mathbf{g} \equiv 0$ is in general of order two unless we have $M_3 = 0$, i.e., $m_3' = \frac{1}{6}$. In this case we obtain a fourth-order scheme. For the leapfrog scheme we have $M_3 = \frac{1}{24}$ because of $m_3' = 0$. Hence, for functions Ψ with $m_3' \in (0, \frac{1}{3})$ we get that the constant M_3 is smaller than for the leapfrog scheme, which in general leads to smaller errors for sufficiently smooth functions. This is confirmed by the numerical examples in the next chapter for the LFC polynomials (3.3).

Example 3.55. For $\mathbf{g} \equiv 0$ and $\Psi(z) = z - \frac{1}{12}z^2$ the general two-step scheme (3.1a) comprises the modified (equation) leapfrog method, which is of order four [SB87]. This is confirmed by the above theorem because we have $m_3' = \frac{1}{6}$ and $m_4 = m_4^\star = 0$. ◇

Proof of Theorem 3.54. The main task consists in deriving the defects and the corresponding bounds in the refined framework. We first observe that Definition 3.53 allows us to define functions $\Theta, \Theta^\star \colon [0, \beta_\Psi^2] \to \mathbb{R}$ satisfying

$$\Theta(z) = \frac{\Psi(z) - z + \frac{1}{2}m_3' z^2}{z^3}, \qquad |\Theta(z)| \le m_4, \tag{3.84a}$$

$$\Theta^\star(z) = \frac{\Psi'(z) - 1 + m_3' z}{z^2}, \qquad |\Theta^\star(z)| \le 3m_4^\star. \tag{3.84b}$$

With these functions, we write the defect in (3.54) for the two-step scheme (3.1a) again with Taylor expansion and $\ddot{\mathbf{q}} = -\mathbf{L}\mathbf{q}$ as

$$\begin{aligned}\mathbf{d}_n &= (\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1}) + \tau^2 \widehat{\mathbf{\Psi}}\mathbf{L}\widetilde{\mathbf{q}}_n = (\mathbf{\Psi} - \tau^2\mathbf{L} + \tfrac{1}{12}\tau^4\mathbf{L}^2)\widetilde{\mathbf{q}}_n + \boldsymbol{\delta}_{n,+}^{(6)} - \boldsymbol{\delta}_{n,-}^{(6)} \\ &= \tfrac{1}{12}\tau^4(1 - 6m_3')\mathbf{L}^2\widetilde{\mathbf{q}}_n + \tau^6\Theta(\tau^2\mathbf{L})\mathbf{L}^3\widetilde{\mathbf{q}}_n + \boldsymbol{\delta}_{n,+}^{(6)} - \boldsymbol{\delta}_{n,-}^{(6)}\end{aligned}$$

and the error $\mathbf{e}_1 = \widehat{\mathbf{d}}_0$ in (3.81) as

$$\begin{aligned}\mathbf{e}_1 = \widehat{\mathbf{d}}_0 &= \tfrac{1}{2}(\mathbf{\Psi} - \tau^2\mathbf{L} + \tfrac{1}{12}\tau^4\mathbf{L}^2)\widetilde{\mathbf{q}}_0 - \tau(\mathbf{\Psi}' - \mathbf{I} + \tfrac{1}{6}\tau^2\mathbf{L})\dot{\mathbf{q}}(0) + \boldsymbol{\delta}_{n,+}^{(5)} \\ &= \tfrac{1}{24}\tau^4(1 - 6m_3')\mathbf{L}^2\widetilde{\mathbf{q}}_0 + \tfrac{1}{2}\tau^6\Theta(\tau^2\mathbf{L})\mathbf{L}^3\widetilde{\mathbf{q}}_0 \\ &\quad - \tfrac{1}{6}\tau^3(1 - 6m_3')\mathbf{L}\dot{\mathbf{q}}(0) - \tau^5\Theta^\star(\tau^2\mathbf{L})\mathbf{L}^2\dot{\mathbf{q}}(0) + \boldsymbol{\delta}_{n,+}^{(5)},\end{aligned}$$

where $\boldsymbol{\delta}_{n,+}^{(5)}$, $\boldsymbol{\delta}_{n,+}^{(6)}$, and $\boldsymbol{\delta}_{n,-}^{(6)}$ are defined in (3.51). Applying (3.84), (3.52), and (3.50) then yields

$$\|\mathbf{d}_n\| \le 2M_3\tau^4 B_{t_n}^{(4)} + (m_4 + \tfrac{1}{360})\tau^6 B_{t_{n+1}}^{(6)}$$

and

$$\|\mathbf{e}_1\| \le M_3\tau^4 B_0^{(4)} + \tfrac{1}{2}m_4\tau^6 B_0^{(6)} + 4M_3\tau^3 B_0^{(3)} + (3m_4^\star + \tfrac{1}{120})\tau^5 B_\tau^{(5)}.$$

The representation formula (3.58) for the error $\mathbf{e}_n$ (with $\mathbf{r}_\ell = 0$ for all ℓ due to $\mathbf{g} \equiv 0$) together with (3.34a) completes the proof. □

As before a similar result as in Theorem 3.54 holds for the discrete energy norm under the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$. Further, other starting values for $\mathbf{q}_1$ yielding a fourth-order approximation, e.g., a fourth-order Taylor approximation, could be used as well to obtain an overall fourth-order scheme.

Remark 3.56. The above error analysis for linear, homogeneous problems can easily be generalized to schemes of higher (even) order if the function Ψ satisfies additional consistency conditions. In [GJ08, JR10] such higher-order schemes were constructed without doing a rigorous stability and error analysis and specifying a starting value $\mathbf{q}_1$. ◇

3.5.2. Averaged approximations

The advantage of the special starting value can be favorably employed only for $\mathbf{g} \equiv 0$. In the following we show how to achieve improved stability results for the scheme (3.1) for general linear problems (2.12).

Motivated by the stability and error results for $\partial_\tau \mathbf{q}_n$ in Theorem 3.28 and $\partial_\tau \mathbf{e}_n$ in Theorem 3.42, respectively, we consider an averaged quantity $\mathbf{q}_n^{\mathrm{a}}$ instead of $\mathbf{q}_n$ as the approximation to $\mathbf{q}(t_n)$. More precisely, we define for $n \ge 1$

$$\mathbf{q}_n^{\mathrm{a}} = \tfrac{1}{4}(\mathbf{q}_{n+1} + 2\mathbf{q}_n + \mathbf{q}_{n-1}) \approx \mathbf{q}(t_n). \tag{3.85}$$

For this quantity we obtain the following representation formula.

Corollary 3.57. *Let $\tau \leq \tau_{\mathrm{SSR}}$ and $n \geq 1$. For the approximations $\mathbf{q}_n$ of the general scheme (3.1) we have*

$$\begin{aligned}\mathbf{q}_n^{\mathrm{a}} = \cos(n\boldsymbol{\Phi})\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\mathbf{q}_0 + \tau\boldsymbol{\mathcal{S}}_n\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\widehat{\boldsymbol{\Psi}}\dot{\mathbf{q}}_0 \\ + \tau^2\sum_{\ell=0}^{n-1}\xi_{\ell,n}\,\boldsymbol{\mathcal{S}}_{n-\ell}\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\widehat{\boldsymbol{\Psi}}\mathbf{g}_\ell + \tfrac{1}{4}\tau^2\widehat{\boldsymbol{\Psi}}\mathbf{g}_n,\end{aligned} \tag{3.86}$$

where the coefficients $\xi_{\ell,n}$, $0 \leq \ell \leq n$, are defined as in (3.26).

Proof. From the two-step scheme (3.1a) we obtain by adding $4\mathbf{q}_n$ and dividing by 4

$$\mathbf{q}_n^{\mathrm{a}} = (\mathbf{I} - \tfrac{1}{4}\Psi)\mathbf{q}_n + \tfrac{1}{4}\tau^2\widehat{\Psi}(\tau^2\mathbf{L})\mathbf{g}_n = \cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\mathbf{q}_n + \tfrac{1}{4}\tau^2\widehat{\Psi}(\tau^2\mathbf{L})\mathbf{g}_n, \tag{3.87}$$

where we used that $\mathbf{I} - \frac{1}{4}\Psi = \cos\left(\frac{1}{2}\boldsymbol{\Phi}\right)^2$ because of (3.30). Application of Corollary 3.20 then yields the result.

Alternatively, one could proceed similarly to the proof of Corollary 3.21 by applying the sum-to-product formulae for sine (B.4a) and cosine (B.4b) to the formula (3.27) in Corollary 3.20. □

In contrast to the approximations $\mathbf{q}_n$, we obtain for $\mathbf{q}_n^{\mathrm{a}}$ besides the additional term with $\mathbf{g}_n$ everywhere the factor $\cos(\frac{1}{2}\boldsymbol{\Phi})^2$. As we will see in a moment, this ensures together with $\widehat{\boldsymbol{\Psi}}$ the favorable stability behavior, since it cancels out all possible zeros of $\sin\boldsymbol{\Phi}$ in $\boldsymbol{\mathcal{S}}_n$ which are stemming from the function Ψ.

To state bounds on $\mathbf{q}_n^{\mathrm{a}}$ we first show estimates for the matrix functions occurring in (3.86).

Lemma 3.58. *Let $\tau \leq \tau_{\mathrm{SSR}}$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\tau\left\|\boldsymbol{\mathcal{S}}_n\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\widehat{\boldsymbol{\Psi}}\mathbf{q}\right\| \leq \min\{t_n, c_{\mathrm{inv}}\}\|\mathbf{q}\|, \tag{3.88a}$$

$$\tau\left\|\boldsymbol{\mathcal{S}}_n\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\widehat{\boldsymbol{\Psi}}\mathbf{q}\right\|_{\mathbf{L}} \leq \|\mathbf{q}\|. \tag{3.88b}$$

Proof. From the definition of $\sin\left(\frac{1}{2}\boldsymbol{\Phi}\right)$ in (3.30) we obtain together with (B.2a)

$$\tau\boldsymbol{\mathcal{S}}_n\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)^2\widehat{\boldsymbol{\Psi}}\mathbf{L}^{1/2} = \boldsymbol{\mathcal{S}}_n\sin\boldsymbol{\Phi}\,\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)\widehat{\boldsymbol{\Psi}}^{1/2} = \sin(n\boldsymbol{\Phi})\cos\left(\tfrac{1}{2}\boldsymbol{\Phi}\right)\widehat{\boldsymbol{\Psi}}^{1/2}.$$

Together with (3.19) this yields the second estimate (3.88b). For the first estimate we multiply by $\mathbf{L}^{-1/2}$ if $\mathbf{L}$ is positive definite to get the bound with c_{inv}. Otherwise, we have again with (3.19) the bound with t_n. □

Theorem 3.59. *Let $\tau \leq \tau_{\mathrm{SSR}}$ and $n \geq 1$. The approximations $\mathbf{q}_n$ obtained by the general scheme (3.1) applied to the linear differential equation (2.12) satisfy for*

$$\|\mathbf{q}_n^{\mathrm{a}}\| \leq \|\mathbf{q}_0\| + \min\{t_n, c_{\mathrm{inv}}\}\|\dot{\mathbf{q}}_0\| + \min\{t_n, c_{\mathrm{inv}}\}\,\tau\sum_{\ell=0}^{n}\|\mathbf{g}_\ell\|, \tag{3.89}$$

$$\|\mathbf{q}_n^{\mathrm{a}}\|_{\mathbf{L}} \leq \|\mathbf{q}_0\|_{\mathbf{L}} + \|\dot{\mathbf{q}}_0\| + \tau\sum_{\ell=0}^{n}\|\mathbf{g}_\ell\|. \tag{3.90}$$

Proof. The proof directly follows from Corollary 3.57, Lemma 3.58, and from the estimates

$$\tfrac{1}{4}\tau^2\|\widehat{\boldsymbol{\Psi}}\mathbf{g}_n\| \leq \tfrac{1}{2}\tau\min\{t_n, c_{\mathrm{inv}}\}\|\mathbf{g}_n\| \qquad \text{and} \qquad \tfrac{1}{4}\tau^2\|\widehat{\boldsymbol{\Psi}}\mathbf{g}_n\|_{\mathbf{L}} \leq \tfrac{1}{2}\tau\|\mathbf{g}_n\|$$

because of $\tau \leq t_n$ and (3.11). □

If we combine (3.90) with (3.39a) we have for the discrete energy norm (with $\|\mathbf{q}_n\|_{\mathbf{L}}$ replaced by $\|\mathbf{q}_n^{\mathrm{a}}\|_{\mathbf{L}}$)

$$\Big(\|\mathbf{q}_n^{\mathrm{a}}\|_{\mathbf{L}}^2 + \|\partial_\tau \mathbf{q}_n\|^2\Big)^{1/2} \leq 2^{1/2}\Big(\|\mathbf{q}_0\|_{\mathbf{L}} + \|\dot{\mathbf{q}}_0\| + \tau \sum_{\ell=0}^{n} \|\mathbf{g}_\ell\|\Big).$$

Thus, if $\mathbf{L}$ is positive definite, we have for both the standard as well as the discrete energy norm uniformly bounded constants similarly to the behavior of the exact solution; see (2.15a) and (2.15b). For the standard norm the averaged approximations even have exactly the same stability bound as the exact solution.

Remark 3.60. From (3.87) we observe that instead of the averaged quantity $\mathbf{q}_n^{\mathrm{a}}$ we could also consider the "filtered" quantity $\cos(\frac{1}{2}\mathbf{\Phi})^2\mathbf{q}_n$. In fact, for $\mathbf{g} \equiv 0$ they are identical. Hence, the averaged quantity $\mathbf{q}_n^{\mathrm{a}}$ acts as a filter which removes possible (linear) instabilities occurring for certain step sizes. For $\partial_\tau \mathbf{q}_n$ defined in (3.20a) we have already used a similar effect, since it can be viewed as the filtered quantity (with filter $\widehat{\mathbf{\Psi}}$) of $\mathbf{p}_n$ stemming from the one-step scheme (3.8); see (3.21). ◇

As for the special starting value (3.74) in the last subsection we are not able to transfer this improved stability behavior to the error analysis. More precisely, similarly to $\mathbf{q}_n^{\mathrm{a}}$ we have for the averaged error $\mathbf{e}_n^{\mathrm{a}}$ with Lemma 3.38

$$\mathbf{e}_n^{\mathrm{a}} = \boldsymbol{\mathcal{S}}_n \cos(\tfrac{1}{2}\mathbf{\Phi})^2 \mathbf{e}_1 + \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell} \cos(\tfrac{1}{2}\mathbf{\Phi})^2 (\tau^2 \widehat{\mathbf{\Psi}} \mathbf{r}_\ell + \mathbf{d}_\ell) + \tfrac{1}{4}\tau^2 (\widehat{\mathbf{\Psi}} \mathbf{r}_n + \mathbf{d}_n).$$

Since the term $\mathbf{\Delta}_\ell$ in $\mathbf{d}_\ell$ defined in (3.53b) admits for $\tau \leq \tau_{\mathrm{SSR}}$ no additional factor $\widehat{\mathbf{\Psi}}$ we cannot profit from Lemma 3.58, since not all possible zeros of $\sin \mathbf{\Phi}$ cancel out. As before, this is mainly problematic for the error analysis in the discrete energy norm. Since the analysis is completely analogous to the one in Section 3.4, we skip the details.

3.6. Modified θ-schemes

We conclude this chapter by considering the scheme (3.1) with $\Psi = \Psi_\theta$ given by the rational functions (3.4). For this functions we state explicit values for some of the constants given in Section 3.1.2. Moreover, we show that the stability bounds in Section 3.3, and thus the error bounds, can be improved due to the special structure of Ψ_θ.

As mentioned in Section 3.1 the choice of these functions is motivated by the so-called θ-schemes, which are symmetric two-step methods of second order; see, e.g., [HNW93, Theorems III.10.3 and III.10.5] for conditions to derive these schemes. Applied to (2.1) the θ-schemes read

$$\begin{aligned} \mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = &-\tau^2\Big(\theta \mathbf{L}\mathbf{q}_{n+1} + (1-2\theta)\mathbf{L}\mathbf{q}_n + \theta \mathbf{L}\mathbf{q}_{n-1}\Big) \\ &+ \tau^2\Big(\theta \mathbf{g}_{n+1} + (1-2\theta)\mathbf{g}_n + \theta \mathbf{g}_{n-1}\Big). \end{aligned} \tag{3.91}$$

Obviously, this two-step scheme and (3.1a) differ for $\theta > 0$. In particular, we have to solve a possibly nonlinear system in each time step. We point out that for $\theta = 0$ the scheme (3.91) reduces to the leapfrog scheme (2.20a) and for $\theta = \frac{1}{4}$ the scheme equipped with a suitable starting value is equivalent to the Crank–Nicolson scheme if applied to (2.4).

However, if we apply (3.91) to the linear problem (2.12) with $\mathbf{g} \equiv 0$ we obtain

$$(\mathbf{I} + \tau^2\theta\mathbf{L})(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1}) = -\tau^2\mathbf{L}\mathbf{q}_n, \qquad n = 1, 2, \dots.$$

Multiplying with $(\mathbf{I} + \tau^2\theta\mathbf{L})^{-1}$ then yields the matrix function $\Psi_\theta(\tau^2\mathbf{L})$ with Ψ_θ given in (3.4).

For an inhomogeneity $\mathbf{g}$ the *modified θ-scheme* (3.1a), (3.4) (with a second-order Taylor approximation (2.20b) as starting value) was recently analyzed in [HHW21] for $\theta \in [0, \frac{1}{2}]$ in the setting of a full discretization in space and time for the linear acoustic wave equation. For the space discretization several variants of dG-FEM are used. The original θ-scheme (3.91) is analyzed for $\theta \in [0, \frac{1}{2}]$ and a linear acoustic wave equation with a dG-FEM approximation in space in [Kar11]. The starting value, which is introduced in [Kar11, equation (2.9)], coincides with the one proposed in (3.1b).

We further note that for $\theta > 0$ the modified θ-scheme can be interpreted as an *IMEX scheme* (implicit-explicit scheme), since the linear part is updated implicitly and the semilinearity explicitly. In particular, for $\theta = \frac{1}{4}$ it can be viewed as a special version of the IMEX scheme considered in [HL21].

3.6.1. Explicit values for constants

We start by stating explicit values for the constants β_Ψ, m_3, $\widetilde{m}_3$ defined in Section 3.1.2 and showing that Assumption 3.16 holds. For the constants m_1, $\widetilde{m}_1$, $\widetilde{m}_2$ as well as $\widehat{\beta}_\Psi$ we do not give explicit values, since we can prove the stability bounds only with β_Ψ.

Lemma 3.61. *For the rational function Ψ_θ defined in (3.4) we have*

$$\beta_\Psi^2 = \beta_\theta^2 = \begin{cases} \frac{4}{1-4\theta}, & \theta \in [0, \frac{1}{4}), \\ \infty, & \theta \geq \frac{1}{4}, \end{cases} \tag{3.92}$$

with β_Ψ given in Definition 3.9.

Proof. Since for all $\theta \geq 0$ the function is monotonically increasing for $z \geq 0$ and $\Psi_\theta(0) = 0$, we directly obtain the result by solving $\Psi_\theta(z) \leq 4$ for $z \geq 0$. □

With Definition 3.17 we have that for $\theta < \frac{1}{4}$ a step-size restriction is necessary to obtain stability for the scheme (3.1), although the scheme is implicit for $\theta \neq 0$. Since such schemes are generally useless in terms of efficiency, we focus on $\theta \geq \frac{1}{4}$ in the following. Nevertheless, the analysis can be extended to the case $\theta < \frac{1}{4}$ with some technical effort.

We further note that $\widehat{\Psi}_\theta(z) = (1 + \theta z)^{-1} > 0$ for all $z \geq 0$. Hence, $\widehat{\Psi}_\theta(z)^{-1}$ exists for every $z \geq 0$. In particular, this allows us to take the inverse of $\widehat{\mathbf{\Psi}}$ for every $\tau > 0$.

Lemma 3.62. *For the rational function Ψ_θ defined in (3.4) the constants given in Definition 3.13 and Lemma 3.14 hold for all $z \geq 0$ and are given by*

$$m_3 = -\Psi_\theta''(0) = 2\theta, \qquad \widetilde{m}_3 = \theta. \tag{3.93}$$

Moreover, Assumption 3.16 is satisfied for all $z \geq 0$.

Proof. The values for m_3 and $\widetilde{m}_3$ immediately follow from

$$\Upsilon(z) = \frac{\widehat{\Psi}_\theta(z) - 1}{z} = -\theta\frac{1}{1 + \theta z} \quad \text{and} \quad \widehat{\Psi}^{-1}(z)\Upsilon(z) = -\theta.$$

Further, Assumption 3.16 obviously holds, since $1 + \theta z \geq 1$ for all $z \geq 0$. □

From this lemma we especially see that the functions Υ and $\widehat{\Psi}^{-1}\Upsilon$ are bounded by the same constant. Additionally, it is worth to mention that the function $\widehat{\Psi}^{-1}\Upsilon$ is in this special case only a constant.

3.6.2. Improved stability results

Next, we show how the stability estimates from Section 3.3 can be improved for this special function. Recall that we focus on the case $\theta \geq \frac{1}{4}$ because it yields unconditionally stable schemes if applied to linear problems (2.12).

Lemma 3.63. *Let $\theta \geq \frac{1}{4}$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}}_\theta \mathbf{q} \right\|_{\mathbf{L}} \leq \|\mathbf{q}\|, \tag{3.94a}$$

and, if additionally $\mathbf{L}$ is positive definite,

$$\tau \left\| \boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}}_\theta \mathbf{q} \right\| \leq c_{\mathrm{inv}} \|\mathbf{q}\|. \tag{3.94b}$$

Proof. As in the proof of Lemma 3.23 it suffices to show the estimates for the eigenvalues of $\tau^2\mathbf{L}$, from which we adopt the notation. Since $\theta \geq \frac{1}{4}$ we have that $\Psi_\theta(z) \in (0,4)$ for $z > 0$. Hence, the function $R_\theta \colon [0,\infty) \to \mathbb{R}$, given by

$$R_\theta(z) = \frac{\widehat{\Psi}_\theta(z)}{\left(\widehat{\Psi}_\theta(z)(1 - \frac{1}{4}\Psi_\theta(z))\right)^{1/2}},$$

is well defined and continuous. From the formula of $\sin \boldsymbol{\Phi}$ in (3.23b) we further have

$$\tau \mathbf{L}^{1/2} \boldsymbol{\mathcal{S}}_n \widehat{\Psi}_\theta(\tau^2\mathbf{L}) = \sin(n\boldsymbol{\Phi}) R_\theta(\tau^2\mathbf{L}).$$

Thus, we have to show that $|R_\theta(z)| \leq 1$ for all $z \geq 0$. From

$$1 - \tfrac{1}{4}\Psi_\theta(z) = (1+\theta z)^{-1}(1 + \theta z - \tfrac{1}{4}z) = \widehat{\Psi}_\theta(z)(1 + \theta z - \tfrac{1}{4}z)$$

we obtain $R_\theta(z) = \left(1 + (\theta - \frac{1}{4})z\right)^{-1/2}$. Hence, we obviously have $|R_\theta(z)| \leq 1$ for all $z \geq 0$ (note that for $\theta = \frac{1}{4}$ we even have $R_\theta(z) = 1$), which finishes the proof. □

In contrast to Lemma 3.23 we see that uniform bounds for $\boldsymbol{\mathcal{S}}_n \widehat{\boldsymbol{\Psi}}_\theta$ exists although $\Psi_\theta(z) \to 4$ for $z \to \infty$ if $\theta = \frac{1}{4}$. Further, note that the bounds are independent of the choice of $\theta \geq \frac{1}{4}$.

We restrict ourselves in the following to the linear case and results in the standard norm $\|\cdot\|$. Similar results can be shown for the discrete energy norm $|||\cdot|||_\tau$. Moreover, as in Section 3.4.3 we can extend the error results to semilinear problems (2.1).

Theorem 3.64. *Let $\theta \geq \frac{1}{4}$ and $n \geq 0$. The approximations $\mathbf{q}_n$ obtained by the modified θ-scheme (3.1), (3.4) applied to the linear problem (2.12) satisfy*

$$\|\mathbf{q}_n\| \leq \|\mathbf{q}_0\| + \min\{t_n, c_{\mathrm{inv}}\}\|\dot{\mathbf{q}}_0\| + \min\{t_n, c_{\mathrm{inv}}\}\, \tau \sum_{\ell=0}^{n-1} \|\mathbf{g}_\ell\|. \tag{3.95}$$

Proof. The estimate with c_{inv} is a direct consequence of the representation formula (3.27) and the previous lemma. The bound with t_n can be done as in Theorem 3.27. □

For the proof of the error estimates we again exploit that $\widehat{\Psi}_\theta(z) > 0$ for $z \geq 0$, yielding together with the improved stability bounds the following.

Theorem 3.65. *Let $\theta \geq \frac{1}{4}$ and $T \in (0, t_*)$. Further, assume that the solution $\mathbf{q}$ of (2.12) satisfies $\mathbf{q} \in C^4([0,T])$. Then, for $t_n \leq T$ we have for the approximations $\mathbf{q}_n$ of the modified θ-scheme (3.1), (3.4)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \leq \min\{T, c_{\mathrm{inv}}\}\Big(C_{1,\theta} + C_{T,\theta} T\Big)\tau^2 \tag{3.96a}$$

where

$$C_{1,\theta} = \theta \max_{0\leq s\leq \tau} \|\mathbf{L}\dot{\mathbf{q}}(s)\| + \tfrac{1}{6} B_\tau^{(3)}, \qquad C_{T,\theta} = \theta \max_{0\leq s\leq T} \|\mathbf{L}\ddot{\mathbf{q}}(s)\| + \tfrac{1}{12} B_T^{(4)}. \tag{3.96b}$$

Proof. The proof is done as in Theorem 3.41, except that we use the bounds from Lemma 3.63 instead of the ones from Lemma 3.23. □

3.6.3. Implementation

To conclude this chapter we shortly turn towards the implementation of the modified θ-scheme (3.1), (3.4). Because of its practical relevance we state the implementation for the general semilinear differential equation (2.2) with a general mass matrix $\mathbf{M}$; see Remark 3.4 for changes in the scheme (3.1). Multiplying the resulting two-step scheme with $\mathbf{M} + \tau^2\theta\mathbf{L}$ then yields

$$(\mathbf{M} + \tau^2\theta\mathbf{L})(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1}) = \tau^2(-\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n), \qquad n = 0, 1, \ldots.$$

The implementation of this two-step scheme is straightforward, for the sake of completeness we still state it in Algorithm 3.1. Clearly, for $\theta > 0$ we have to solve a linear system with the matrix $\mathbf{M} + \tau^2\theta\mathbf{L}$ in each time step, which renders the scheme considerably more costly in general. Nevertheless, since we have for $\theta \geq \frac{1}{4}$ no restriction to the step size at all, the scheme can be beneficial over the leapfrog scheme in some situations. For the starting value (3.1b) a similar algorithm can be derived.

Algorithm 3.1.: Computation of nth time step of the modified θ-scheme (3.1a), (3.4) applied to semilinear problems (2.2).

1: Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
2: $\mathbf{v} = -\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n$
3: Solve $(\mathbf{M} + \tau^2\theta\mathbf{L})\widetilde{\mathbf{v}} = \mathbf{v}$
4: $\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n-1} + \tau^2\widetilde{\mathbf{v}}$

In the next chapter we compare the efficiency of the modified θ-scheme (3.1a), (3.4) with the two-step version (2.20a) of the leapfrog scheme and the LFC schemes (3.1a), (3.3).

Chapter 4

Leapfrog-Chebyshev schemes

After the abstract stability and error analysis for general functions Ψ in the last chapter we now focus on the two-step scheme (3.1a) equipped with the leapfrog-Chebyshev polynomials (3.3). In particular, we show that these functions fit into the framework of the last chapter, and discuss the efficiency of these schemes. Recall that we are interested in the situation that the stiffness of the differential equation (2.1) is induced by the matrix $\mathbf{L}$ and $\mathbf{g}$ is a function with a rather small Lipschitz constant but costly to evaluate in comparison to a matrix-vector multiplication with $\mathbf{L}$.

For convenience we concisely rewrite the *leapfrog-Chebyshev* (LFC) *schemes*

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2 \widehat{P}_p(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \qquad n = 1, 2, \dots, \tag{4.1a}$$

where

$$\Psi(z) = P_p(z) = \widehat{P}_p(z)z = 2 - \frac{2}{T_p(\nu)} T_p\Big(\nu - \frac{z}{\alpha_p}\Big), \qquad \alpha_p = 2\frac{T_p'(\nu)}{T_p(\nu)}, \qquad \nu \geq 1. \tag{4.1b}$$

Recall that T_p denotes the pth Chebyshev polynomial of first kind so that P_p is a polynomial of degree $p \geq 1$; see Section B.2 in the appendix. To complete the scheme we equip (4.1a) with the starting value

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau P_p'(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2 \widehat{P}_p(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_0 + \mathbf{g}_0), \tag{4.1c}$$

which is an extension of (3.74) to $\mathbf{g} \not\equiv 0$. Nevertheless, as we have seen in Section 3.5.1 there is no "optimal" choice for the starting value in general. We thus could have also equipped the LFC scheme with the starting value (3.1b), yielding

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau \widehat{P}_p(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2 \widehat{P}_p(\tau^2\mathbf{L})(-\mathbf{L}\mathbf{q}_0 + \mathbf{g}_0). \tag{4.2}$$

Further, regarding the stabilization parameter ν, we pay special attention to the choice

$$\nu = \nu_{p,\eta} = 1 + \frac{\eta^2}{2p^2} \tag{4.3}$$

with yet another stabilization parameter $\eta \geq 0$. This choice is motivated by stabilized/damped Runge–Kutta–Chebyshev methods; see, e.g., [VHS90, HV03].

In Section 4.1 we start with a short overview about the origin of the LFC polynomials and a motivation of their construction. In addition we show that Assumption 3.2 is satisfied. Afterwards in Section 4.2, we state explicit values for all constants arising for Ψ in Sections 3.1.2 and 3.5.1. Moreover, we show the advantages of the special stabilization parameter $\nu_{p,\eta}$. In Section 4.3 we point out a close relation between the *unstabilized* LFC schemes, i.e., with $\nu = 1$, and the leapfrog method (2.20) for linear, homogeneous problems. We continue with the implementation of the LFC schemes and show the efficiency in the situation described above compared to the leapfrog scheme (2.20) and the modified θ-schemes considered in Section 3.6. Finally, we confirm our theoretical findings with some numerical examples in Section 4.5.

Most of the results in this chapter are published either in [CHS20] or [CH21]. The constants involving ν in Section 4.2.1 where mainly shown by the authors of [HS18] with exception of Theorem 4.11. In addition, Lemma 4.12 contains considerably improved bounds compared to the one in [CH21, Corollary 5.2].

4.1. Motivation and some first observations

We begin with a short motivation about the construction of the LFC scheme (4.1a),(4.1b). For linear problems (2.12) with $\mathbf{g} \equiv 0$ the unstabilized scheme has been constructed (without specifying $\mathbf{q}_1$) in [GJ08, JR10]. To see this, we point out that for $\nu = 1$ we have $\alpha_p = 2p^2$ due to $T_p(1) = 1$ and $T_p'(1) = p^2$; see (B.15). Hence, we obtain

$$P_p(z) = P_{p,*}(z) = 2 - 2T_p\Big(1 - \frac{z}{2p^2}\Big) \tag{4.4}$$

in accordance with [GJ08, JR10]. In these papers it is further shown that the unstabilized polynomials (4.1b) are optimal in the following sense (note that $P_{p,*}$ satisfies (3.10) for every $p \in \mathbb{N}$).

Lemma 4.1. *For all polynomials $Q \not\equiv P_{p,*}$ of degree $p \in \mathbb{N}$ satisfying (3.10), i.e., $Q(0) = 0$ and $Q'(0) = 1$, we have*

$$\beta_Q < \beta_{P_{p,*}} = 2p,$$

where β_Q, $\beta_{P_{p,*}}$ are given as in Definition 3.9(a).

The lemma states that among all polynomials of a fixed degree $p \in \mathbb{N}$ satisfying the consistency conditions (3.10), the polynomials (4.1b) with $\nu = 1$ admit the maximum value for β_Ψ in Definition 3.9(a). Thus, by Definition 3.17 the step-size restriction $\tau \le \tau_{\mathrm{SSR}} = 2p/\|\mathbf{L}\|^{1/2}$ is the weakest one we can achieve with polynomials of degree $p \in \mathbb{N}$ fulfilling (3.10). This result is not surprising at all if one brings to mind the "optimality properties" of Chebyshev polynomials; cf. Lemma B.7 for a similar result.

Proof. The proof closely follows the lines of the proof of Lemma B.7; see also [HV03, Theorem 1.1] for a related proof for Runge–Kutta–Chebyshev methods. The value $\beta_{P_p} = 2p$ is a direct consequence of Lemma B.4(a) and Lemma B.5; see also Theorem 4.4 below. □

Remark 4.2. The optimality of these polynomials in terms of β_Ψ has implicitly already been shown in [JN81, Theorem 5.1]. By using this theorem we obtain for the test equation

$$y' = \mathrm{i}\lambda y, \qquad \lambda \in \mathbb{R},$$

with $\mu = \tau \mathrm{i}\lambda$ and $k \in \mathbb{N}$, that

$$y_{n+1} + (-1)^k y_{n-1} = \kappa_k(\mu) y_n, \quad \kappa_k(\mu) = 2\,\mathrm{i}^k\, T_k\Big(-\mathrm{i}\frac{\mu}{k}\Big), \qquad n = 1, 2, \dots.$$

Considering only even $k = 2p$, $p \in \mathbb{N}$, yields by using the identity (B.13) (with $k = 2$)

$$y_{n+1} + y_{n-1} = 2\,(-1)^p\, T_{2p}\Big(\frac{\tau\lambda}{2p}\Big) y_n = 2T_p\Big(1 - \frac{\tau^2\lambda^2}{2p^2}\Big) y_n, \qquad n = 1, 2, \dots,$$

where in the second step we additionally used that Chebyshev polynomials are even or odd according to the parity of p; see Lemma B.2. Hence, for $\nu = 1$ this coincides with the scheme (4.1a), (4.1b) applied to the second-order test equation

$$y'' = -\lambda^2 y, \qquad \lambda \in \mathbb{R},$$

which is obtained from the first-order test equation by differentiating. ◇

As we have shown in the last chapter, for a correct long-time behavior it is in general favorable to use the stronger step-size restriction $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, which forbids the polynomials to be equal or too close to 0 or 4 except for $z = 0$; see Definitions 3.22 and 3.9(b). In Theorem 4.9 below we show that for $\nu = 1$ this yields $\widehat{\beta}_\Psi^2 \ll 4p^2$, or, equivalently, $\widehat{\tau}_{\mathrm{SSR}} \ll \tau_{\mathrm{SSR}}$, since $P_p(z) \in \{0, 4\}$ for p values of $z \in (0, 4p^2]$; see also Figure 4.2. Clearly, without having to go into details, this renders the scheme considerably less efficient.

As a remedy to this problem, we introduce a stabilization parameter $\nu \ge 1$, which is motivated by damped/stabilized Runge–Kutta–Chebyshev (RKC) methods; see, e.g., [HS80, Ver82, VHS90] or [HV03, Chapter V]. Unstabilized/undamped RKC methods are explicit s-stage Runge–Kutta methods which have the largest stability interval along the negative real axis among all explicit s-stage Runge–Kutta methods. The construction relies on adjusting the stability polynomials, which again involves scaled and shifted Chebyshev polynomials. Similarly as for $\nu = 1$ the unstabilized/undamped RKC methods suffer from instabilities at some discrete points, which led to the construction of damped/stabilized variants.

We point out that the scheme (4.1) can be viewed as a *multiple time-stepping scheme*, i.e., as a time integration scheme, where one part of the differential equation is numerically integrated with smaller step sizes than the other. More precisely, each time step requires p matrix-vector multiplications with $\mathbf{L}$ (P_p is a polynomial of degree p), and only one evaluation of $\mathbf{g}$. Hence, the linear part inducing the stiffness is evaluated p times more often than the "nice" nonlinear part. In Section 4.3 we will see that our scheme is for $\nu = 1$ indeed closely related to a multiple time-stepping scheme.

We conclude this section with some basic properties of LFC schemes. For $p = 1$ the general scheme (4.1) reduces to the leapfrog scheme (2.20), since $\alpha_1 = \frac{2}{\nu}$ due to $T_1(x) = x$ and, thus,

$$P_1(z) = 2 - \frac{2}{T_1(\nu)} T_1\Big(\nu - \frac{z}{\alpha_1}\Big) = 2 - 2\Big(1 - \frac{z}{2}\Big) = z. \tag{4.5}$$

In particular, P_1 is independent of ν. Obviously, in this case both starting values (4.1c) and (4.2) coincide with the Taylor starting value (2.20b) of the leapfrog scheme.

Last, we show that Assumption 3.2 holds.

Lemma 4.3. *The LFC polynomials* (4.1b) *satisfy Assumption* 3.2 *for every* $p \in \mathbb{N}$ *and* $\nu \ge 1$.

Proof. Condition $\widehat{\Psi}(0) = 1$ in Assumption 3.2 is equivalent to the consistency conditions (3.10). Since the derivative of P_p is given with the definition of α_p in (4.1b) by

$$P_p'(z) = \frac{2}{\alpha_p T_p(\nu)} T_p'\Big(\nu - \frac{z}{\alpha_p}\Big) = \frac{1}{T_p'(\nu)} T_p'\Big(\nu - \frac{z}{\alpha_p}\Big), \tag{4.6}$$

the consistency conditions are fulfilled. □

4.2. Constants

In this section we state explicit values for all constants arising in Sections 3.1.2 and 3.5.1 in case of LFC polynomials (4.1b) for Ψ. Moreover, we show that these polynomials satisfy Assumption 3.16. Afterwards we have a closer look at the dependency of these values on ν and the polynomial degree p. Concluding, we present the advantages of the specific choice $\nu = \nu_{p,\eta}$ defined in (4.3). In particular, we state values for some of the constants, which only depend on η but are independent of the polynomial degree p.

4.2.1. Values in dependence of the stabilization parameter ν

We first focus on constants depending on β_Ψ before we turn towards constants involving $\widehat{\beta}_\Psi$. Recall that the values for β_Ψ and $\widehat{\beta}_\Psi$ determine the step-size restrictions $\tau \le \tau_{\mathrm{SSR}}$ and $\tau \le \widehat{\tau}_{\mathrm{SSR}}$, respectively, which are required to obtain (in some sense) stability of the schemes; see Definitions 3.17 and 3.22 in the previous chapter.

Constants involving β_Ψ

We start by stating an explicit value for β_Ψ.

Theorem 4.4. *For $p \in \mathbb{N}$ and $\nu \ge 1$ the polynomials P_p defined in (4.1b) satisfy $\beta_\Psi = \beta_{p,\nu}$, where*

$$\beta_{p,\nu}^2 = 2\alpha_p \nu \tag{4.7}$$

and β_Ψ is defined in Definition 3.9(a).

From the theorem we directly obtain that for $\nu = 1$ we have $\beta_{p,1}^2 = 4p^2$, since $\alpha_p = 2p^2$ due to (B.15). Moreover, we have $\beta_{p,\nu}^2 < \beta_{p,1}^2 = 4p^2$ for every $\nu > 1$ in accordance with Lemma 4.1, because $\beta_{p,\nu}$ is strictly monotonically decreasing in ν; see Lemma A.4.

In order to prove this and the next theorems we frequently make use of the following change of variables

$$x = \nu - \frac{z}{\alpha_p} \in [-\nu, \nu] \qquad \Longleftrightarrow \qquad z = \alpha_p(\nu - x) \in [0, \beta_{p,\nu}^2]. \tag{4.8}$$

Observe that $z = \beta_{p,\nu}^2$ and $z = 0$ corresponds to $x = -\nu$ and $x = \nu$, respectively.

Proof. We have to show that the inequalities (3.11) hold true for all $z \in [0, \beta_{p,\nu}^2]$ and that $\beta_{p,\nu}^2$ is the largest value such that these inequalities hold. Lemmas B.2 to B.5 yield that Chebyshev polynomials T_p satisfy for $\nu \ge 1$

$$-T_p(\nu) \le T_p(x) \le T_p(\nu) \qquad \text{for } x \in [-\nu, \nu].$$

From this we obtain by subtracting $T_p(\nu) \geq 1$ and multiplying with $-2/T_p(\nu)$ that

$$0 \leq 2 - \frac{2}{T_p(\nu)} T_p(x) \leq 4 \qquad \text{for } x \in [-\nu, \nu],$$

which shows with the transformation (4.8) that (3.11) is at least satisfied for all $z \in [0, \beta_{p,\nu}^2]$. The optimality of $\beta_{p,\nu}^2$ follows in the case of p odd from $P_p(\beta_{p,\nu}^2) = 4$ and the strictly monotonic growth of T_p for $x \leq -1$; see Lemma B.5. For p even we have that $P_p(\beta_{p,\nu}^2) = 0$ and T_p is strictly monotonically decreasing for $x \leq -1$. □

With this result we are in a position to state explicit values for all (error) constants defined in Sections 3.1.2 and 3.5.1 depending on $\beta_\Psi = \beta_{p,\nu}$. Additionally, we show that Assumption 3.16 holds for the LFC polynomials (4.1b).

Theorem 4.5. *Let $p \in \mathbb{N}$ and $\nu \geq 1$. For the polynomials P_p defined in (4.1b) the constants in Definitions 3.13, 3.51, and 3.53 are given by*

$$m_3 = m_3^\star = m_3' = m_3^{p,\nu} = -P_p''(0) = \frac{T_p''(\nu)}{\alpha_p T_p'(\nu)} \tag{4.9a}$$

and

$$m_4 = m_4^\star = m_4^{p,\nu} = \tfrac{1}{6} P_p'''(0) = \frac{T_p'''(\nu)}{6\alpha_p^2 T_p'(\nu)}. \tag{4.9b}$$

Moreover, Assumption 3.16 is satisfied for all $z \in [0, \beta_{p,\nu}^2]$.

We point out that this theorem is the main reason why we have included the factors $\frac{1}{2}$ in Definition 3.13 and 3 in Definition 3.53; see also Definition 3.51. Further, because of $T_p'' \equiv T_p''' \equiv 0$ for $p = 1$ and every $\nu \geq 1$ we have $m_3 = m_3^\star = m_3' = m_4 = m_4^\star = 0$ in accordance with the corresponding definitions; cf. Example 3.15. For $p = 2$ we still get $m_4 = m_4^\star = 0$.

Proof. The values for the constants are proven separately. Throughout the proof we use several times that

$$T_p^{(k)}(x) \leq |T_p^{(k)}(x)| \leq T_p^{(k)}(\nu) \qquad \text{for } x \in [-\nu, \nu],\ k \in \mathbb{N}. \tag{4.10}$$

This can be deduced from (B.15) (Markov brothers' inequality) together with Lemmas B.2 and B.5 due to $\nu \geq 1$.

(i) We first prove (3.18) in Assumption 3.16. Choosing $k = 1$ in (4.10) yields $T_p'(x) \leq T_p'(\nu)$ for $x \in [-\nu, \nu]$. Thus, with (4.6) and the transformation (4.8) we obtain $P_p'(z) \leq 1$ for all $z \in [0, \beta_{p,\nu}^2]$. Integrating from 0 to $z \in [0, \beta_{p,\nu}^2]$ leads to $P_p(z) - P_p(0) \leq z$, which shows (3.18) because of $P_p(0) = 0$.

(ii) By Definition 3.53 we have $m_3' = -\Psi''(z) = -P_p''(0)$. From (4.6) we obtain

$$P_p''(z) = -\frac{1}{\alpha_p T_p'(\nu)} T_p''(x), \tag{4.11}$$

which yields the formula for $P_p''(0)$ in (4.9a).

(iii) Next, we show that the bounds (3.15) and (3.79) in Definitions 3.13 and 3.51, respectively, hold with $m_3 = m_3^\star = -P_p''(0)$. This means that we have to prove

$$|P_p(z) - z| \leq \tfrac{1}{2}|P_p''(0)|\, z^2 \quad \text{and} \quad |P_p'(z) - 1| \leq |P_p''(0)|\, z \qquad \text{for } z \in [0, \beta_{p,\nu}^2]. \tag{4.12}$$

To do so, we integrate (4.10) for $k = 2$ twice from $x \in [-\nu, \nu]$ to ν. Integrating once leads to

$$T_p'(\nu) - T_p'(x) \leq T_p''(\nu)(\nu - x). \tag{4.13}$$

Dividing this inequality by $-T_p'(\nu)$ yields with (4.6), (4.8), and the definition of α_p in (4.1b)

$$P_p'(z) = \frac{1}{T_p'(\nu)} T_p'(x) \geq 1 - \frac{T_p''(\nu)}{T_p'(\nu)}(\nu - x) = 1 - \frac{T_p''(\nu)}{\alpha_p T_p'(\nu)} z = 1 + P_p''(0) z.$$

Integrating (4.13) a second time yields

$$-T_p(\nu) + T_p(x) \leq -T_p'(\nu)(\nu - x) + \frac{T_p''(\nu)}{2}(\nu - x)^2,$$

from which we obtain similarly as before by multiplication with $-2/T_p(\nu)$

$$P_p(z) = 2 - \frac{2}{T_p(\nu)} T_p(x) \geq \alpha_p(\nu - x) - \frac{T_p''(\nu)}{T_p(\nu)}(\nu - x)^2 = z - \frac{T_p''(\nu)}{2\alpha_p T_p'(\nu)} z^2 = z + \tfrac{1}{2} P_p''(0) z^2.$$

Together with

$$P_p(z) - z \leq 0 \qquad \text{and} \qquad P_p'(z) - 1 \leq 0$$

for $z \in [0, \beta_{p,\nu}^2]$, which follows from part (i) of this proof, this shows (4.12).

(iv) It remains to show that the bounds (3.82) in Definition 3.53 hold with $m_4 = m_4^\star$ as given in (4.9b), i.e.,

$$\left| P_p(z) - z - \tfrac{1}{2} P_p''(0) z^2 \right| \leq \tfrac{1}{6} P_p'''(0) z^3, \quad \left| P_p'(z) - 1 - P_p''(0) z \right| \leq \tfrac{1}{2} P_p'''(0) z^2$$

for all $z \in [0, \beta_{p,\nu}^2]$. The value for $P_p'''(0)$ can be computed by differentiating (4.11) once more.

We proceed similarly to part (iii). Choosing $k = 3$ in (4.10) and integrating twice and three times from $x \in [-\nu, \nu]$ to ν yields

$$-T_p'(\nu) + T_p'(x) \leq -T_p''(\nu)(\nu - x) + \frac{T_p'''(\nu)}{2}(\nu - x)^2$$

and

$$T_p(\nu) - T_p(x) \leq T_p'(\nu)(\nu - x) - \frac{T_p''(\nu)}{2}(\nu - x)^2 + \frac{T_p'''(\nu)}{6}(\nu - x)^3.$$

Rearranging both inequalities as before leads to

$$P_p'(z) \leq 1 - \frac{T_p''(\nu)}{T_p'(\nu)}(\nu - x) + \frac{T_p'''(\nu)}{2T_p'(\nu)}(\nu - x)^2 = 1 + P_p''(0) z + \tfrac{1}{2} P_p'''(0) z^2$$

and

$$P_p(z) \leq \alpha_p(\nu - x) - \frac{T_p''(\nu)}{T_p(\nu)}(\nu - x)^2 + \frac{T_p'''(\nu)}{3T_p(\nu)}(\nu - x)^3 = z + \tfrac{1}{2} P_p''(0) z^2 + \tfrac{1}{6} P_p'''(0) z^3.$$

This completes the proof, since we have from part (iii) that $P_p(z) - z - \frac{1}{2} P_p''(0) z^2 \geq 0$ and $P_p'(z) - 1 - P_p''(0) z \geq 0$ for all $z \in [0, \beta_{p,\nu}^2]$. □

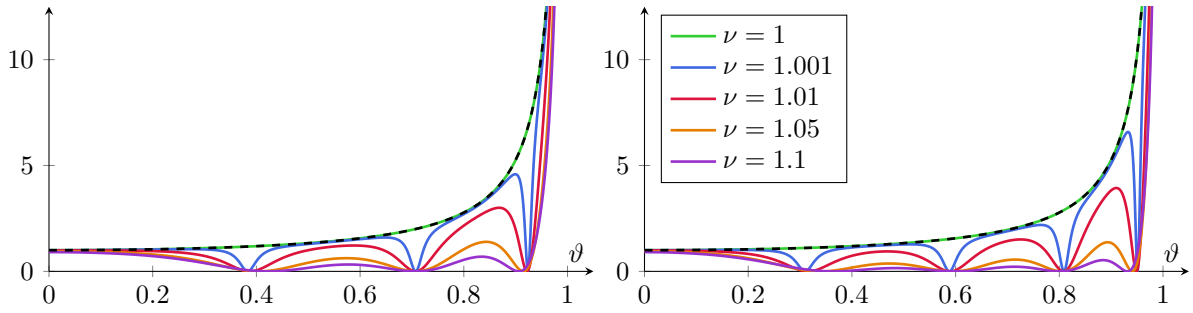


Figure 4.1.: Illustration of Conjecture 4.8. The function $\widehat{\Gamma}\colon (0,1) \to \mathbb{R}$, $\vartheta \mapsto \Gamma(\beta_{p,\nu}^2 \vartheta^2)$, with Γ defined in (3.76) is plotted for the LFC polynomials (4.1b) with $p = 4$ (left) and $p = 5$ (right) and stabilization parameters $\nu = 1$, $\nu = 1.001$, $\nu = 1.01$, $\nu = 1.05$, $\nu = 1.1$. The black dashed line represents $(1 - \vartheta^2)^{-1}$.

Remark 4.6. As stated in Theorem 3.54, the scheme (4.1) with special starting value (4.1c) applied to a linear homogeneous problem (2.12) is of order four if and only if $M_3 = 0$, i.e., $m_3' = \frac{1}{6}$. As we will see in Section 4.2.2, this is always possible for the LFC polynomials (4.1b) for $p \geq 2$ by adjusting $\nu = \nu_{p*}$. For instance, for $p = 2, \dots, 5$ the choices

$$\nu_{2*} = \tfrac{1}{2}\sqrt{6} \approx 1.224745, \qquad \nu_{3*} = \left(\tfrac{1}{2} + \tfrac{1}{4}\sqrt{5}\right)^{1/2} \approx 1.029086,$$
$$\nu_{4*} \approx 1.008261, \qquad \nu_{5*} \approx 1.003233,$$

fulfill $m_3' = \frac{1}{6}$. Note that in the case of $p = 2$ and $\nu_{2*} = \frac{1}{2}\sqrt{6}$, we retrieve $P_2(z) = z - \frac{1}{12}z^2$ with $\beta_2^2 = 12$, yielding the modified (equation) leapfrog method; see Example 3.55. ◇

Next, we turn towards the constants $C^\star$ and $\Psi'_{\max}$ in Lemma 3.48 and Theorem 3.49 in Section 3.5.1, respectively, stemming from the special starting value (4.1c). We start with the latter one.

Theorem 4.7. *For the polynomials P_p defined in (4.1b) we have*

$$\Psi'_{\max} = \max_{z \in [0, \beta_{p,\nu}^2]} |P_p'(z)| = 1$$

for every $p \in \mathbb{N}$ and $\nu \geq 1$.

Proof. Using again (4.10) with $k = 1$ leads to $|T_p'(x)| \leq T_p'(\nu)$ for $x \in [-\nu, \nu]$. Together with (4.6) and (4.8) this shows the claim. □

Unfortunately, for the constant $C^\star$ we are not able to prove sharp bounds for $\nu > 1$ and $p \in \mathbb{N}$. Instead we state the following conjecture.

Conjecture 4.8. *Let $\vartheta \in (0,1)$. Lemma 3.48 holds for the polynomials P_p defined in (4.1b) for every $p \in \mathbb{N}$ and $\nu \geq 1$ with $C^\star(\vartheta, P_p) = (1 - \vartheta^2)^{-1/2}$.*

Below we show a proof for the special case $\nu = 1$. However, we believe that this result also holds for all $\nu > 1$, indicated in Figure 4.1. Recall that we have to show that the function Γ defined in (3.76) satisfies $|\Gamma(z)| \leq C^\star(\vartheta, P_p)^2 = (1 - \vartheta^2)^{-1}$ for all $z \in [0, \beta_{p,\nu}^2 \vartheta^2]$.

Proof for $\nu = 1$. Let $\vartheta \in (0,1)$. Since $\alpha_p = 2p^2$ and $\beta_{p,\nu}^2 = 4p^2$ due to $\nu = 1$, we get from the transformation (4.8) that $z = 2p^2(1-x)$ and $x \in [1-2\vartheta^2, 1] \subset (-1,1]$. Thus, we can write $x = \cos\psi$ for some $\psi \in [0,\pi)$.

By using the LFC polynomials (4.1b) in the definition (3.76) of Γ we obtain with (4.6) and (B.21)

$$\Gamma(z) = \frac{2(1-x)\,T_p'(x)^2}{p^2(1-T_p(x))(1+T_p(x))} = \frac{2}{1+x}$$

where in the first step we additionally used that $T_p(1) = 1$ and $T_p'(1) = p^2$; see Lemma B.4 and (B.15). Hence, resubstituting z for x leads with $z \le 4p^2\vartheta^2$ to

$$0 \le \Gamma(z) = \frac{2}{1+x} = \frac{1}{(1-z/(4p^2)} \le \frac{1}{1-\vartheta^2},$$

which completes the proof. □

Constants involving $\widehat{\beta}_\Psi$

In the next two theorems we focus on constants involving $\widehat{\beta}_\Psi$ defined in Definition 3.9(b). Since the LFC polynomials (4.1b) for $p = 1$ yield the leapfrog scheme and, thus, the constants are already given in the previous chapter, we only consider the LFC polynomials P_p with $p \ge 2$. We start by stating specific choices for m_1, $\widetilde{m}_1$, $\widetilde{m}_2$ to obtain an explicit expression for $\widehat{\beta}_\Psi$.

Theorem 4.9. *Let $p \ge 2$. For the polynomial P_p defined in (4.1b), Definition 3.9(b) holds*
(a) *for $\nu = 1$ and every $m_1, \widetilde{m}_1, \widetilde{m}_2 > 0$ with $\widehat{\beta}_\Psi^2 < 2p^2(1-\cos\frac{\pi}{p}) < \pi^2$,*
(b) *for $\nu > 1$ and*

$$m_1 = \widetilde{m}_1 = m_1^{p,\nu} = \frac{1}{2}\Big(1 - \frac{1}{T_p(\nu)}\Big), \qquad \widetilde{m}_2 = \widetilde{m}_2^{p,\nu} = \frac{4\widetilde{m}_1}{\alpha_p(\nu-1)} = \frac{T_p(\nu)-1}{T_p'(\nu)(\nu-1)}, \tag{4.14a}$$

with

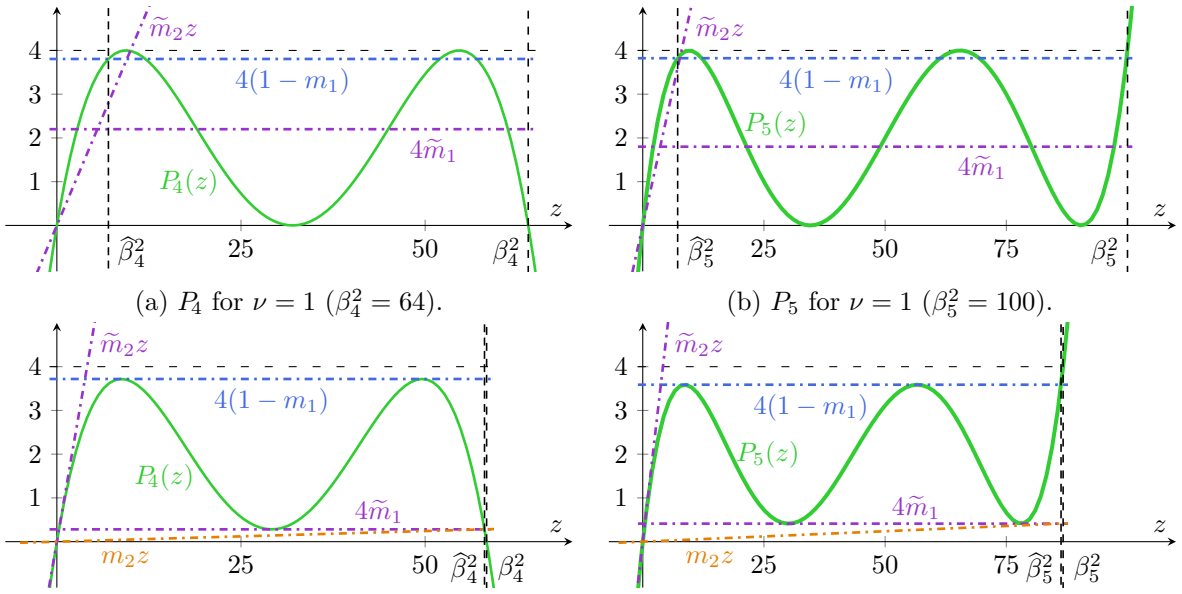
$$\widehat{\beta}_\Psi^2 = \widehat{\beta}_{p,\nu}^2 = \alpha_p(\nu+1). \tag{4.14b}$$

The lemma states that for $\nu = 1$ we always have $\widehat{\beta}_\Psi^2 < \pi^2$ regardless of the choice of m_1, $\widetilde{m}_1$, $\widetilde{m}_2$, and the polynomial degree p. Hence, in this case the stronger step-size restriction condition $\tau \le \widehat{\tau}_{\mathrm{SSR}} = \widehat{\beta}_\Psi/\|\mathbf{L}\|^{1/2}$ is much more restrictive than the weaker one $\tau \le \tau_{\mathrm{SSR}} = \beta_\Psi/\|\mathbf{L}\|^{1/2}$ and independent of the polynomial degree, whereas $\beta_\Psi = \beta_{p,1} = 2p$ grows linearly in p. In contrast to this, for $\nu > 1$ we can choose the constants for m_1, $\widetilde{m}_1$, $\widetilde{m}_2$ in such a way that we receive a value $\widehat{\beta}_\Psi = \widehat{\beta}_{p,\nu}$ which is only slightly smaller than $\beta_{p,\nu}$. More precisely, we have $\widehat{\beta}_{p,\nu}^2 = \alpha_p(\nu+1) < 2\alpha_p\nu = \beta_{p,\nu}^2$ for all $\nu > 1$ (note that $\lim_{\nu\to 1} m_1^{p,\nu} = 0$ and $\lim_{\nu\to 1}\widetilde{m}_2 = 1$). An illustration of this behavior is given in Figure 4.2.

Proof. (a) Let $\nu = 1$. As mentioned before, we obtain $\alpha_p = 2p^2$ and $\beta_{p,\nu}^2 = 4p^2$. Since $p \ge 2$, we can use that for $x \in (-1,1)$ the local extrema of T_p are given by $x_k = \cos(k\frac{\pi}{p})$, $k = 1,\dots,p-1$; see Lemma B.4. This yields due to $\nu = 1$

$$P_p(z_k) = 2 - 2T_p\Big(1 - \frac{z_k}{2p^2}\Big) \in \{0,4\} \qquad \text{for } z_k = 2p^2(1-x_k) \in (0,\beta_{p,\nu}^2),$$

$k = 1,\dots,p-1$. Since z_1 is the smallest extremum point in $z \in (0,\beta_{p,\nu}^2)$ we have that $\widehat{\beta}_{p,\nu}^2 < z_1$ because of $P_p(z_1) = 4$ and $m_1 > 0$ in Definition 3.9(b). The estimate $z_1 < \pi^2$ follows from $\cos(\zeta) > 1 - \frac{1}{2}\zeta^2$, $\zeta \in (0, \frac{1}{2}\pi]$.



(a) P_4 for $\nu = 1$ ($\beta_4^2 = 64$).

(b) P_5 for $\nu = 1$ ($\beta_5^2 = 100$).

(c) P_4 for $\nu = 1.01 > 1$.

(d) P_5 for $\nu = 1.01 > 1$.

Figure 4.2.: Illustration of Theorem 4.9 for the LFC polynomials for $p = 4, 5$ and $\nu = 1$ (Figures 4.2a and 4.2b) and $\nu = 1.01$ (Figures 4.2c and 4.2d). The values of m_1, $\widetilde{m}_1$, and $\widetilde{m}_2$ in Figures 4.2c and 4.2d are those of (4.14a). Additional to these constants the line $m_2 z$ given in (4.15) is plotted.

(b) Let $\nu > 1$. We have to prove that the lower and upper bounds in (3.12) hold true for all $z \in [0, \widehat{\beta}_{p,\nu}^2]$ with constants (4.14a). Note that from the variable transformation (4.8) we have that $z \in [0, \widehat{\beta}_{p,\nu}^2]$ is equivalent to $x \in [-1, \nu]$. Additionally, we have for $\sigma_{p,\nu} = \alpha_p(\nu - 1)$ that $z = \sigma_{p,\nu}$ is equivalent to $x = 1$.

Since we have $T_p(x) \geq -1$ for $x \in [-1, \nu]$ (see Lemmas B.4 and B.5), we get

$$P_p(z) = 2 - \frac{2}{T_p(\nu)} T_p(x) \leq 2 + \frac{2}{T_p(\nu)} = 4(1 - m_1) \qquad \text{for } z \in [0, \widehat{\beta}_{p,\nu}^2],$$

which is the desired bound with m_1 given in (4.14a).

To show the lower bounds in (3.12), we distinguish the cases of $z \in [0, \sigma_{p,\nu}]$ and $z \in [\sigma_{p,\nu}, \widehat{\beta}_{p,\nu}^2]$. For $z \in [\sigma_{p,\nu}, \widehat{\beta}_{p,\nu}^2]$ we can show the bound with $\widetilde{m}_1$ similarly as for m_1. Using that $T_p(x) \leq 1$ for $x \in [-1, 1]$ yields

$$P_p(z) = 2 - \frac{2}{T_p(\nu)} T_p(x) \geq 2 - \frac{2}{T_p(\nu)} = 4\widetilde{m}_1 \qquad \text{for } z \in [\sigma_{p,\nu}, \widehat{\beta}_{p,\nu}^2].$$

For $z \in [0, \sigma_{p,\nu}]$ we have that the polynomials P_p are concave, which can be seen from (4.11) because for $x \in [1, \nu]$ we have that $T_p''(x) \geq 0$; see Lemma B.5. Thus, we have with $P_p(0) = 0$ that

$$P_p(z) \geq \frac{z}{\sigma_{p,\nu}} P_p(\sigma_{p,\nu}) = \frac{1}{\sigma_{p,\nu}} \left(2 - \frac{2}{T_p(\nu)} T_p(1)\right) z = \frac{4\widetilde{m}_1}{\sigma_{p,\nu}} z = \widetilde{m}_2 z \qquad \text{for } z \in [0, \sigma_{p,\nu}],$$

which finishes the proof. □

Remark 4.10. (i) The previous lemma directly implies a value $m_2^{p,\nu}$ for the constant m_2 given in Remark 3.10, since we obtain with (3.13)

$$P_p(z) \geq \min\left\{\frac{4m_1^{p,\nu}}{\widehat{\beta}_{p,\nu}^2}, \widetilde{m}_2\right\} z = \min\left\{\frac{4m_1^{p,\nu}}{\widehat{\beta}_{p,\nu}^2}, \frac{4m_1^{p,\nu}}{\alpha_p(\nu-1)}\right\} z = \frac{4m_1^{p,\nu}}{\widehat{\beta}_{p,\nu}^2} z = m_2^{p,\nu} z. \tag{4.15}$$

Obviously, we have $\widetilde{m}_2^{p,\nu} > m_2^{p,\nu}$ in this case; see also Figures 4.2c and 4.2d. Further, as $m_2^{p,\nu}$ depends on $\widehat{\beta}_{p,\nu}$, the constant deteriorates if one increases the polynomial degree p. A different proof of this estimate can be found in [CHS20, Theorem 5.2].

(ii) It is possible to slightly increase the stability bound $\widehat{\beta}_{p,\nu}^2$ given in (4.14b) for $\nu > 1$ (up to $\beta_{p,\nu}^2$). However, we have to degrade either m_1 or $\widetilde{m}_1$ depending on whether the polynomial degree p is odd or even; cf. Figures 4.2c and 4.2d. More precisely, to obtain $\widehat{\beta}_{p,\nu}^2 = \alpha_p(\kappa+\nu)$ with $\kappa \in [1,\nu)$ one has to replace the constants in (4.14a) with

$$m_1 = \frac{1}{2}\Big(1 - \frac{T_p(\kappa)}{T_p(\nu)}\Big) \qquad \text{or} \qquad \widetilde{m}_1 = \frac{1}{2}\Big(1 - \frac{T_p(\kappa)}{T_p(\nu)}\Big),$$

depending on the parity of p. A more detailed proof for m_1 (and m_2 instead of $\widetilde{m}_2$, $\widetilde{m}_1$) is given in [HS18, Lemma 5.4]. ◇

It remains tho show an explicit value for $\widetilde{m}_3$ in Lemma 3.14.

Theorem 4.11. *Let $p \geq 2$ and $\nu > 1$. The polynomials P_p defined in (4.1b) satisfy the inequality (3.17) in Lemma 3.14 for $z \in [0, \widehat{\beta}_{p,\nu}^2]$ with*

$$\widetilde{m}_3 = \widetilde{m}_3^{p,\nu} = \frac{1}{4m_1^{p,\nu}} = \frac{1}{2}\frac{T_p(\nu)}{T_p(\nu)-1},$$

where $\widehat{\beta}_{p,\nu}$ and $m_1^{p,\nu}$ are defined as in (4.14).

Proof. To simplify the notation in the following we define the *continuous* function

$$\widetilde{\Upsilon}\colon [0,\widehat{\beta}_{p,\nu}^2] \to \mathbb{R}, \qquad \widetilde{\Upsilon}(z) = \widehat{\Psi}(z)^{-1}\Upsilon(z) = \begin{cases} \frac{P_p(z)-z}{P_p(z)z}, & z > 0, \\ \frac{1}{2}P_p''(0), & z = 0. \end{cases} \tag{4.16}$$

Moreover, as in the last proof we abbreviate $\sigma_{p,\nu} = \alpha_p(\nu-1)$. We first observe that $\widetilde{\Upsilon}(z) \leq 0$ for $z \in [0,\widehat{\beta}_{p,\nu}^2]$, since $\widehat{\Psi}(z) > 0$ and $\Upsilon(z) \leq 0$ for $z \in [0,\widehat{\beta}_{p,\nu}^2]$ by (3.19) due to Assumption 3.16. Thus, we have to bound $\widetilde{\Upsilon}$ from below by $-\widetilde{m}_3$ for all $z \in [0,\widehat{\beta}_{p,\nu}^2]$.

For $z \in [\sigma_{p,\nu},\widehat{\beta}_{p,\nu}^2]$ we have

$$\widetilde{\Upsilon}(z) = \frac{1}{z} - \frac{1}{P_p(z)} \geq -\frac{1}{P_p(z)} \geq -\frac{1}{4m_1^{p,\nu}},$$

where the last inequality follows as for $\widetilde{m}_1$ in the previous proof. For $z \in [0,\sigma_{p,\nu}]$ we employ that $\widetilde{\Upsilon}$ is monotonically decreasing; see Lemma A.3 below. This leads to

$$\widetilde{\Upsilon}(z) \geq \widetilde{\Upsilon}(\sigma_{p,\nu}) \geq -\frac{1}{P_p(\sigma_{p,\nu})} = -\Big(2 - \frac{2}{T_p(\nu)}T_p(1)\Big)^{-1} = -\frac{1}{4m_1^{p,\nu}},$$

which concludes the proof. □

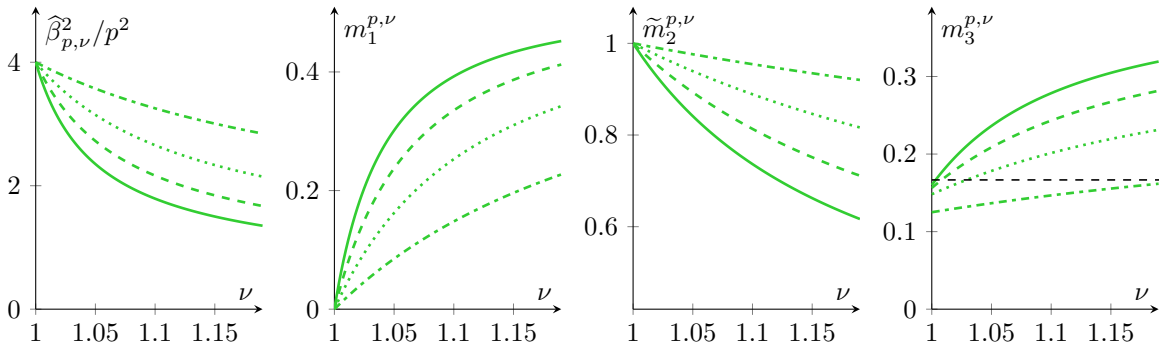


Figure 4.3.: Dependence of $\widehat{\beta}^2_{p,\nu}/p^2$, $m_1^{p,\nu}$, $\widetilde{m}_2^{p,\nu}$ in (4.14) and $m_3^{p,\nu}$ in (4.9a) on ν for $p = 2, 3, 4, 5$ (dash-dotted, dotted, dashed, solid). The black dashed horizontal line in the plot for $m_3^{p,\nu}$ represents the value $\frac{1}{6}$.

4.2.2. Qualitative behavior and a special choice of ν

After the derivation of the explicit values $\beta_{p,\nu}$, $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, $\widetilde{m}_2^{p,\nu}$, $m_3^{p,\nu}$, and $m_4^{p,\nu}$, we investigate their dependency on the stabilization parameter ν and the polynomial degree $p \in \mathbb{N}$. Moreover, we analyze the influence of the special choice of $\nu_{p,\eta}$ defined in (4.3). For a more concise presentation all of the subsequent calculations are postponed to Sections A.1 and A.2.

In Figure 4.3 the constants $\widehat{\beta}_{p,\nu}$ (scaled by p^2), $m_1^{p,\nu}$, $\widetilde{m}_2^{p,\nu}$ given in Theorem 4.9, and $m_3^{p,\nu}$ given in Theorem 4.5 are plotted against the stabilization parameter $\nu \geq 1$ for $p = 2, 3, 4, 5$. We observe that the stability constant $m_1^{p,\nu}$ improves, whereas $\widetilde{m}_2^{p,\nu}$ and $\widehat{\beta}^2_{p,\nu}$ degrade with increasing ν. In fact, in Lemmas A.4 and A.5 we show that these constants are monotone in ν for every $p \in \mathbb{N}$. For $\beta_{p,\nu}$ a similar behavior as for $\widehat{\beta}_{p,\nu}$ can be observed (and proven in Lemma A.4), since they only differ marginally; see (4.7) and (4.14b). In the limit cases, i.e., for $\nu \to 1$ and $\nu \to \infty$, respectively, these constants take the following values

$$\begin{aligned} &\lim_{\nu\to 1} \beta^2_{p,\nu} = 4p^2, \quad \lim_{\nu\to\infty} \beta^2_{p,\nu} = 4p, && \lim_{\nu\to 1} \widehat{\beta}^2_{p,\nu} = 4p^2, \quad \lim_{\nu\to\infty} \widehat{\beta}^2_{p,\nu} = 2p, \\ &\lim_{\nu\to 1} m_1^{p,\nu} = 0, \quad \lim_{\nu\to\infty} m_1^{p,\nu} = \tfrac{1}{2}, && \lim_{\nu\to 1} \widetilde{m}_2^{p,\nu} = 1, \quad \lim_{\nu\to\infty} \widetilde{m}_2^{p,\nu} = \tfrac{1}{p} \end{aligned}$$

(recall that $T_p'(1) = p^2$). We notice that, in accordance with the first plot in Figure 4.3, the values for $\widehat{\beta}_{p,\nu}$, and also for $\beta_{p,\nu}$, drastically decay for growing ν, which renders the LFC scheme (4.1) considerably less efficient due to stronger step-size restrictions; see Section 4.4.2. More precisely, we obtain $2p \leq \widehat{\beta}^2_{p,\nu} \leq \beta^2_{p,\nu} \leq 4p^2$ because of the monotonicity and the definition of $\widehat{\beta}_{p,\nu}$ and $\beta_{p,\nu}$.

Further, we see in Figure 4.3 that the error constant $m_3^{p,\nu}$ given in (4.9a) grows with increasing ν. We believe that this constant is also monotone in ν but unfortunately we could not prove it. For $m_3^{p,\nu}$ we have the limits

$$\lim_{\nu\to 1} m_3^{p,\nu} = \frac{p^2-1}{6p^2}, \qquad \lim_{\nu\to\infty} m_3^{p,\nu} = \frac{p-1}{2p}, \tag{4.17}$$

where we used (B.15) for the values for $T_p^{(k)}(1)$, $k = 1, 2$. From this we obtain for $p \geq 2$ that

$$\tfrac{1}{8} \leq m_3^{p,1} < \tfrac{1}{6} \qquad \text{and} \qquad \tfrac{1}{4} \leq \lim_{\nu\to\infty} m_3^{p,\nu} < \tfrac{1}{2}.$$

Since $m_3^{p,\nu}$ continuously depends on ν, we thus have

$$[\tfrac{1}{6}, \tfrac{1}{4}] \subset \{m_3^{p,\nu} \mid \nu \geq 1\}$$

for every $p \geq 2$. Hence, because of the continuity of $m_3^{p,\nu}$ in ν for every $p \geq 2$ there exists a $\nu = \nu_{p*}$ such that $m_3^{p,\nu_{p*}} = \frac{1}{6}$, yielding a fourth-order scheme for $\mathbf{g} \equiv 0$; see Remark 4.6 and Theorem 3.54. Note that ν_{p*} tends to 1 for $p \to \infty$, since $\lim_{p\to\infty} m_3^{p,\nu} = \frac{1}{6}$ for $\nu = 1$. Moreover, we see in the plot of $m_3^{p,\nu}$ that the constant is smaller than $\frac{1}{3}$ for a relatively large set of $\nu \geq 1$ (depending on p). Thus, as mentioned in the comments after Theorem 3.54, the LFC scheme has in these cases a smaller error constant than the leapfrog scheme for linear problems (2.12) with $\mathbf{g} \equiv 0$.

For $m_4^{p,\nu}$ one observes a similar behavior as for $m_3^{p,\nu}$. In particular, the constant also seems to be monotonically increasing in ν. The limits are given by

$$\lim_{\nu\to 1} m_4^{p,\nu} = \frac{(p^2-1)(p^2-4)}{360p^4}, \qquad \lim_{\nu\to\infty} m_4^{p,\nu} = \frac{(p-1)(p-2)}{24p^2},$$

where we again recall (B.15) for the values of $T_p^{(k)}(1)$, $k = 1, 2, 3$. However, since this values are only of interest for $\nu = \nu_{p*}$ and $\lim_{p\to\infty} \nu_{p*} = 1$, we approximately have $m_4^{p,\nu_{p*}} \approx m_4^{p,1}$ for p large.

We now turn towards the special choice $\nu = \nu_{p,\eta}$ defined in (4.3) which is motivated by stabilized/damped Runge–Kutta–Chebyshev methods; see, e.g., [VHS90, HV03]. From Figure 4.3 we observe that for a fixed $\nu > 1$ the constants strongly depend on the polynomial degree p of the underlying Chebyshev polynomial. However, since we want to have stability and error constants which do not (strongly) depend on the chosen LFC polynomial, we want to get rid of this dependency on p. For Runge–Kutta–Chebyshev methods a similar problem occurs, which is resolved by using the same scaling of the stabilization parameter (with $\eta^2/2$ replaced by a single constant). In fact, the next lemma shows that the special choice $\nu = \nu_{p,\eta}$ remedies this dependency on p for the (stability) constants $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, $\widetilde{m}_2^{p,\nu}$ (and thus also $\beta_{p,\nu}$) in a satisfactory manner.

Lemma 4.12. *Let $p \in \mathbb{N}$ and $\eta > 0$. For $\nu = \nu_{p,\eta}$ defined in (4.3) we have*

$$\frac{\widehat{\beta}_{p,\nu}^2}{4p^2} \geq \frac{(1+\frac{1}{4}\eta^2)^{1/2}}{1+\frac{1}{2}\eta^2}, \qquad m_1^{p,\nu} \geq \frac{\eta^2}{4+2\eta^2}, \qquad \widetilde{m}_2^{p,\nu} \geq (1+\tfrac{1}{4}\eta^2)^{-1/2}, \tag{4.18}$$

with $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ defined in (4.14).

A proof of this lemma is contained in Lemma A.7. Here, we only state estimates from below for these constants, since the analysis in the previous chapter and also in the next one only requires estimates from above of $1/m_1^{p,\nu}$ and $1/\widetilde{m}_2^{p,\nu}$. Note that the bounds hold for all $\eta > 0$ and $p \in \mathbb{N}$. Further, since $\widetilde{m}_3^{p,\nu} = (4m_1^{p,\nu})^{-1}$, we directly obtain

$$\widetilde{m}_3^{p,\nu} \leq \frac{2+\eta^2}{2\eta^2}.$$

In Figure 4.4 the constants for this special choice for ν are plotted. We observe that in contrast to Figure 4.3 the values for $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ are much closer to each other for different polynomial degrees if η is not too large. Moreover, the uniform bounds in p given

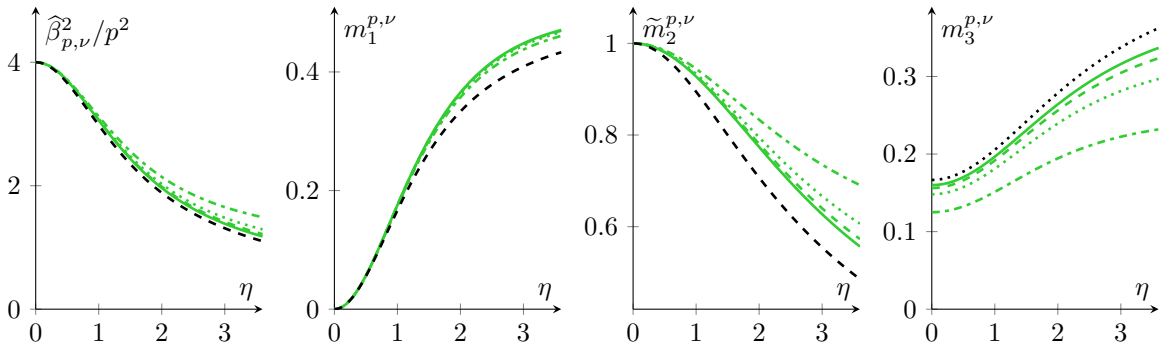


Figure 4.4.: Dependence of $\widehat{\beta}^2_{p,\nu}$, $m_1 = \widetilde{m}_1$, $\widetilde{m}_2$ in (4.14) and m_2 in (4.15) on $\nu_{p,\eta} = 1 + \frac{\eta^2}{2p^2}$ for the LFC polynomials (4.1b) with $p = 2, 3, 4, 5$ (dash-dotted, dotted, dashed, solid). The black dashed lines represents the lower bounds given in Lemma 4.12, the black dotted line the conjectured one in (A.15).

in (4.18) seem to be quite good approximations at least for small values for η. For the error constant $m_3^{p,\nu}$ the special choice $\nu_{p,\eta}$ yields no benefit compared to a general ν, which is not surprising because of (4.17). We refer to Section A.2 for more information about the dependency of these constants on $\nu_{p,\eta}$ and p.

Remark 4.13. The same bound for $m_1^{p,\nu}$ in (4.18) is proven in [GMS21, Lemma A.4]. In this lemma the authors also state a bound for $\widetilde{m}_2^{p,\nu}$, which is unfortunately wrong since it relies on the erroneous estimate (A.2) in [GMS21, Lemma A.1]. This can be seen from the estimates (A.3) and (A.7) in [GMS21], which contradict each other (only (A.3) is true). For the same reason the bound they state for $\widetilde{m}_3^{p,\nu}$ does not hold true. ◇

4.3. Equivalence to the leapfrog scheme in specific cases

In this section we focus on the LFC scheme (4.1) with the unstabilized polynomials (4.1b), i.e., $\nu = 1$, applied to linear problems (2.12) with $\mathbf{g} \equiv 0$. For this, we have the following remarkable result.

Theorem 4.14. *Let* $\mathbf{g} \equiv 0$. *For* $k, m \in \mathbb{N}$ *we denote by*

(a) $\mathbf{q}_k$ *the solution of the LFC scheme* (4.1) *with starting value* (4.1c), $\nu = 1$, *and polynomial degree* $p \in \mathbb{N}$ *after* k *time steps with step size* τ *and by*

(b) $\mathbf{q}_{k,*}$ *the solution of the leapfrog scheme* (2.20) *after* k *time steps with step size* τ_*.

If $\tau \leq \tau_{\mathrm{SSR}} = 2p/\|\mathbf{L}\|^{1/2}$ *and* $\tau_* = \tau/p$, *we have*

$$\mathbf{q}_n = \mathbf{q}_{pn,*}, \qquad n = 1, 2, \dots \ .$$

The theorem states that the leapfrog scheme (2.20) applied to linear homogeneous problems yields the same approximations as the unstabilized LFC schemes with the special starting value (4.1c) if the step sizes are chosen correctly. We point out that the equivalence only holds for $\nu = 1$ and it fails to be true if we choose other starting values for the leapfrog or the LFC scheme.

Proof. The proof relies on the representation formula of the numerical solution in Theorem 3.18 by inserting the special starting value (4.1c); cf. (3.78).

We first recall that for $p = 1$ the scheme (4.1a), (4.1c) comprises the leapfrog method (2.20), since $P_1(z) = z$. Moreover, the step-size restriction $\tau^2 \leq \tau_{\mathrm{SSR}}^2 = 4p^2/\|\mathbf{L}\|$ for the unstabilized LFC scheme implies the weaker step-size restriction of the leapfrog scheme $\tau_*^2 = \tau^2/p^2 \leq 4/\|\mathbf{L}\|$; see Examples 3.11 and 3.26 or Lemma 2.16. Hence, (3.78) holds for the leapfrog scheme. In particular, we have due to $P_1' \equiv 1$

$$\mathbf{q}_{k,*} = \cos(k\mathbf{\Phi}_*)\mathbf{q}_0 + \tau_* \frac{\sin(k\mathbf{\Phi}_*)}{\sin \mathbf{\Phi}_*}\dot{\mathbf{q}}_0$$

with a matrix $\mathbf{\Phi}_*$ with spectrum in $[0, \pi]$ satisfying $\cos \mathbf{\Phi}_* = \mathbf{I} - \frac{1}{2}\tau_*^2\mathbf{L}$.

From (3.23b) and definition (4.1b) of $P_p(\tau^2\mathbf{L})$ for $\nu = 1$ we get with (B.10)

$$\cos(\mathbf{\Phi}) = \mathbf{I} - \tfrac{1}{2}P_p(\tau^2\mathbf{L}) = T_p\Big(\mathbf{I} - \tfrac{1}{2p^2}\tau^2\mathbf{L}\Big) = \cos\Big(p \arccos(\mathbf{I} - \tfrac{1}{2}\tau_*^2\mathbf{L})\Big) = \cos(p\mathbf{\Phi}_*).$$

From this, we obtain for $n \in \mathbb{N}_0$

$$\cos(n\mathbf{\Phi}) = \cos(np\mathbf{\Phi}_*), \qquad \frac{\sin(n\mathbf{\Phi})}{\sin \mathbf{\Phi}} = \frac{\sin(np\mathbf{\Phi}_*)}{\sin(p\mathbf{\Phi}_*)},$$

by using induction arguments together with the formulae (B.5). Furthermore, we get from (4.6) and (B.12)

$$P_p'(\tau^2\mathbf{L}) = \frac{1}{T_p'(1)} T_p'\left(\mathbf{I} - \tfrac{1}{2p^2}\tau^2\mathbf{L}\right) = \frac{1}{p^2} T_p'(\cos \mathbf{\Phi}_*) = \frac{\sin(p\mathbf{\Phi}_*)}{p \sin \mathbf{\Phi}_*}.$$

Inserting these identities in (3.78) for the LFC scheme yields

$$\mathbf{q}_n = \cos(n\mathbf{\Phi})\mathbf{u}_0 + \tau \frac{\sin(n\mathbf{\Phi})}{\sin(\mathbf{\Phi})} P_p'(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 = \cos(np\mathbf{\Phi}_*)\mathbf{u}_0 + \frac{\tau}{p}\frac{\sin(np\mathbf{\Phi}_*)}{\sin(\mathbf{\Phi}_*)}\dot{\mathbf{q}}_0 = \mathbf{q}_{np,*},$$

which completes the proof. □

As a direct consequence of this theorem we obtain

$$\mathbf{q}_{p,*} = \mathbf{q}_1 = \mathbf{q}_0 + \tau P_p'(\tau^2\mathbf{L})\dot{\mathbf{q}}_0 - \tfrac{1}{2}\tau^2 P_p(\tau^2\mathbf{L})\mathbf{q}_0,$$

if $\nu = 1$ and $\mathbf{g} = 0$. This shows that our approach is indeed very similar to *impulse methods* (multiple time-stepping schemes), if the stiff linear part $\ddot{\mathbf{q}} = -\mathbf{L}\mathbf{q}$ in the "oscillation step" is solved by p time steps with the leapfrog scheme; see, e.g., [HLW06, Equation (XIII.1.14)] or [GSS99].

4.4. Implementation and efficiency

This section is devoted to the efficient implementation of the LFC schemes (4.1). Further, we heuristically compare the computational effort of these schemes to the leapfrog scheme (2.20) and the modified θ-schemes discussed in Section 3.6. Because of its practical relevance we consider the implementation and efficiency for the application of the LFC schemes to the semilinear differential equation (2.2) with a general mass matrix $\mathbf{M}$; see Remarks 2.2 and 3.4.

In the following we focus on the implementation and efficiency of the two-step scheme (4.1) but similar considerations hold for the corresponding one-step schemes (3.8) and (3.9). In particular, the (main) computational costs coincide since we can reuse the computations in the last step of the one-step schemes in the first step of the next time step; cf. the discussion for the leapfrog scheme before Remark 2.10.

4.4.1. Implementation

The implementation of one time step of the two-step LFC scheme (4.1a) applied to the semilinear problem (2.2) is stated in Algorithm 4.1. We treat the case $p = 1$ separately, since in this case the scheme reduces to the leapfrog scheme (2.20a) for every $\nu \geq 1$; see (4.5). Obviously, for $p \geq 2$ the main challenge consists in the computation of the expression $\widehat{P}_p(\tau^2\mathbf{L})\mathbf{q}$.

We point out that the algorithm slightly changes if the matrix-vector product $\mathbf{g}_n^* = \mathbf{M}\mathbf{g}_n$ is given explicitly, which is the case, for instance, in the modified FPUT problem in Section 2.2.1 but also possible in the spatially discretized wave equation in Section 2.2.2 if $\mathbf{g}$ depends only on time. In such situations one can replace the second and third line with $\widehat{\mathbf{v}} = -\mathbf{L}\mathbf{q}_n + \mathbf{g}_n^*$ and the solving of $\mathbf{M}\widetilde{\mathbf{v}} = \widehat{\mathbf{v}}$, respectively.

Algorithm 4.1.: Computation of nth time step of two-step LFC scheme (4.1a) applied to semilinear problems (2.2).

1: Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
2: Solve $\mathbf{M}\widehat{\mathbf{v}} = -\mathbf{L}\mathbf{q}_n$
3: $\widetilde{\mathbf{v}} = \widehat{\mathbf{v}} + \mathbf{g}_n$
4: **if** $p = 1$ **then**
5: $\quad \mathbf{v} = \widetilde{\mathbf{v}}$
6: **else**
7: $\quad$ Compute $\mathbf{v} = \widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ by Algorithm 4.2 below
8: **end if**
9: $\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n-1} + \tau^2\mathbf{v}$

For the computation of $\widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ one could naively use Horner's method. However, since the LFC polynomials P_p are based on the Chebyshev polynomials T_p evaluated on $[-\nu, \nu]$, it is beneficial in terms of stability to employ the linear three-term recurrence relation (B.9) of the Chebyshev polynomials. In addition, for Horner's method we would have to explicitly compute the coefficients in the monomial basis, which vary for different p and ν; cf. Lemma A.9.

We first derive a three-term recurrence relation for P_p, which allows us to deduce a recurrence relation for $\widehat{P}_p$.

Lemma 4.15. *Let $p \in \mathbb{N}$, $k \in \mathbb{N}_0$. The polynomials $P_{k,p}\colon \mathbb{R} \to \mathbb{R}$, defined by*

$$P_{k,p}(z) = 2 - \frac{2}{T_k(\nu)} T_k\Big(\nu - \frac{z}{\alpha_p}\Big), \tag{4.19}$$

satisfy the linear recurrence relation

$$P_{0,p}(z) = 0, \qquad P_{1,p}(z) = \frac{2}{\alpha_p \nu} z,$$
$$T_k(\nu)P_{k,p}(z) = 2\nu T_{k-1}(\nu)P_{k-1,p}(z) + \frac{2}{\alpha_p} T_{k-1}(\nu)\, z\,(2 - P_{k-1,p}(z)) - T_{k-2}(\nu)P_{k-2,p}(z),$$

for $k \geq 2$.

Proof. The statements for $k = 0$ and $k = 1$ are trivially satisfied because of $T_0 \equiv 1$ and $T_1(x) = x$, respectively; see Definition B.1.

Let $k \geq 2$. First, we observe that (4.19) can be rewritten as

$$2T_k(x) = T_k(\nu)\big(2 - P_{k,p}(z)\big), \tag{4.20}$$

where we again make use of the variable transformation (4.8) for a clearer presentation. We further deduce from (4.19) and the recurrence relation of Chebyshev polynomials (B.9)

$$T_k(\nu)P_{k,p}(z) = 2T_k(\nu) - 2T_k(x) = 2T_k(\nu) - 4xT_{k-1}(x) + 2T_{k-2}(x).$$

Inserting (4.20) in this formula yields again with (4.8)

$$\begin{aligned} T_k(\nu)P_{k,p}(z) &= 2T_k(\nu) - 2xT_{k-1}(\nu)\big(2 - P_{k-1,p}(z)\big) + T_{k-2}(\nu)\big(2 - P_{k-2,p}(z)\big) \\ &= 2T_k(\nu) - 2\nu T_{k-1}(\nu)\big(2 - P_{k-1,p}(z)\big) + \tfrac{2}{\alpha_p}T_{k-1}(\nu)\, z\,\big(2 - P_{k-1,p}(z)\big) \\ &\quad + T_{k-2}(\nu)\big(2 - P_{k-2,p}(z)\big), \end{aligned}$$

which finishes the proof, since we have $2T_k(\nu) - 4\nu T_{k-1}(\nu) + 2T_{k-2}(\nu) = 0$ by the Chebyshev recurrence relation (B.9). □

By definition we have $P_p = P_{p,p}$ for all $p \in \mathbb{N}$. Thus, the recurrence relation for the polynomials (4.19) can be used for the implementation of the alternative two-step scheme (3.7) if equipped with the LFC polynomials; see [CHS20, Section 6.1]. Another application are linear problems (2.12) with $\mathbf{g} \equiv 0$, since in this case (4.1a) reduces to

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = -\tau^2 \widehat{P}_p(\tau^2\mathbf{L})\mathbf{L}\mathbf{q}_n = -P_p(\tau^2\mathbf{L})\mathbf{q}_n, \qquad n = 1, 2, \dots.$$

We further draw attention to the fact that for all $p, k \in \mathbb{N}$ we have $P_{k,p}(z) = cz + z^2 Q(z)$ with a constant $c = c(k,p) > 0$ and a polynomial Q of degree $k-2$ ($Q \equiv 0$ for $k = 1$), which is a direct consequence of the definition of the recurrence relation. Hence, we obtain the following.

Corollary 4.16. *Let $p \in \mathbb{N}$, $k \in \mathbb{N}_0$. The polynomials $\widehat{P}_{k,p}\colon \mathbb{R} \to \mathbb{R}$, given by $\widehat{P}_{k,p}(z) = \frac{P_{k,p}(z)}{z}$, satisfy the linear recurrence relation*

$$\begin{aligned} \widehat{P}_{0,p}(z) = 0, \qquad \widehat{P}_{1,p}(z) &= \frac{2}{\alpha_p \nu}, \\ T_k(\nu)\widehat{P}_{k,p}(z) = 2\nu T_{k-1}(\nu)\widehat{P}_{k-1,p}(z) + \frac{2}{\alpha_p}T_{k-1}(\nu)\big(2 - z\widehat{P}_{k-1,p}(z)\big) &- T_{k-2}(\nu)\widehat{P}_{k-2,p}(z), \end{aligned}$$

for $k \geq 2$.

Similarly as before the corollary yields that $\widehat{P}_p = \widehat{P}_{p,p}$ for every $p \in \mathbb{N}$. Hence, we have a three-term recurrence relation yielding the polynomials $\widehat{P}_p$ which allows us to compute $\mathbf{v} = \widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ in Algorithm 4.1 in a stable and efficient way. Algorithm 4.2 presents the details of the computation. Note that $\mathbf{w}_k = \widehat{P}_{k,p}(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$. Further, we emphasize that for a fixed $\nu \geq 1$, the parameters $T_0(\nu), \dots, T_p(\nu)$ have to be computed only once by means of the Chebyshev recurrence relation (B.9). For the same reason we compute α_p only once, where we additionally employ (B.20) together with the recurrence relation (B.18) to obtain $T_p'(\nu)$.

Algorithm 4.2.: Computation of $\widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ in Algorithm 4.1 for LFC polynomials (4.1b).

1: $\mathbf{w}_0 = 0$, $\mathbf{w}_1 = \frac{2}{\alpha_p \nu}\widetilde{\mathbf{v}}$
2: **for** $k = 2, \dots, p$ **do**
3: Solve $\mathbf{M}\widetilde{\mathbf{w}}_{k-1} = \mathbf{L}\mathbf{w}_{k-1}$
4: $\mathbf{w}_k = 2\nu \frac{T_{k-1}(\nu)}{T_k(\nu)}\mathbf{w}_{k-1} + \frac{2}{\alpha_p}\frac{T_{k-1}(\nu)}{T_k(\nu)}\left(2\widetilde{\mathbf{v}} - \tau^2\widetilde{\mathbf{w}}_{k-1}\right) - \frac{T_{k-2}(\nu)}{T_k(\nu)}\mathbf{w}_{k-2}$
5: **end for**
6: $\mathbf{v} = \mathbf{w}_p$

The implementation of the starting values (4.1c) as well as (4.2) is done in a similar way as for the the two-step scheme (4.1a) in Algorithm 4.1. Clearly, if we use the special starting value (4.1c), we additionally need to compute $P_p'(\tau^2\mathbf{M}^{-1}\mathbf{L})\dot{\mathbf{q}}_0$ once. As for $\widehat{\Psi}$ we do this via a three-term recurrence relation.

Lemma 4.17. *Let $p \in \mathbb{N}$, $k \in \mathbb{N}$, and U_p be the pth Chebyshev polynomial of the second kind, defined in Definition B.10. The polynomials $P_{k,p}^\star \colon \mathbb{R} \to \mathbb{R}$, defined by*

$$P_{k,p}^\star(z) = \frac{1}{U_{k-1}(\nu)} U_{k-1}\Big(\nu - \frac{z}{\alpha_p}\Big),$$

satisfy the linear recurrence relation

$$P_{1,p}^\star(z) = 1, \qquad P_{2,p}^\star(z) = \frac{1}{2\nu} 2\Big(\nu - \frac{z}{\alpha_p}\Big) = 1 - \frac{z}{\alpha_p \nu},$$
$$U_{k-1}(\nu) P_{k,p}^\star(z) = 2\nu U_{k-2}(\nu) P_{k-1,p}^\star(z) - \frac{2}{\alpha_p} U_{k-2}(\nu) z P_{k-1,p}^\star(z) - U_{k-3}(\nu) P_{k-2,p}^\star(z),$$

for $k \geq 3$.

Proof. The recurrence relation immediately follows from the Chebyshev recurrence relation (B.18) for U_p. □

With this lemma we obtain a three-term recurrence relation for P_p' because we have with (4.6) and (B.20) that

$$P_p'(z) = \frac{1}{U_{p-1}(\nu)} U_{p-1}\Big(\nu - \frac{z}{\alpha_p}\Big) = P_{p,p}^\star(z).$$

For the computation of $P_p'(\tau^2\mathbf{M}^{-1}\mathbf{L})\dot{\mathbf{q}}_0$ one then proceeds similarly as for $\widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ in Algorithm 4.2. As before, the values for $U_k(\nu)$, $k = 0, 1, \dots, p-1$, and α_p have to be computed only once beforehand.

4.4.2. Costs and efficiency

For the comparison of the efficiency of the leapfrog scheme (2.20), the LFC scheme (4.1), and the modified θ-schemes, where (3.1) is equipped with (3.4), we focus on the two-step schemes. The costs of the starting value are neglectable because they have to be computed only once. Recall that we assume that the evaluation of $\mathbf{g}$ is approximately as expensive as or more expensive than one matrix-vector multiplication with $\mathbf{L}$.

Table 4.1.: Comparison of main costs per time step of leapfrog scheme, LFC scheme (with polynomial of degree p), and modified θ-scheme in terms of matrix-vector multiplications (MVM), solutions of linear systems, and evaluations of $\mathbf{g}$, if implemented as in Algorithms 4.1 and 3.1.

leapfrog scheme (2.20a)	LFC scheme (4.1a)	modified θ-scheme (3.1a), (3.4)
1 evaluation of $\mathbf{g}$	1 evaluation of $\mathbf{g}$	1 evaluation of $\mathbf{g}$
1 MVM with $\mathbf{L}$	p MVMs with $\mathbf{L}$	1 MVM with $\mathbf{L}$ and $\mathbf{M}$
1 linear system with $\mathbf{M}$	p linear systems with $\mathbf{M}$	1 linear system with $\mathbf{M} + \tau^2\theta\mathbf{L}$

Remark 4.18 (Efficiency versus accuracy). We mainly consider efficiency as the ratio between the maximum step size, for which the schemes are stable, and the related cost, although the size of τ is also related to the accuracy of the approximations. This is motivated by the fact that in applications errors obtained from the maximum step size for which the leapfrog scheme is stable are often smaller than required. In particular, for differential equations stemming from the space discretization of a partial differential equation the error of the space discretization is often larger than the error of the numerical time integration scheme; see the numerical examples at the end of this chapter. ◇

We start by looking at the main computational effort, stemming from matrix-vector multiplications, solving of linear systems, and evaluations of $\mathbf{g}$. In Table 4.1 the main costs per time step of the LFC scheme is compared to the costs of one time step of the leapfrog scheme and the modified θ-schemes. We see that for all three schemes the function $\mathbf{g}$ is evaluated only once in accordance with the construction of the schemes, whereas the number of matrix-vector multiplication with $\mathbf{L}$ and the linear systems which has to be solved differ for the schemes. We point out that, if the matrix-vector product $\mathbf{g}_n^* = \mathbf{M}\mathbf{g}_n$ is given explicitly, the main costs stay the same for the leapfrog and the LFC scheme; see the comments above Algorithm 4.1. For the modified θ-scheme, however, one saves the multiplication with $\mathbf{M}$ in the first step in Algorithm 3.1.

For a fair comparison of the costs we next consider the maximum step sizes which yield stable schemes. For the modified θ-schemes we restrict ourselves to the reasonable choices $\theta \geq \frac{1}{4}$ yielding unconditionally stable schemes; cf. Section 3.6. For LFC schemes we consider the stronger step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ with $\widehat{\tau}_{\mathrm{SSR}}$ defined in (3.33). Using Lemma 4.12 for the special choice of $\nu = \nu_{p,\eta}$ given in(4.3) with $\eta > 0$ then yields the following sufficient step-size restriction

$$\tau^2 \leq \frac{C_\eta^2 4p^2}{\|\mathbf{L}\|}, \qquad C_\eta^2 = \frac{(1 + \frac{1}{4}\eta^2)^{1/2}}{1 + \frac{1}{2}\eta^2} < 1.$$

Clearly, for $\eta > 0$ small we have $C_\eta \approx 1$. Hence, since we require $\tau^2 < 4/\|\mathbf{L}\|$ for the leapfrog scheme to gain stability, the LFC scheme (4.1) allows an approximately p times larger time step than the leapfrog scheme (2.20).

We now combine these considerations to show the benefits of the LFC scheme. For this, we first compare the overall cost of the LFC scheme to the leapfrog scheme. We neglect the computational cost of solving with $\mathbf{M}$ for simplicity. Moreover, as noted in Remark 2.10, linear systems with $\mathbf{M}$ are often cheap to solve.

Let $\varrho = \mathcal{C}_{\mathbf{g}}/\mathcal{C}_{\mathbf{L}}$ be the ratio of the cost of one evaluation of $\mathbf{g}$ to the cost of one matrix-vector multiplication with $\mathbf{L}$. The main effort of the leapfrog and the LFC scheme per time step is

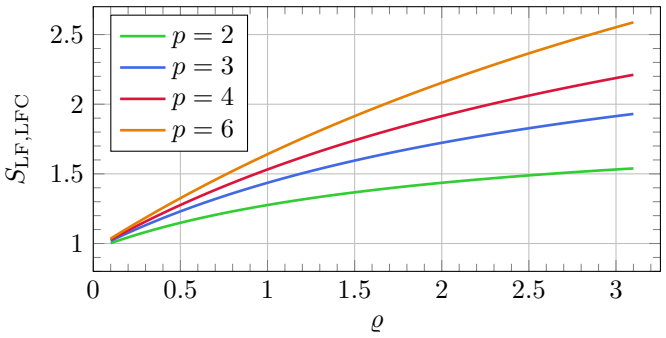


Figure 4.5.: Relative theoretical speedup $S_{\mathrm{LF,LFC}}$ defined in (4.21) of LFC scheme (4.1) compared to the leapfrog scheme (2.20) for polynomial degrees $p = 2, 3, 4, 6$. With ϱ we denote the ratio of the cost of one evaluation of $\mathbf{g}$ to the cost of one matrix-vector multiplication with $\mathbf{L}$ ($\eta = 0.5$).

then given by

$$\mathcal{C}_{\mathrm{LF}} = \mathcal{C}_{\mathbf{L}} + \mathcal{C}_{\mathbf{g}} = (1 + \varrho)\mathcal{C}_{\mathbf{L}} \qquad \text{and} \qquad \mathcal{C}_{\mathrm{LFC}} = p\mathcal{C}_{\mathbf{L}} + \mathcal{C}_{\mathbf{g}} = (p + \varrho)\mathcal{C}_{\mathbf{L}},$$

respectively, where p is the polynomial degree of the used LFC polynomial. Since for the LFC scheme we can choose a step size which is (at least) $C_\eta\, p$ larger than the one of the leapfrog scheme, the LFC scheme requires only $N/(C_\eta\, p)$ time steps to reach a certain simulation time T if the leapfrog scheme requires minimum N time steps. Thus, we get the following relative speedup of the LFC scheme compared to the leapfrog scheme

$$S_{\mathrm{LF,LFC}} = \frac{N\mathcal{C}_{\mathrm{LF}}}{N/(C_\eta\, p)\,\mathcal{C}_{\mathrm{LFC}}} = \frac{(C_\eta\, p)\,\mathcal{C}_{\mathrm{LF}}}{\mathcal{C}_{\mathrm{LFC}}} = C_\eta \frac{1 + \varrho}{1 + \varrho/p}, \tag{4.21}$$

Note that $S_{\mathrm{LF,LFC}} > 1$ means that the LFC scheme is less costly than the leapfrog scheme.

In Figure 4.5 the relative speedup $S_{\mathrm{LF,LFC}}$ is plotted against ϱ for $p = 2, 3, 4, 6$ and $\eta = 0.5$, yielding $C_\eta \approx 0.9572$. We observe that already for $\varrho \approx \frac{1}{2}$ one gains efficiency compared to the leapfrog scheme. Moreover, an increase of the polynomial degree p leads to a further improvement of the efficiency. Clearly, the larger η is chosen, the smaller $\widehat{\beta}^2_{p,\nu}$, and, hence, the greater ϱ has to be.

Remark 4.19 (Choice of η). For a general function $\mathbf{g} \not\equiv 0$ we suggest $\eta = 0.5$ as initial guess for the stabilization parameter. This value yields a good trade-off between improved stability behavior and a step-size restriction which is only slightly smaller than the "optimal" one $\tau^2\|\mathbf{L}\| \leq \beta^2_{p,1} = 4p^2$. More precisely, the maximum step size, for which the scheme is stable, is only $\sim 4.43\%$ smaller than $\beta_{p,1}/\|\mathbf{L}\|^{1/2}$. Moreover, the value is large enough to compensate small to moderate instabilities occurring from the semilinearity; see Section 3.3.3 and especially Remark 3.34. ◇

For $\mathbf{g} \equiv 0$ there seems to be no benefit (or even a small disadvantage) of the LFC scheme (4.1), since the p times larger step size is fully compensated by p times more matrix-vector multiplications with $\mathbf{L}$. In particular, for $\nu = 1$ we have shown in Section 4.3 that the schemes are then equivalent. Nevertheless, we will see in the numerical examples in the next section that the LFC scheme with $\nu > 1$ can pay off in such a case because of smaller error constants; cf. Remark 4.6 and Theorem 3.54 with the subsequent comments.

Last, we point out that there is a breakeven point in which evaluating $\widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\tilde{\mathbf{v}}$ becomes more expensive than solving a linear system with $\mathbf{M}+\tau^2\theta\mathbf{L}$ and possibly computing a matrix-vector multiplication with $\mathbf{M}$. Hence, in such cases it is beneficial to use the modified θ-schemes with $\theta \geq \frac{1}{4}$ instead of the LFC scheme. However, as mentioned before, one has to take into account that with increasing step sizes the error of the approximations to the exact solution deteriorates and, thus, a too large step size does not pay off; cf. Remark 4.18.

4.5. Numerical examples

Concluding, we illustrate the theoretical findings on LFC schemes shown in this and the last chapter by some numerical examples. The first example and the ones for the acoustic wave equations are (slight) modifications and extensions from the ones in [CHS20, Section 7.1 and 7.2.1]. The codes for reproducing the numerical results are available on https://doi.org/10.5445/IR/1000147744.

4.5.1. Influence of starting value

In this first example we demonstrate the influence of the starting value on the stability behavior of the LFC scheme, or, more general, of the two-step scheme (3.1a). To show this, we consider the test problem (2.25) for $\omega > 0$, i.e., the harmonic oscillator. Recall that the solution q is bounded uniformly in time; cf. (2.15) and (2.16) with $c_{\mathrm{inv}} = \omega^{-1}$.

We now apply the two-step LFC scheme (4.1a), (4.1b) to this equation with four different starting values, namely

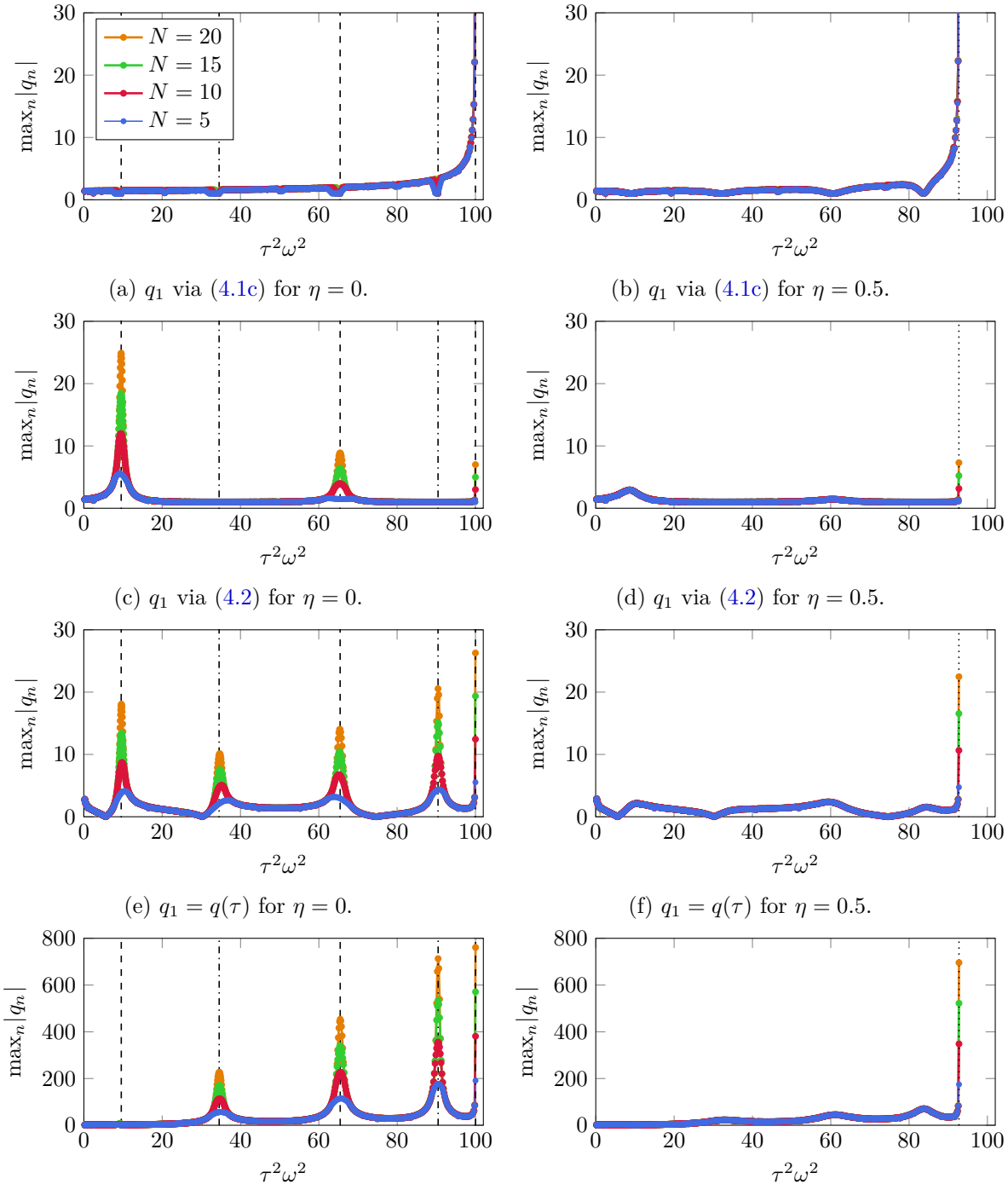
(i) the special one (4.1c) motivated in Section 3.5.1,
(ii) the starting value (4.2) proposed for the general two-step scheme (3.1),
(iii) the second-order Taylor approximation (2.20b) used in the leapfrog scheme (2.20),
(iv) and the exact starting value $q_1 = q(\tau)$.

We use the LFC polynomial of degree $p = 5$ with stabilization parameter $\eta = 0$ (unstabilized case) and $\eta = 0.5$ (stabilized case). For the harmonic oscillator we choose the initial values $q_0 = 1$ and $\dot{q}_0 = 1$ and the frequency $\omega = 1$.

In Figure 4.6 the absolute value of q_n is plotted for all variants against a range of step sizes $\tau > 0$ satisfying $\tau \leq \tau_{\mathrm{SSR}}$ defined in (3.22) i.e., we consider $0 < \tau^2\omega^2 = \tau^2 \leq \beta^2_{5,\nu}$. Note that for $\eta = 0$ we have $\nu = \nu_{p,\eta} = 1$ and, thus, $\beta^2_{5,\nu} = 100$.

In Figures 4.6a, 4.6c, 4.6e, and 4.6g the unstabilized LFC scheme is applied to the harmonic oscillator. We clearly see that in Figure 4.6a the approximations computed with the special starting value (4.1c) stay bounded uniformly in time if we stay away from $\beta^2_{5,1}$. In contrast to this we observe in Figure 4.6c that for the starting value (4.2) resonances appear at points $z = \tau^2\omega^2$, where $P_5(z) = 4$. Even worse, for the exact and the Taylor starting value the solution grows linearly in time at points $z = \tau^2\omega^2$, where $P_5(z) = 4$ or $P_5(z) = 0$. This perfectly fits to our analysis in Section 3.5.1, where we have shown that for general starting values one cannot achieve uniform bounds under the weaker step-size restriction.

In contrary, we see in Figures 4.6b, 4.6d, 4.6f, and 4.6h that for the stabilized LFC scheme all starting values yield uniformly bounded approximations if $\tau^2\omega^2 < \beta^2_{5,\nu_{p,\eta}}$. Nevertheless, the special starting value (4.1c), the starting value (4.2), as well as the exact starting value yield quantitatively better results than the Taylor starting value. This confirms our results in Section 3.3 for the stronger step-size restriction $\tau \leq \widehat{\tau}_{\mathrm{SSR}}$ defined in (3.33). We further see that



(a) q_1 via (4.1c) for $\eta = 0$.

(b) q_1 via (4.1c) for $\eta = 0.5$.

(c) q_1 via (4.2) for $\eta = 0$.

(d) q_1 via (4.2) for $\eta = 0.5$.

(e) $q_1 = q(\tau)$ for $\eta = 0$.

(f) $q_1 = q(\tau)$ for $\eta = 0.5$.

(g) q_1 via (2.20b) for $\eta = 0$.

(h) q_1 via (2.20b) for $\eta = 0.5$.

Figure 4.6.: Maxima of absolute values of the approximations q_n, $n = 1, \ldots, N$, of the LFC scheme (4.1a), (4.1b) plotted against a range of step sizes for different starting values and stabilization parameter $\eta = 0$ (Figures 4.6a, 4.6c, 4.6e, and 4.6g) and $\eta = 0.5$ (Figures 4.6b, 4.6d, 4.6f, and 4.6h). The dashed vertical and dash-dotted lines represent points $\tau^2\omega^2$ with $P_5(\tau^2\omega^2) = 4$ and $P_5(\tau^2\omega^2) = 0$, respectively. The dotted lines in Figures 4.6b, 4.6d, 4.6f, and 4.6h represent the point $\tau^2\omega^2 = \beta^2_{5,\nu_{p,\eta}}$.

for $\eta = 0.5$ the maximum step size we can choose to obtain approximations growing at least linearly in time is slightly smaller than the one for $\eta = 0$ in agreement with our results.

4.5.2. Modified Fermi–Pasta–Ulam–Tsingou problem

Next, we consider two numerical examples for the modification of the FPUT β-problem introduced in Section 2.2.1. For both examples we use $m = 200$ with $\mu_i = 1$ and $k_i = 99^2$ for all $i = 1, \dots, m$, whereas the values for β_i^* differ. For the starting values $\mathbf{q}_0 = (q_{0,1}, \dots, q_{0,m})^T$ and $\dot{\mathbf{q}}_0 = (\dot{q}_{0,1}, \dots, \dot{q}_{0,m})^T$ we use

$$\dot{q}_{0,i} = 0.5 \qquad \text{and} \qquad \dot{q}_{0,i} = (-1)^{i-1} \qquad \text{for all } i = 1, \dots, m.$$

Note that the choice $\mu_i = 1$ for all i results in $\mathbf{M} = \mathbf{I}_m$.

Linear case

With the first example we show that the LFC scheme converges with order four in case of $\mathbf{g} \equiv 0$ and $\nu = \nu_{p*}$ given in Remark 4.6. Thus, we set $\beta_i^* = 0$ for all i in the modified FPUT β-problem. The problem then reduces to a system of coupled harmonic oscillators.

For the time integration we use the leapfrog scheme (2.20) and the LFC scheme (4.1) with polynomial degree $p = 3$ and $p = 4$ and five different values for the stabilization parameter ν (either via $\nu_{p,\eta}$ in (4.3) or via ν_{p*}). For determining the errors we use as reference solution the numerical solution obtained by the leapfrog scheme with step size $\tau = 10^{-5}$.

In Figure 4.7 we plot the maximum error over all time steps up to the ending time $T = 1.2$. We clearly observe that the error is in general of order two unless we use the special choice $\nu = \nu_{p*}$, which yields order four in agreement with Theorem 3.54 and Remark 4.6. In addition, the error constants for the LFC schemes are smaller compared to the one of the leapfrog scheme because the choices for η yield $m_3^{p,\nu} \leq \frac{1}{3}$; cf. Theorem 3.54 and the subsequent paragraph. The good behavior of the error for $p = 4$ and $\eta = 0.5$ is due to the fact that $\nu_{4,\eta} \approx \nu_{4*}$. We further observe that the errors for the leapfrog scheme and the LFC scheme for $\eta = 0$ are only translated by a factor p confirming Theorem 4.14, which states that for $\mathbf{g} \equiv 0$ and $\eta = 0$ the leapfrog and the LFC scheme are equivalent. Moreover, we see that the maximum step size for which the LFC scheme is stable is approximately p times larger than the one for the leapfrog scheme in accordance with our theory.

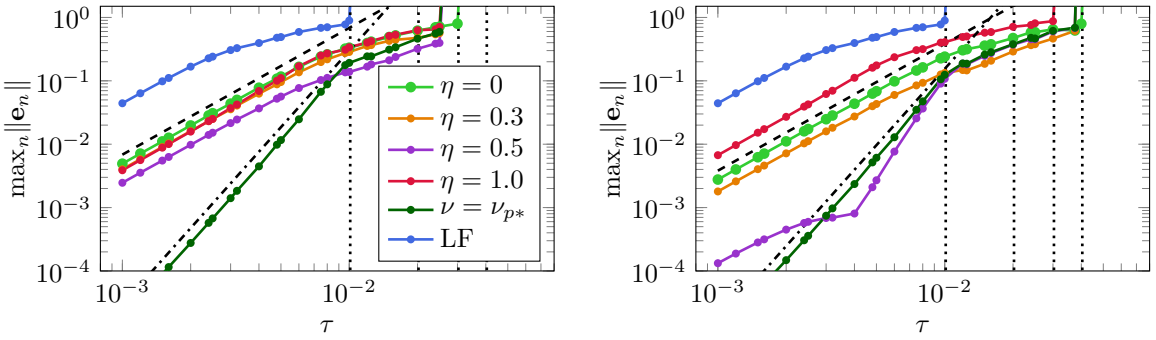
Semilinear case

In the second example we are interested in the behavior of the Hamiltonian (2.11) with U given in (2.17) over time. To incorporate the nonlinear effects, we choose a rather large value for all nonlinear spring constants, namely $\beta_i^* = 20$ for all $i = 1, \dots, m$.

In Figure 4.8 the relative error of the Hamiltonian (2.11),(2.17)

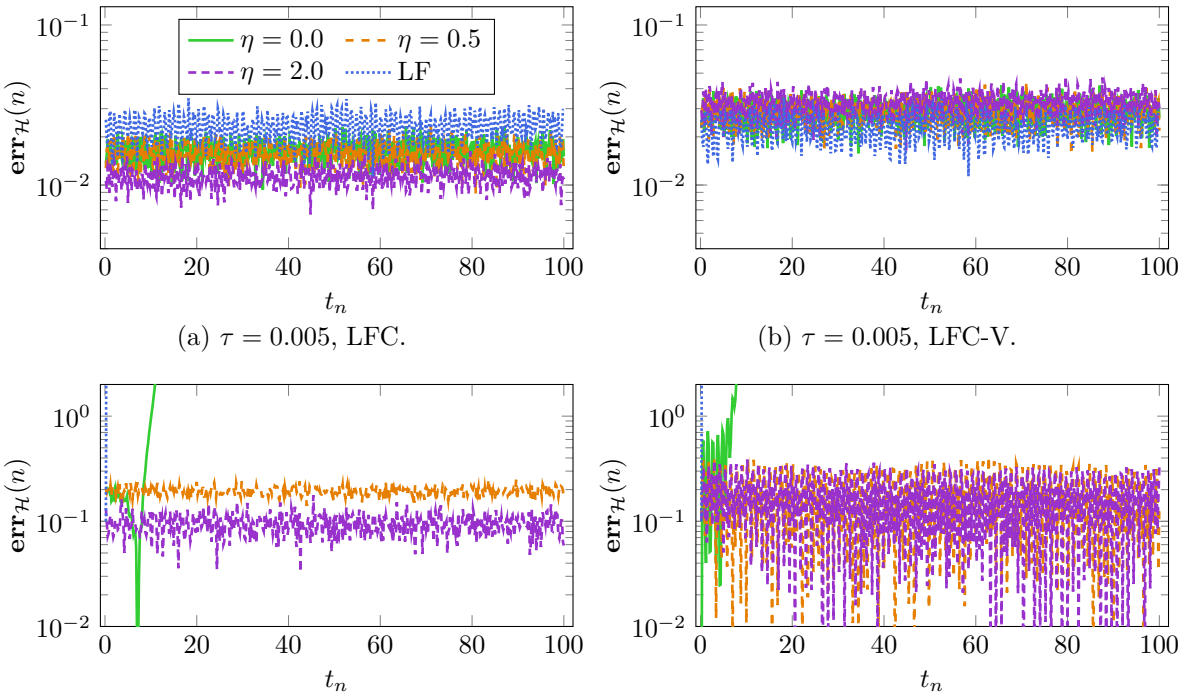
$$\mathbf{err}_{\mathcal{H}}(n) = \frac{|\mathcal{H}(\mathbf{p}_n, \mathbf{q}_n) - \mathcal{H}(\mathbf{p}_0, \mathbf{q}_0)|}{\mathcal{H}(\mathbf{p}_0, \mathbf{q}_0)} \tag{4.22}$$

is plotted over time until $T = 100$ for two different step sizes. As time integration schemes we employ the leapfrog scheme (2.20) and (variants of) the LFC schemes (4.1). Since the computation of the Hamiltonian requires approximations $\mathbf{p}_n$ to $\mathbf{p}(t_n)$, we use the one-step formulation (2.21) of the leapfrog scheme and the one-step formulations (3.8), (4.1b) of the LFC



(a) $p = 3$. (b) $p = 4$.

Figure 4.7.: Error for the numerical solution of a linear FPUT β-problem ($\beta_i^* = 0$) in Section 2.2.1 computed with the LFC scheme (4.1) up to $T = 1.2$. As stabilization parameter we use $\nu = \nu_{p,\eta}$ with $\eta = 0.0$, $\eta = 0.3$, $\eta = 0.5$, $\eta = 1$, and $\nu = \nu_{p*}$ (cf. Remark 4.6). The blue line represents the leapfrog (LF) scheme. The dashed black lines indicate order two, the dash-dotted order four. The vertical dotted lines correspond to integer multiples (1, 2, 3, 4) of the maximum step size for which the leapfrog scheme (2.20) is stable.



(a) $\tau = 0.005$, LFC. (b) $\tau = 0.005$, LFC-V.

(c) $\tau = 0.02$, LFC scheme. (d) $\tau = 0.02$, LFC-V .

Figure 4.8.: Relative error in Hamiltonian for the numerical solution of the FPUT β-problem ($\beta_i^* = 20$) computed with one-step formulations of the leapfrog scheme (blue, dotted), the LFC scheme (3.8), (4.1b), and its variant (3.9), (4.1b) (LFC-V) for two different step sizes τ. For the LFC polynomial P_p we use polynomial degree $p = 4$ and different values for the stabilization parameter ($\eta = 0.0$ (solid), $\eta = 0.5$ (dashed), $\eta = 2.0$ (densely dashed)). The relative error of the Hamiltonian is only plotted at times $t = 0.2k$, $k = 1, \ldots, 500$, for the sake of clarity.

scheme (4.1a), (4.1b) with starting value (4.2) as well as its variant (3.9), (4.1b) (abbreviated with LFC-V). For the LFC polynomial P_p we use $p = 4$ and different values for the stabilization parameter η. Recall that in Section 3.1.1 we have shown that the one-step formulation (3.8) is symplectic in contrast to its variant (3.9).

In Figures 4.8a and 4.8b we observe that for the small step size $\tau = 0.005$ the relative error does not have a drift in time for all schemes including the non-symplectic LFC-V scheme. The error for the LFC scheme is smaller than the one for the leapfrog scheme and improves for increasing η, whereas the error for the variant is slightly larger than for the leapfrog scheme independent of the choice of η.

For the larger step size $\tau = 0.02$ in Figures 4.8a and 4.8b the situation changes. Since the step-size restriction is violated, the leapfrog scheme is not stable anymore; cf. Figure 4.7. Further, the LFC scheme and the variant are unstable for $\eta = 0$ because for one eigenvalue λ of $\mathbf{L}$ we have that $\tau^2\lambda$ is near the first maximum point $z_1 \approx 9.3726$ of P_4, where $P_4(z_1) = 4$. This confirms the results in Section 3.3.3 stating that stabilization is not only required to prohibit linear growth in time of the numerical approximations but also (exponential) instabilities due to the semilinearity. In agreement with our theory, a sufficient increase of the stabilization parameter prevents this instability. In these cases the LFC scheme and its variant nearly preserve the Hamiltonian for long times without having a visible drift. Moreover, as for the smaller step size the error of the LFC scheme decreases for increasing η and is slightly smaller than the one of its variant.

4.5.3. Spatially discretized acoustic wave equation

Finally, we show two examples for the LFC schemes applied to (2.2) stemming from a space discretization of the acoustic wave equation (2.18) as described in Section 2.2.2. For both examples we consider the domain $\Omega = [0, 1]^2$ and material parameter $c \equiv 1$. The initial values and the semilinearity g are chosen in such a way that the exact solution of (2.18) is given by

$$q(t, x) = \sin(\pi x_1)\sin(\pi x_2)\Big(\cos(t\omega) + \sin(t\omega)\Big), \qquad x = (x_1, x_2)^T \in \Omega,\ t \geq 0, \tag{4.23}$$

with $\omega = \left(2\pi^2 + \delta\right)^{1/2}$, where the parameter $\delta \geq 0$ differs in these examples. This implies

$$g(t, x, q(t, x)) = -\delta\, q(t, x), \qquad x \in \Omega,\ t \geq 0. \tag{4.24}$$

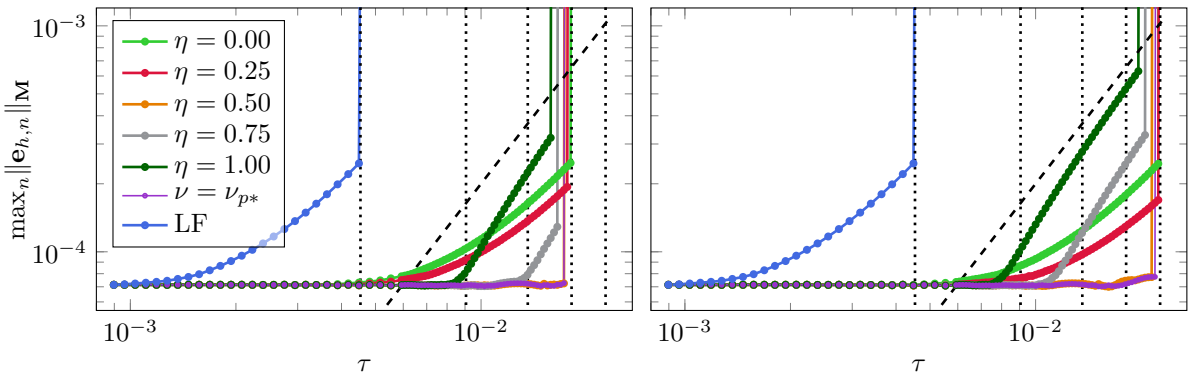
In particular, we have $g \equiv 0$ for $\delta = 0$.

As mentioned in Section 2.2.2, we discretize (2.18) in space with a symmetric interior penalty dG-FEM [Arn82, GSS06]. For both examples we use piecewise polynomials of degree 2 and the same (unstructured) triangulation of Ω consisting of 432 triangles with minimum and maximum diameter $h_{\min} \approx 0.0552$ and $h_{\max} \approx 0.1059$, respectively. Recall that the dG-FEM leads to (2.2) with a block-diagonal mass matrix $\mathbf{M}$, which can be inverted at low costs.

Although we have not analyzed the error of the full discretization in space and time of (2.18), we include the error from the space discretization in the subsequent error plots for a more representative illustration. More precisely, we consider the error

$$\mathbf{e}_{h,n} = \mathbf{q}_h(t_n) - \mathbf{q}_n \tag{4.25}$$

between the $L^2(\Omega)$-orthogonal projection $\mathbf{q}_h(t_n)$ of the exact solution onto the discontinuous Galerkin space and the approximation $\mathbf{q}_n$ of the leapfrog or LFC scheme. We emphasize that we measure the error in the weighted norm $\|\cdot\|_{\mathbf{M}}$ because of $\mathbf{M} \neq \mathbf{I}$; cf. Remark 2.2 and the precedent paragraph as well as Remark 3.4.



(a) $p = 4$. (b) $p = 5$.

Figure 4.9.: Error for the numerical solution of the (spatially discretized) wave equation (2.18) with $g \equiv 0$ (and exact solution (4.23)) plotted against the step size for LFC schemes (4.1) with polynomial degree $p = 4$ and $p = 5$. For the stabilization parameter we use $\nu = \nu_{p,\eta}$ with $\eta = 0$, $\eta = 0.25$, $\eta = 0.5$, $\eta = 0.75$, $\eta = 1$, and $\nu = \nu_{p*}$. The blue line represents the leapfrog scheme. The dashed black line indicates order two. The vertical dotted lines correspond to integer multiples (1, 2, 3, 4, 5) of the maximum step size for which the leapfrog scheme (2.20) is stable.

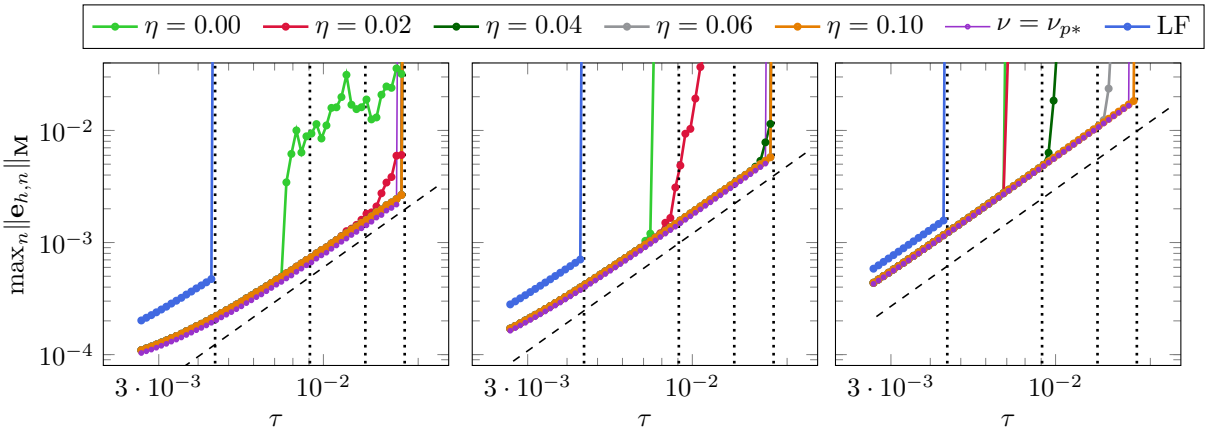
The case $\delta = 0$

In the case of $\delta = 0$ we have $g \equiv 0$ in (2.18) and, hence, also $\mathbf{g} \equiv 0$. As before, we use the leapfrog scheme (2.20) and the LFC scheme (4.1), but this time with polynomial degree $p = 4$ and $p = 5$ and six different values for the stabilization parameter ν (either via $\nu_{p,\eta}$ in (4.3) or via ν_{p*}).

In Figure 4.9a, we plot the maximum error $\|\mathbf{e}_{h,n}\|_{\mathbf{M}}$ over all time steps up to the ending time $T = 4.6$. As for the example of the modified FPUT problem with $\beta_i^* = 0$, we observe that the LFC method allows us to choose an approximately p times larger step size compared to the leapfrog method; see the dotted lines marking integer multiples of the maximum step size for which the leapfrog scheme is stable. Further, in agreement with the results in Section 4.2.2 an increase of the stabilization parameter η (or ν) slightly reduces the maximum step size, since $\widehat{\beta}_{p,\nu}$ decreases. Moreover, one can clearly see the influence of the stabilization parameter ν on the error constant. The closer $\nu_{p,\eta}$ approaches the value ν_{p*}, the better the error constant becomes. This again confirms the theoretical result in Theorem 3.54 and Remark 4.6. In contrast to the previous example the fourth-order convergence is not visible since the time discretization error is already dominated by the space discretization error.

The case $\delta > 0$

In this last example we look at the behavior of the numerical approximations in the case of an increasing value of $\delta > 0$, i.e., we have $g \not\equiv 0$. The spatially discretized problem (2.2), (4.24) then serves as a simple model showing the influence of the semilinearity to the stability of the LFC schemes. In particular, it fits into the setting of Section 3.3.3. In Figure 4.10 we plot the maximum error $\|\mathbf{e}_{h,n}\|_{\mathbf{M}}$ over all time steps up to the ending time $T = 5$ for different values of $\delta > 0$. As time integration schemes we employ besides the leapfrog scheme (2.20) and the LFC scheme (4.1) also the variant (3.7), (4.1b) (abbreviated with LFC-G). As starting value for



(a) LFC scheme, $\delta = 10$. (b) LFC, $\delta = 20$. (c) LFC, $\delta = 50$.

(d) LFC-G, $\delta = 10$. (e) LFC-G, $\delta = 20$. (f) LFC-G, $\delta = 50$.

Figure 4.10.: Errors for the numerical solution of the (spatially discretized) wave equation (2.18) with g defined in (4.24) (and exact solution (4.23)) plotted against the step size for different values of δ. For the time integration we employ the leapfrog scheme (2.20), the LFC scheme (4.1) and the variant (3.7), (4.1b) (LFC-G) with a modified starting value. We use the LFC polynomial of degree $p = 4$ with stabilization parameters $\nu = \nu_{p,\eta}$ with $\eta = 0$, $\eta = 0.02$, $\eta = 0.04$, $\eta = 0.06$, $\eta = 0.1$, and $\nu = \nu_{p*}$. The blue line represents the leapfrog scheme. The dashed black line indicates order two. The vertical dotted lines correspond to integer multiples (1, 2, 3, 4) of the maximum step size for which the leapfrog scheme (2.20) is stable.

this variant we modify (4.1c) in the same way as the two-step method. Note that this variant is analyzed in [CHS20]. For the LFC polynomial we use $p = 4$ and different values for the stabilization parameter ν.

In agreement with the results in Section 3.3.3 we observe that without (enough) stabilization the LFC scheme and also its variant cannot achieve an approximately p times larger step size than the leapfrog scheme. Moreover, we see that the larger δ is, the greater the stabilization parameter ν (or η) has to be chosen to gain an optimal step-size restriction. Nevertheless, this is already achieved for rather small values for ν and η. We can further clearly observe the second-order convergence of all schemes in accordance with our theory. In contrast to the previous example with $g \equiv 0$ the influence of the stabilization parameter ν to the error constant is almost neglectable, even for the choice $\nu = \nu_{p*}$.

By a comparison of the LFC scheme (4.1) and its variant LFC-G we see that the LFC scheme yields not only slightly smaller errors than the LFC-G scheme but also has a better stability behavior. More precisely, for the same stabilization parameter the stability behavior of the LFC scheme is at least as good as of its variant, if not better. This better behavior can be explained by the fact that due to (3.12) and (3.18) we have

$$\widehat{P}_p(z) \le \min\{1, 4(1-m_1)/z\} \le 1 \qquad \text{for all } z \in [0, \widehat{\beta}^2_{p,\nu}],$$

leading to $\widehat{P}_p(z) < 1$ for $z \ge 4$ (recall that $\lim_{\nu\to 1} \widehat{\beta}^2_{p,\nu} = 4p^2$).

Chapter 5

Multirate leapfrog-type two-step schemes

This last chapter is devoted to the construction of multirate schemes for differential equations of the form (2.1). Recall that multirate schemes are designed for situations where only a (very) small part of the differential equation is responsible of the severe stiffness of (2.1). To overcome this problem, multirate schemes either employ in the stiff part the same time integration scheme as in the non-stiff part but with smaller step size, or use even a completely different scheme in the stiff part including implicit ones. The multirate scheme we consider in this chapter are based on the leapfrog scheme (as integrator for the non-stiff part) and the general class of schemes considered in Chapter 3 (for the small stiff part). Again, we mainly focus on the LFC schemes and the modified θ-schemes.

We start with a short motivation in Section 5.1. In Section 5.2 we present general multirate schemes, where we make use of the function $\widehat{\Psi}$ introduced in Chapter 3. In addition, we also show some basic (geometric) properties of these schemes. Afterwards we analyze the stability and errors of the general schemes. In contrast to Chapter 3, we show here results only in the standard norm $\|\cdot\|$. In Section 5.5 we show that for the special case of the rational function Ψ_θ (θ-*function*) defined in (3.4) the stability bounds can be improved and a weaker step-size restriction is required. Section 5.6 contains the implementation for the special case of Ψ being an LFC polynomial (4.1b) or a θ-function (3.4). Moreover, we show the beneficial efficiency of these schemes compared to the standard leapfrog scheme for a large class of applications. We conclude this chapter by some numerical examples confirming our theoretical findings.

Most of the following results are published in [CH21]. Besides some small additional results, the improved results for the special case of θ-functions in Section 5.5 have been not published so far. We further point out that in [GMS21] the special case of the leapfrog scheme combined with the LFC schemes is analyzed for $\mathbf{g} \equiv 0$. In contrast to their work our results hold for general semilinear problems (2.2), need less regularity in time, can be applied also to semidefinite matrices $\mathbf{L}$, and require a weaker step-size restriction.

5.1. Motivation

The situation that only a small principle submatrix of $\mathbf{L}$ is responsible for the main stiffness of differential equations of the form (2.1) is a problem which often occurs in applications, especially

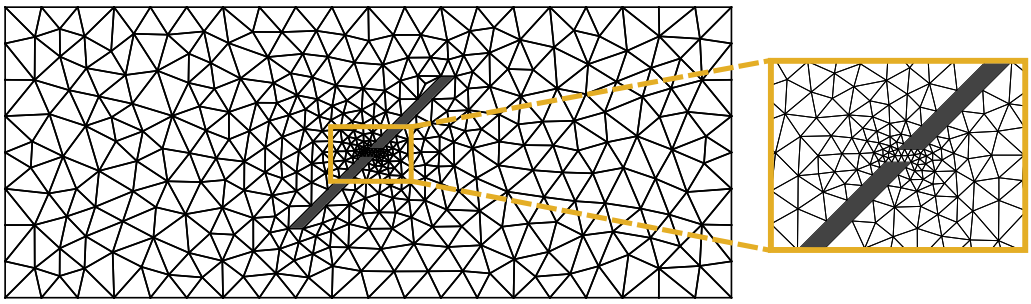


Figure 5.1.: Locally refined mesh based on a geometric constraint (picture taken from [Stu17, Figure 5.2]).

for spatially discretized wave-type equations such as the acoustic wave equation (2.18) presented in Section 2.2.2. For these the origin of such situations can be traced back to two reasons.

The first one are locally refined meshes; see Figure 5.1 for an example. Such meshes consist mostly of coarse elements but also contain a few (very) fine elements. They are required, for instance, to resolve small geometric features in the underlying domain such as narrow gaps. Using an appropriate space discretization method for such meshes, for instance, (discontinuous Galerkin) finite element methods, then leads to differential equations (2.1) (possibly after retransformation) where only a small part of $\mathbf{L}$ causes the severe stiffness.

The other case where such situation can arise are heterogeneous media with material parameters of different magnitude, for instance, if the parameter c in the acoustic wave equation (2.18) is only large in a small part of the domain compared to the remaining part of the domain. In such cases the ratio of the material parameter between both parts of the domains transfers to to the matrix $\mathbf{L}$. Hence, if the material parameter is only large in a small part of the domain, again only a small submatrix of $\mathbf{L}$ induces the main stiffness.

5.2. Construction and basic properties

As in Chapter 3, we first show the construction of general multirate schemes for the problem (2.1) with $\mathbf{M} = \mathbf{I}$ and comment afterwards on changes for general $\mathbf{M}$. Motivated by the examples presented in the last section, we assume the following structure for the symmetric and positive semidefinite matrix $\mathbf{L}$.

Assumption 5.1. *Possibly after permutation let* $\mathbf{L}$ *be partitioned as*

$$\mathbf{L} = \begin{pmatrix} \mathbf{S} & \mathbf{K}^T \\ \mathbf{K} & \mathbf{N} \end{pmatrix}, \tag{5.1}$$

where the norms of the "nonstiff" *and* "stiff" *submatrices* $\mathbf{N} \in \mathbb{R}^{(m-s)\times(m-s)}$ *and* $\mathbf{S} \in \mathbb{R}^{s\times s}$, *respectively, satisfy* $\|\mathbf{S}\| = r\|\mathbf{N}\|$ *with* $r \gg 1$ *and* $m \gg s$. *For the* coupling *matrix* $\mathbf{K} \in \mathbb{R}^{(m-s)\times s}$ *it holds* $\|\mathbf{K}\| = \kappa\|\mathbf{N}\|$ *with* $0 \leq \kappa \ll r^{1/2}$.

We point out that the exact form (5.1) of $\mathbf{L}$ is only for the sake of presentation. In applications it is sufficient to know the corresponding entries of the stiff and nonstiff (or not that stiff) part of the differential equation. With these assumptions on $\mathbf{L}$ the stiffness of (2.1) is induced by the submatrix $\mathbf{S}$ which is of much smaller size than $\mathbf{N}$. Moreover, the symmetry and positive

semidefiniteness of $\mathbf{L}$ is inherited from $\mathbf{S}$ and $\mathbf{N}$. The assumption $\kappa^2 \ll r$ originates from the fact that the positive semidefiniteness implies $\kappa^2 \leq r$ and that the coupling of $\mathbf{S}$ and $\mathbf{N}$ should not be too strong.

In order to split the differential equation (2.1) into a stiff and a nonstiff part, we further define a *restriction matrix* to the stiff part of the differential equation

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_s & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}. \tag{5.2}$$

For a clearer presentation we write here and in the following $\mathbf{I}_m$, $\mathbf{I}_s$, and $\mathbf{I}_{m-s}$ for identity matrices of size m, s, and $m-s$, respectively. With the restriction matrix we have

$$\mathbf{LR} = \begin{pmatrix} \mathbf{S} & 0 \\ \mathbf{K} & 0 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\mathbf{S}} = \mathbf{RLR} = \begin{pmatrix} \mathbf{S} & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.3}$$

Note that all eigenvalues of the matrix $\mathbf{LR}$ are real and nonnegative due to the symmetry and positive semidefiniteness of $\mathbf{L}$. More precisely, the eigenvalues of $\mathbf{LR}$ are those of $\mathbf{S}$ and (additionally) zero with multiplicity $m-s$.

5.2.1. Two-step formulation

With these definitions at hand, we are able to construct the multirate schemes. For this, we recall the general two-step schemes (3.1a) in Chapter 3. In the previous two chapters we showed for these schemes that with appropriate choices of $\widehat{\Psi}$, for instance, the LFC polynomials (4.1b), the maximum step size for which the schemes are stable can be significantly enlarged.

We now transfer this idea to the multirate case. Since we would like to retain the computational cheap leapfrog scheme for the large nonstiff part of the differential equation (2.1), we multiply the right-hand side of the leapfrog scheme (2.20) by the matrix function $\widehat{\Psi}(\tau^2\mathbf{LR})$, where $\widehat{\Psi}$ is given as in Assumption 3.2. Thus, we propose the scheme

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2 \widehat{\Psi}(\tau^2\mathbf{LR})(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \qquad n = 1, 2, \ldots, \tag{5.4a}$$

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2 \widehat{\Psi}(\tau^2\mathbf{LR})(-\mathbf{L}\mathbf{q}_0 + \mathbf{g}_0). \tag{5.4b}$$

For $\widehat{\Psi} \equiv 1$ or $\mathbf{R} = 0$ we obtain by Assumption 3.2 that the scheme reduces to the leapfrog scheme (2.20). Further, the matrix functions $\widehat{\Psi}(\tau^2\mathbf{LR})$ are well-defined for all $\tau \geq 0$ because of (5.3) and the definition of $\widehat{\Psi}$ in Assumption 3.2; cf. the Definition B.13 for matrix functions. The insertion of the matrix function $\widehat{\Psi}(\tau^2\mathbf{LR})$ is motivated by the following observation, which is essential for our analysis.

Lemma 5.2. *The matrix $\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}$ is symmetric.*

Proof. With Lemmas B.14(a), B.16, and the symmetry of $\mathbf{L}$, $\mathbf{R}$ we obtain

$$\big(\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\big)^T = \mathbf{L}^T\widehat{\Psi}(\tau^2\mathbf{LR})^T = \mathbf{L}\widehat{\Psi}(\tau^2\mathbf{RL}) = \widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L},$$

which shows the symmetry. □

In contrary, if we had inserted the symmetric matrix function $\widehat{\Psi}(\tau^2\mathbf{RLR})$ instead of $\widehat{\Psi}(\tau^2\mathbf{RL})$, the resulting matrix $\widehat{\Psi}(\tau^2\mathbf{RLR})\mathbf{L}$ would be not symmetric in general.

Similarly to the variant (3.7) of the two-step scheme (3.1a), we also could consider the following variant to (5.4)

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = -\tau^2 \widehat{\Psi}(\tau^2 \mathbf{L}\mathbf{R})\mathbf{L}\mathbf{q}_n + \tau^2 \mathbf{g}_n, \qquad n = 1, 2, \dots, \tag{5.5a}$$

$$\mathbf{q}_1 = \mathbf{q}_0 + \tau \dot{\mathbf{q}}_0 - \tfrac{1}{2}\tau^2 \widehat{\Psi}(\tau^2 \mathbf{L}\mathbf{R})\mathbf{L}\mathbf{q}_0 + \tfrac{1}{2}\tau^2 \mathbf{g}_0, \tag{5.5b}$$

where $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$ is only applied to the matrix $\mathbf{L}$ but not to $\mathbf{g}$. The analysis of this scheme can be done analogously as below with similar assumptions. However, the additional factor $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$ in front of $\mathbf{g}$ leads to an improved stability constant, especially for the special case of θ-functions (3.4) for Ψ in Section 5.5. Hence, we focus on the scheme (5.4) instead of its variant (5.5) in the following.

Remark 5.3 (General mass matrix $\mathbf{M}$). In Lemma 2.1 we have shown that the general differential equation (2.2) can be reformulated to the form (2.1) by using the transformation $\widehat{\mathbf{q}} = \mathbf{C}_\mathbf{M}^T \mathbf{q}$ where $\mathbf{C}_\mathbf{M}$ denotes the Cholesky factor of $\mathbf{M}$. Hence, we can apply the multirate scheme (5.4) to the recast equation (2.3) leading to

$$\widehat{\mathbf{q}}_{n+1} - 2\widehat{\mathbf{q}}_n + \widehat{\mathbf{q}}_{n-1} = \tau^2 \widehat{\Psi}(\tau^2 \widehat{\mathbf{L}}\mathbf{R})(-\widehat{\mathbf{L}}\widehat{\mathbf{q}}_n + \widehat{\mathbf{g}}(t_n, \widehat{\mathbf{q}}_n)), \qquad n = 1, 2, \dots,$$

and similarly for the starting value. Transforming these back to the original variables then yields the multirate scheme for the general problem (2.2)

$$\mathbf{M}(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1}) = \tau^2 \widehat{\Psi}(\tau^2 \mathbf{L}\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1})(-\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n), \qquad n = 1, 2, \dots, \tag{5.6a}$$

$$\mathbf{M}\mathbf{q}_1 = \mathbf{M}\mathbf{q}_0 + \tau \mathbf{M}\dot{\mathbf{q}}_0 + \tfrac{1}{2}\tau^2 \widehat{\Psi}(\tau^2 \mathbf{L}\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1})(-\mathbf{L}\mathbf{q}_0 + \mathbf{M}\mathbf{g}_0). \tag{5.6b}$$

Owing to the lower triangular structure of the Cholesky factor $\mathbf{C}_\mathbf{M}$ we have

$$\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{M}_S^{-1} & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{for} \qquad \mathbf{M} = \begin{pmatrix} \mathbf{M}_S & \mathbf{M}_K^T \\ \mathbf{M}_K & \mathbf{M}_N \end{pmatrix}, \tag{5.7}$$

which is particularly beneficial for the implementation; see Section 5.6.1.

As noted in Remark 2.2 and the comments above, the subsequent results also hold in this more general situation by replacing the standard norm $\|\cdot\|$ with $\|\cdot\|_\mathbf{M}^2 = (\,\cdot\,, \mathbf{M}\,\cdot\,)$. ◇

Last, we point out that for $\mathbf{g} \equiv 0$ the multirate scheme (5.4a) (without the starting value) equipped with the leapfrog-Chebyshev polynomials (4.1b) indeed coincides with the stabilized local time-stepping scheme proposed in [GMS21] and, for $\nu = 1$, with the one in [DG09]. The equivalence of these schemes for $\mathbf{g} = 0$ is shown in [GMS21]. Moreover, the multirate scheme is also closely related to the locally implicit scheme proposed and analyzed in [Ver11] and [HS16], respectively, if we use the θ-function with $\theta = \frac{1}{4}$ for Ψ.

5.2.2. One-step formulations and geometric properties

Before we start with the stability and error analysis of the general multirate scheme (5.4), we first present an equivalent one-step formulation of this scheme and the variant (5.5). Further, we investigate the symmetry and the (non-)symplecticity of these schemes.

For the derivation of the equivalent one-step formulation of (5.4) we proceed as in the derivation of the leapfrog scheme (2.21) in Section 2.3.1. Defining $\mathbf{p}_{n+1/2} = \frac{1}{\tau}(\mathbf{q}_{n+1} - \mathbf{q}_n)$ and

$\mathbf{p}_{n+1} = \frac{1}{2}(\mathbf{p}_{n+3/2} + \mathbf{p}_{n+1/2})$ for $n \geq 0$ yields the equivalent scheme

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{LR})(-\mathbf{L}\mathbf{q}_n + \mathbf{g}_n), \tag{5.8a}$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\mathbf{p}_{n+1/2}, \qquad n = 0, 1, 2, \ldots, \tag{5.8b}$$

$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{LR})(-\mathbf{L}\mathbf{q}_{n+1} + \mathbf{g}_{n+1}), \tag{5.8c}$$

if we set $\mathbf{p}_0 = \dot{\mathbf{q}}_0$. As before, $\mathbf{p}_n$ can be interpreted as an approximation to $\dot{\mathbf{q}}(t_n) = \mathbf{p}(t_n)$ in the first-order system (2.4).

With the same arguments we obtain for the variant (5.5) the equivalent one-step scheme

$$\widehat{\mathbf{p}}_{n+1/2} = \widehat{\mathbf{p}}_n - \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\mathbf{q}_n + \tfrac{1}{2}\tau\mathbf{g}_n, \tag{5.9a}$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \tau\widehat{\mathbf{p}}_{n+1/2}, \qquad n = 0, 1, 2, \ldots, \tag{5.9b}$$

$$\widehat{\mathbf{p}}_{n+1} = \widehat{\mathbf{p}}_{n+1/2} - \tfrac{1}{2}\tau\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\mathbf{q}_{n+1} + \tfrac{1}{2}\tau\mathbf{g}_{n+1}, \tag{5.9c}$$

where we again set $\widehat{\mathbf{p}}_0 = \dot{\mathbf{q}}_0$.

Next, we turn towards geometric properties of these schemes. For the two-step schemes the symmetry is a direct consequence of Definition 2.12.

Lemma 5.4. *The two-step schemes* (5.4a) *and* (5.5a) *are symmetric.*

To investigate the symplecticity of the one-step method (5.8) and its variant (5.9), the semilinear problem (2.1) has to be of Hamiltonian structure; see Definition 2.14. Thus, let $\mathbf{g}(\cdot, \mathbf{q}) = \mathbf{g}(\mathbf{q})$ and $\mathbf{g}$ additionally satisfy (2.8) in the remaining part of this section.

Lemma 5.5. *The scheme* (5.8) *is symmetric but not symplectic in general.*

Proof. If we proceed as in the proof of Lemma 3.7, we see that the scheme (5.8) is equivalent to the leapfrog scheme (2.21) applied to the modified equation

$$\ddot{\mathbf{q}} = -\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\mathbf{q} + \widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{g}(\mathbf{q}), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0,$$

Hence, it inherits the symmetry of the one-step formulation of the leapfrog method. However, there exists no function $V\colon \mathbb{R}^m \to \mathbb{R}$ such that $\nabla V(\mathbf{q}) = -\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{g}(\mathbf{q})$ because $\widehat{\Psi}(\tau^2\mathbf{LR})$ is not symmetric in general. Thus, the problem is not Hamiltonian and we cannot conclude the symplecticity. Checking the symplecticity condition (2.24) by calculating the Jacobian of the numerical flow as in the proof of Lemma 3.8 shows that the scheme (5.8) is indeed not symplectic in general. □

After this negative result for the symplecticity of the one-step formulation of the multirate scheme (5.4), we conclude this section by showing that the variant (5.9) is symplectic.

Lemma 5.6. *The scheme* (5.9) *is symmetric and symplectic.*

Proof. Proceeding again as in the proof of Lemma 3.7 shows that the scheme (5.9) is equivalent to the leapfrog scheme (2.21) applied to the modified equation

$$\ddot{\mathbf{q}} = -\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\mathbf{q} + \mathbf{g}(\mathbf{q}), \qquad \mathbf{q}(0) = \mathbf{q}_0, \quad \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0. \tag{5.10}$$

Moreover, we can rewrite (5.10) as Hamiltonian problem (2.9a) with Hamiltonian

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2}(\mathbf{p}, \mathbf{p}) + \tfrac{1}{2}(\mathbf{q}, \widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}\mathbf{q}) + U(\mathbf{q}),$$

since $\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}$ is symmetric; see Lemma 5.2. Hence, the one-step scheme (5.9) inherits the symmetry and symplecticity of the leapfrog method. □

Since both schemes are very similar, we expect that the scheme (5.8) at least approximately conserves the favorable properties of symplectic schemes. In fact, we will see in the numerical example in Section 5.7.2 that it still nearly preserves the Hamiltonian over a long time.

5.3. Stability analysis

In this section we show stability of the scheme (5.4) under the general conditions for the function Ψ defined in Section 3.1.2. To do so, we first present some preliminary considerations and investigate the matrices $\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}$ as well as $\widehat{\Psi}(\tau^2\mathbf{LR})$ more closely. For similar reasons as in Section 3.3 we state stability results only for linear problems (2.12).

In the following let Assumptions 3.2, 3.16, and 5.1 hold without mentioning it explicitly everywhere. Moreover, we abbreviate

$$\mathbf{L}_{\Psi,\tau} = \widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{L}. \tag{5.11}$$

5.3.1. Preliminary considerations

A main challenge in the subsequent stability and error analysis is that the bounds from Ψ in Section 3.1.2 cannot be directly transferred to $\mathbf{L}_{\Psi,\tau}$ since the matrices $\mathbf{L}$ and $\mathbf{L}_{\Psi,\tau}$ in general do not share the same eigenvectors anymore due to the restriction matrix $\mathbf{R}$. Moreover, the matrix $\mathbf{LR}$ (and also $\mathbf{RL}$) are non-symmetric and thus not necessarily unitarily diagonalizable. However, as shown in Lemma 5.2, the matrix $\mathbf{L}_{\Psi,\tau}$ is symmetric.

A further crucial observation towards the stability analysis is that we can rewrite the two-step scheme (5.4a) in a "leapfrog like" manner. More precisely, defining $\widetilde{\mathbf{g}}_n = \widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{g}_n$ yields with the definition of $\mathbf{L}_{\Psi,\tau}$ for (5.4a)

$$\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1} = \tau^2(-\mathbf{L}_{\Psi,\tau}\mathbf{q}_n + \widetilde{\mathbf{g}}_n), \qquad n = 1, 2, \dots,$$

which has the same structure as the two-step leapfrog scheme (2.20a). Recall that for stability of the leapfrog scheme we require besides of the symmetry of $\mathbf{L}$ that the spectrum of $\tau^2\mathbf{L}$ is a subset of $[0, 4]$; cf. Example 3.11 and Definition 3.17 together with Theorem 3.27 in Chapter 3. The symmetry of $\mathbf{L}_{\Psi,\tau}$ we have already shown. Hence, for stability we have to prove that the eigenvalues of $\tau^2\mathbf{L}_{\Psi,\tau}$ are in $[0, 4]$ (under a step-size restriction). Moreover, the subsequent stability analysis relies on the same tools as for the general two-step schemes in Chapter 3.

For a closer investigation of $\mathbf{L}_{\Psi,\tau}$ we first state some elementary results which are essential for the subsequent stability analysis. We again point out that all eigenvalues of the matrix $\mathbf{LR}$ and, thus, also $\mathbf{RL}$ as well as $\widetilde{\mathbf{S}}$ are real and nonnegative due to the symmetry and positive definiteness of $\mathbf{L}$. In particular, all occurring matrix functions to Ψ, $\widehat{\Psi}$, and Υ are well defined.

Next, we present two alternative representations for $\mathbf{L}_{\Psi,\tau}$ which turn out to be useful.

Lemma 5.7. *Let $f\colon [0,\infty) \to \mathbb{R}$ be a sufficiently smooth function. Then we have*

$$\tau^2 f(\tau^2\mathbf{LR})\,\mathbf{LR} = \tau^2\mathbf{LR}\, f(\tau^2\widetilde{\mathbf{S}})\mathbf{R}, \tag{5.12}$$

where $\widetilde{\mathbf{S}}$ is defined in (5.3).

Proof. With $\mathbf{R} = \mathbf{R}^2$ we obtain from Lemma B.16 (with $\mathbf{A} = \tau^2\mathbf{LR}$ and $\mathbf{B} = \mathbf{R}$)

$$\tau^2 f(\tau^2\mathbf{LR})\,\mathbf{LR} = \tau^2 f(\tau^2\mathbf{LR}^2)\,\mathbf{LR}^2 = \tau^2\mathbf{LR}\, f(\tau^2\mathbf{RLR})\,\mathbf{R} = \tau^2\mathbf{LR}\, f(\tau^2\widetilde{\mathbf{S}})\mathbf{R},$$

which shows the claim. □

This lemma allows us to determine a block structure of $\mathbf{L}_{\Psi,\tau}$ which is based on the partition of $\mathbf{L}$ in Assumption 5.1.

Lemma 5.8. *For $\mathbf{L}_{\Psi,\tau}$ defined in (5.11) with $\mathbf{L}$ given as in (5.1) we have*

$$\tau^2\mathbf{L}_{\Psi,\tau} = \begin{pmatrix} \Psi(\tau^2\mathbf{S}) & \tau^2\widehat{\Psi}(\tau^2\mathbf{S})\mathbf{K}^T \\ \tau^2\mathbf{K}\widehat{\Psi}(\tau^2\mathbf{S}) & \tau^2\mathbf{N} + \tau^4\mathbf{K}\Upsilon(\tau^2\mathbf{S})\mathbf{K}^T \end{pmatrix}. \tag{5.13}$$

Proof. From Definition 3.12 and Lemma B.15 we get with relation (5.12) and Lemma B.14(c)

$$\begin{aligned} \widehat{\Psi}(\tau^2\mathbf{LR}) &= \mathbf{I}_m + \tau^2\Upsilon(\tau^2\mathbf{LR})\,\mathbf{LR} = \mathbf{I}_m + \tau^2\mathbf{LR}\,\Upsilon(\tau^2\widetilde{\mathbf{S}})\mathbf{R} \\ &= \mathbf{I}_m + \tau^2\begin{pmatrix} \mathbf{S} & 0 \\ \mathbf{K} & 0 \end{pmatrix}\begin{pmatrix} \Upsilon(\tau^2\mathbf{S}) & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \widehat{\Psi}(\tau^2\mathbf{S}) & 0 \\ \tau^2\mathbf{K}\Upsilon(\tau^2\mathbf{S}) & \mathbf{I}_{m-s} \end{pmatrix}. \end{aligned} \tag{5.14}$$

Using (5.11), (3.2), and again (3.14) completes the proof. □

The lemma implies that matrix functions only have to be evaluated on the symmetric, positive semidefinite submatrix $\mathbf{S}$, which is important for an efficient implementation; see Section 5.6. Hence, for estimates of the single blocks we can simply use the scalar estimates on the eigenvalues of $\mathbf{S}$. For the LFC polynomials (4.1b) a different proof of (5.13) is given in [GMS21].

For a second alternative representation of $\mathbf{L}_{\Psi,\tau}$ we recall that the *generalized Schur complement of* $\mathbf{S}$ *in* $\mathbf{L}$ is given by $\mathbf{A}_\mathbf{S} = \mathbf{N} - \mathbf{K}\mathbf{S}^+\mathbf{K}^T$, where $\mathbf{S}^+$ denotes the *Moore-Penrose inverse* of $\mathbf{S}$. Since $\mathbf{L}$ is positive semidefinite, so is $\mathbf{A}_\mathbf{S}$; see [Alb69] or [HZ05, Theorem 1.20]. Moreover, $\mathbf{L}$ admits a block decomposition of the form

$$\mathbf{L} = \mathbf{C}_\mathbf{S}\mathbf{C}_\mathbf{S}^T, \qquad \mathbf{C}_\mathbf{S} = \begin{pmatrix} \mathbf{S}^{\frac{1}{2}} & 0 \\ \mathbf{K}\mathbf{S}^+\mathbf{S}^{\frac{1}{2}} & \mathbf{A}_\mathbf{S}^{\frac{1}{2}} \end{pmatrix}, \tag{5.15}$$

because of $\mathbf{K}\mathbf{S}^+\mathbf{S} = \mathbf{K}$; see again [Alb69] or [HZ05, Theorem 1.19].

Lemma 5.9. *We have $\mathbf{L}_{\Psi,\tau} = \mathbf{C}_\mathbf{S}\widehat{\Psi}(\tau^2\widetilde{\mathbf{S}})\mathbf{C}_\mathbf{S}^T$ with $\widetilde{\mathbf{S}}$ defined in (5.3) and $\mathbf{C}_\mathbf{S}$ in (5.15).*

Proof. Inserting the decomposition (5.15) into $\mathbf{L}_{\Psi,\tau}$ yields together with Lemma B.16

$$\mathbf{L}_{\Psi,\tau} = \widehat{\Psi}(\tau^2\mathbf{C}_\mathbf{S}\mathbf{C}^T\mathbf{R})\mathbf{C}_\mathbf{S}\mathbf{C}_\mathbf{S}^T = \mathbf{C}_\mathbf{S}\widehat{\Psi}(\tau^2\mathbf{C}_\mathbf{S}^T\mathbf{R}\mathbf{C}_\mathbf{S})\mathbf{C}_\mathbf{S}^T.$$

A simple computation shows that $\mathbf{C}_\mathbf{S}^T\mathbf{R}\mathbf{C}_\mathbf{S} = \widetilde{\mathbf{S}}$ which concludes the proof. □

5.3.2. Bounds for matrices and matrix functions

After these preliminary considerations we turn towards bounds for $\tau^2\mathbf{L}_{\Psi,\tau}$ and also $\widehat{\Psi}(\tau^2\mathbf{LR})$. Recall that for stability of the scheme (5.4) we have to show that the spectrum of $\tau^2\mathbf{L}_{\Psi,\tau}$ is a subset of $[0,4]$. To achieve this, we require step-size restriction(s).

Definition 5.10. *For a fixed $\vartheta \in (0,1]$ we define $\tau_{\mathrm{SSR}}(\vartheta) > 0$ via*

$$\tau_{\mathrm{SSR}}(\vartheta)^2 = \min\left\{\frac{\widehat{\beta}_\Psi^2}{\|\mathbf{S}\|}, \frac{4\gamma\vartheta^2}{\|\mathbf{N}\|}\right\}, \qquad \gamma = \frac{2}{1 + (1 + 4\kappa^2 m_1^{-1})^{1/2}}, \tag{5.16a}$$

and

$$\tau_{\mathrm{SSR}} = \tau_{\mathrm{SSR}}(1), \tag{5.16b}$$

where $\widehat{\beta}_\Psi$, m_1 are defined in Definition 3.9(b) and $\mathbf{S}$, $\mathbf{N}$, κ in Assumption 5.1.

From this definition we get the step-size restriction $\tau \le \tau_{\mathrm{SSR}}(\vartheta)$, $\vartheta \in (0,1]$, or equivalently

$$\tau^2\|\mathbf{S}\| \le \widehat{\beta}_\Psi^2, \tag{5.17a}$$
$$\tau^2\|\mathbf{N}\| \le 4\gamma\vartheta^2. \tag{5.17b}$$

Obviously, $\tau_{\mathrm{SSR}}(\vartheta)$ depends on the norms of the submatrices $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{K}$ (via κ) but is independent of the norm of $\mathbf{L}$. Hence, the step-size restrictions are only influenced by the submatrices and the function Ψ. We further point out that for $\kappa = 0$, which implies $\mathbf{K} = 0$, we have $\gamma = 1$. Thus, (5.17b) corresponds in this case to the (standard) step-size restriction for the leapfrog scheme applied to (2.1) with $\mathbf{S} = \mathbf{K} = 0$; cf. Lemma 2.16 and Example 3.26. For general $\kappa > 0$ we have, however, $\gamma < 1$. In particular, the step-size restriction (5.17b) becomes stronger with increasing κ. For instance, if $\kappa = 1$, we have for $m_1 \le \frac{1}{2}$ that $\gamma \le \frac{1}{2}$. Consequently, for $\kappa > 0$ (5.17b) is stronger than the (standard) step-size restriction of the leapfrog scheme applied to (2.1) with $\mathbf{S} = \mathbf{K} = 0$.

In Section 5.7 we will present a simple example which demonstrates the dependency of γ on κ. Nevertheless, in the more realistic numerical experiments the step-size restriction (5.17b) turns out to be rather pessimistic; cf. again Section 5.7 and also the discussion in Section 5.6.2.

Lemma 5.11. *Let $\vartheta \in (0,1]$ and $\tau \le \tau_{\mathrm{SSR}}(\vartheta)$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$0 \le \tau^2(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) \le 4(1 - m_1(1-\vartheta^2))\|\mathbf{q}\|^2. \tag{5.18}$$

In particular, we have $\tau^2\|\mathbf{L}_{\Psi,\tau}\| \le 4(1 - m_1(1-\vartheta^2)) \le 4$.

Because of the symmetry of $\mathbf{L}_{\Psi,\tau}$ the lemma implies that under the step-size restriction $\tau \le \tau_{\mathrm{SSR}}$ the spectrum of $\tau^2\mathbf{L}_{\Psi,\tau}$ is contained in $[0,4]$. In [GMS21] similar estimates are shown for the special case of $\kappa = 1$ and the LFC polynomials (4.1b). However, they require a stronger step-size restriction than (5.17) to show their estimates.

Proof. (i) We first show the upper bound in equation (5.18). To do so, we split

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_S \\ \mathbf{q}_N \end{pmatrix} \in \mathbb{R}^m \qquad \text{with} \quad \mathbf{q}_S \in \mathbb{R}^s,\ \mathbf{q}_N \in \mathbb{R}^{m-s}, \tag{5.19}$$

into two subvectors in accordance with Assumption 5.1. From the block formula of $\mathbf{L}_{\Psi,\tau}$ in Lemma 5.8 we then obtain

$$\begin{aligned} \tau^2(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) = {} & (\Psi(\tau^2\mathbf{S})\mathbf{q}_S, \mathbf{q}_S) + \tau^2(\widehat{\Psi}(\tau^2\mathbf{S})\mathbf{K}^T\mathbf{q}_N, \mathbf{q}_S) \\ & + \tau^2(\mathbf{K}\widehat{\Psi}(\tau^2\mathbf{S})\mathbf{q}_S, \mathbf{q}_N) + \tau^2((\mathbf{N} + \tau^2\mathbf{K}\Upsilon(\tau^2\mathbf{S})\mathbf{K}^T)\mathbf{q}_N, \mathbf{q}_N). \end{aligned} \tag{5.20}$$

We investigate the single terms on the right side separately. The first term can be estimated by

$$(\Psi(\tau^2\mathbf{S})\mathbf{q}_S, \mathbf{q}_S) \le 4(1 - m_1)\|\mathbf{q}_S\|^2,$$

since the upper bound in (3.12) holds due to the step-size restriction (5.17a) (recall that $\mathbf{S}$ is symmetric and positive semidefinite).

For the second and third term in (5.20) we exploit that $\|\widehat{\Psi}(\tau^2\mathbf{S})\| \le 1$ under the step-size restriction (5.17a) because of (3.19). This yields together with the Cauchy-Schwarz inequality, a scaled Young's inequality and Assumption 5.1

$$\tau^2(\widehat{\Psi}(\tau^2\mathbf{S})\mathbf{K}^T\mathbf{q}_N, \mathbf{q}_S) \le \tau^2\|\mathbf{K}\|\|\mathbf{q}_N\|\|\mathbf{q}_S\| \le \tfrac{1}{2}\tau^2\kappa\|\mathbf{N}\|(\gamma_*\|\mathbf{q}_N\|^2 + \gamma_*^{-1}\|\mathbf{q}_S\|^2)$$

with a parameter $\gamma_* > 0$ yet to be determined.

For the last term in (5.20) we get again with (3.19) and (5.17a)

$$\tau^2(\mathbf{N}\mathbf{q}_N, \mathbf{q}_N) + \tau^4(\Upsilon(\tau^2\mathbf{S})\mathbf{K}^T\mathbf{q}_N, \mathbf{K}^T\mathbf{q}_N) \leq \tau^2(\mathbf{N}\mathbf{q}_N, \mathbf{q}_N) \leq \tau^2\|\mathbf{N}\|\|\mathbf{q}_N\|^2.$$

Collecting and inserting these estimates into (5.20) yields

$$\tau^2(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) \leq (4(1-m_1) + \tau^2\kappa\gamma_*^{-1}\|\mathbf{N}\|)\|\mathbf{q}_S\|^2 + \tau^2(1+\kappa\gamma_*)\|\mathbf{N}\|\|\mathbf{q}_N\|^2.$$

For $\kappa = 0$ the result follows immediately with the step-size restriction (5.17b), since $\gamma = 1$ and $\max\{1-m_1, \vartheta^2\} \leq 1 - m_1 + m_1\vartheta^2$ due to $m_1 \in (0,1)$ and $\vartheta \in (0,1]$. If $\kappa > 0$, we set $\gamma_* = \kappa m_1^{-1}\gamma$ with γ defined in (5.16a). This leads with (5.17b) and $1 + \kappa^2 m_1^{-1}\gamma = \gamma^{-1}$ to

$$\begin{aligned}
\tau^2(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) &\leq 4(1 - m_1 + m_1\vartheta^2)\|\mathbf{q}_S\|^2 + (1 + \kappa^2 m_1^{-1}\gamma)4\gamma\vartheta^2\|\mathbf{q}_N\|^2 \\
&\leq 4(1 - m_1 + m_1\vartheta^2)\|\mathbf{q}_S\|^2 + 4\vartheta^2\|\mathbf{q}_N\|^2 \\
&\leq 4(1 - m_1 + m_1\vartheta^2)\|\mathbf{q}\|^2,
\end{aligned}$$

where we used in the last estimate again $\vartheta^2 \leq 1 - m_1 + m_1\vartheta^2$ and $\|\mathbf{q}_S\|^2 + \|\mathbf{q}_N\|^2 = \|\mathbf{q}\|^2$.

(ii) To show the lower bound in (5.18) we exploit the block decomposition (5.15) or, more precisely, Lemma 5.9. With this, $\|\widetilde{\mathbf{S}}\| = \|\mathbf{S}\|$, (3.16), and the step-size restriction (5.17a) we obtain

$$(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) = (\widehat{\Psi}(\tau^2\widetilde{\mathbf{S}})\mathbf{C}_\mathbf{S}^T\mathbf{q}, \mathbf{C}_\mathbf{S}^T\mathbf{q}) \geq 0 \qquad \text{for all } \mathbf{q} \in \mathbb{R}^m,$$

which finishes the proof. □

From the proof we see that the upper bound in (3.12) is essential to obtain $\tau^2\|\mathbf{L}_{\Psi,\tau}\| \leq 4$. In contrast to this, the lower bound in (5.18) would still hold if $\Psi(z) \geq 0$ for all $z \in [0, \widehat{\beta}_\Psi^2]$ instead of the lower bound in (3.12). As in Chapter 3 we need this stricter condition on the lower bound of Ψ in (3.12) to obtain a positive definite $\mathbf{L}_{\Psi,\tau}$ in case of a positive definite $\mathbf{L}$; see Lemma 5.14 below.

Remark 5.12. With Lemma 5.9, $\|\widetilde{\mathbf{S}}\| = \|\mathbf{S}\|$, the step-size restriction (5.17a), (3.19), and (5.15) we have

$$\tau^2(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) = \tau^2(\widehat{\Psi}(\tau^2\widetilde{\mathbf{S}})\mathbf{C}_\mathbf{S}^T\mathbf{q}, \mathbf{C}_\mathbf{S}^T\mathbf{q}) \leq \tau^2(\mathbf{C}_\mathbf{S}^T\mathbf{q}, \mathbf{C}_\mathbf{S}^T\mathbf{q}) = \tau^2(\mathbf{L}\mathbf{q}, \mathbf{q}).$$

Hence, if additionally to (5.17a) the standard step-size restriction $\tau^2\|\mathbf{L}\| \leq 4$ of the leapfrog scheme (2.20) – see Lemma 2.16 and Example 3.11 – holds, the spectrum of $\tau^2\mathbf{L}_{\Psi,\tau}$ is contained in $[0,4]$. Further, for $\widehat{\beta}_\Psi^2 \geq 4$ the step-size restriction (5.17a) is weaker than $\tau^2\|\mathbf{L}\| \leq 4$ because of $\|\mathbf{S}\| \leq \|\mathbf{L}\|$. Thus, if Ψ satisfies (3.18) with $\widehat{\beta}_\Psi^2 \geq 4$, the scheme (5.4) is stable for at least all step sizes for which the leapfrog scheme is. In particular, $\widehat{\beta}_\Psi^2 \geq 4$ is fulfilled for all functions Ψ of interest, e.g., LFC polynomials for $p \geq 2$ and θ-functions for $\theta \geq \frac{1}{4}$. ◇

We now turn towards properties of $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$. We show that the matrix function is nonsingular under the step-size restriction (5.17), which we need at several points in our analysis. Moreover, we show a bound for $\|\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})\|$, since we cannot directly employ (3.19) because of the non-symmetry of $\mathbf{L}\mathbf{R}$.

Lemma 5.13. *Let $\vartheta \in (0,1]$ and $\tau \leq \tau_{\mathrm{SSR}}(\vartheta)$. Then the inverse of $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$ exists and we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$\|\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})\mathbf{q}\| \leq \widehat{c}_\Psi\|\mathbf{q}\|, \qquad \widehat{c}_\Psi = 1 + 2m_3\kappa\gamma\vartheta^2. \tag{5.21}$$

Proof. From the block formula (5.14) we discern that the inverse of $\widehat{\Psi}(\tau^2\mathbf{LR})$ exists if and only if $\widehat{\Psi}(\tau^2\mathbf{S})$ is nonsingular. Using (3.16) together with the step-size restriction (5.17a) yields that all eigenvalues of $\widehat{\Psi}(\tau^2\mathbf{S})$ are positive, hence, the inverse exists.

To estimate $\|\widehat{\Psi}(\tau^2\mathbf{LR})\|$ we again use (5.14), the Cauchy-Schwarz inequality, (3.15) and (3.19) together with the step-size restriction (5.17a), and Young's inequality to obtain

$$\begin{aligned}\|\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}\|^2 &= \|\widehat{\Psi}(\tau^2\mathbf{S})\mathbf{q}_S\|^2 + \|\tau^2\mathbf{K}\Upsilon(\tau^2\mathbf{S})\mathbf{q}_S\|^2 + 2(\tau^2\mathbf{K}\Upsilon(\tau^2\mathbf{S})\mathbf{q}_S, \mathbf{q}_N) + \|\mathbf{q}_N\|^2\\ &\le \|\mathbf{q}_S\|^2 + \rho^2\|\mathbf{q}_S\|^2 + 2\rho\|\mathbf{q}_S\|\|\mathbf{q}_N\| + \|\mathbf{q}_N\|^2\\ &\le (1+\rho+\rho^2)\|\mathbf{q}_S\|^2 + (1+\rho)\|\mathbf{q}_N\|^2\\ &\le (1+\rho)^2\|\mathbf{q}\|^2\end{aligned}$$

for all $\mathbf{q} \in \mathbb{R}^m$, where we abbreviate $\rho = \frac{1}{2}m_3\tau^2\|\mathbf{K}\|$. Employing $\|\mathbf{K}\| = \kappa\|\mathbf{N}\|$ from Assumption 5.1 and the second step-size restriction (5.17b) completes the proof. □

Next, we show the positive definiteness of $\mathbf{L}_{\Psi,\tau}$ for a positive definite $\mathbf{L}$. To prove this we employ that $\widehat{\Psi}(\tau^2\mathbf{LR})$ is nonsingular.

Lemma 5.14. *Let $\tau \le \tau_{\mathrm{SSR}}$ and $\mathbf{L}$ be positive definite. Then the inverse of $\mathbf{L}_{\Psi,\tau}$ exists and we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$(\mathbf{L}_{\Psi,\tau}\mathbf{q}, \mathbf{q}) \ge (c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3)^{-1}\|\mathbf{q}\|^2. \tag{5.22}$$

Moreover, we have $\|\mathbf{L}_{\Psi,\tau}^{-1}\| \le c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3$.

As remarked after Lemma 3.23, the dependency on the step size τ is not a problem at all since one is usually only interested in step sizes $\tau < 1$. For relevant applications we have $\tau < 1$ due to the step-size restriction (5.17b) anyway.

Proof. The existence of the inverse of $\mathbf{L}_{\Psi,\tau}$ directly follows from the nonsingularity of $\mathbf{L}$ due to the positive definiteness and the nonsingularity of $\widehat{\Psi}(\tau^2\mathbf{LR})$ shown in the previous lemma.

With the definition of Υ in (3.14) and relation (5.12) we obtain

$$\begin{aligned}\mathbf{L}_{\Psi,\tau}^{-1} &= \mathbf{L}^{-1}\widehat{\Psi}(\tau^2\mathbf{LR})^{-1} = \mathbf{L}^{-1} + \mathbf{L}^{-1}\Big(\widehat{\Psi}(\tau^2\mathbf{LR})^{-1} - \mathbf{I}_m\Big)\\ &= \mathbf{L}^{-1} - \tau^2\mathbf{L}^{-1}\widehat{\Psi}(\tau^2\mathbf{LR})^{-1}\,\Upsilon(\tau^2\mathbf{LR})\,\mathbf{LR}\\ &= \mathbf{L}^{-1} - \tau^2\mathbf{R}\,\widehat{\Psi}(\tau^2\widetilde{\mathbf{S}})^{-1}\,\Upsilon(\tau^2\widetilde{\mathbf{S}})\,\mathbf{R}.\end{aligned}$$

Hence, we get with (2.13), $\|\widetilde{\mathbf{S}}\| = \|\mathbf{S}\|$, and (3.17) under the step-size restriction (5.17a)

$$\begin{aligned}(\mathbf{L}_{\Psi,\tau}^{-1}\mathbf{q}, \mathbf{q}) &= (\mathbf{L}^{-1}\mathbf{q}, \mathbf{q}) + \tau^2(-\widehat{\Psi}^{-1}(\tau^2\widetilde{\mathbf{S}})\Upsilon(\tau^2\widetilde{\mathbf{S}})\mathbf{Rq}, \mathbf{Rq})\\ &\le c_{\mathrm{inv}}^2\|\mathbf{q}\|^2 + \tau^2\widetilde{m}_3\|\mathbf{Rq}\|^2 \le (c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3)\|\mathbf{q}\|^2,\end{aligned}$$

which yields (5.22) by replacing $\mathbf{q}$ with $\mathbf{L}_{\Psi,\tau}^{1/2}\mathbf{q}$. □

5.3.3. Stability for linear schemes

With these preliminary results we are able to show stability of the scheme (5.4). To do so, we first derive a representation formula of the numerical solution of the scheme (5.4) as in Section 3.2 for the general class of two-step schemes (3.1).

Theorem 5.15. *Let $\tau \leq \tau_{\mathrm{SSR}}$. For the approximations of the scheme (5.4) we have*

$$\mathbf{q}_n = \cos(n\mathbf{\Phi}_\star)\mathbf{q}_0 + \tau\boldsymbol{\mathcal{S}}_{n,\star}\dot{\mathbf{q}}_0 + \tau^2 \sum_{\ell=0}^{n-1} \xi_{\ell,n}\, \boldsymbol{\mathcal{S}}_{n-\ell,\star}\widehat{\Psi}(\tau^2\mathbf{LR})\, \mathbf{g}_\ell, \quad \boldsymbol{\mathcal{S}}_{k,\star} = \frac{\sin(k\mathbf{\Phi}_\star)}{\sin \mathbf{\Phi}_\star}, \tag{5.23a}$$

where $\xi_{\ell,n}$, $\ell = 0, \dots, n-1$, is defined as in (3.26) and $\mathbf{\Phi}_\star \in \mathbb{R}^{m\times m}$ is a symmetric matrix with spectrum in $[0,\pi]$ which is uniquely defined by

$$\cos \mathbf{\Phi}_\star = \mathbf{I}_m - \tfrac{1}{2}\tau^2\mathbf{L}_{\Psi,\tau} \qquad \textit{and satisfies} \quad \sin \mathbf{\Phi}_\star = \tau\big(\mathbf{L}_{\Psi,\tau}(\mathbf{I}_m - \tfrac{1}{4}\tau^2\mathbf{L}_{\Psi,\tau})\big)^{1/2}. \tag{5.23b}$$

Proof. The proof follows the same lines as the proof of Theorem 3.18 by replacing $\mathbf{\Psi}$ with $\mathbf{L}_{\Psi,\tau}$ and $\widehat{\mathbf{\Psi}}$ with $\widehat{\Psi}(\tau^2\mathbf{LR})$, since $\mathbf{L}_{\Psi,\tau}$ is a symmetric matrix whose spectrum is a subset of $[0,4]$ under the step-size restrictions (5.17) through Lemma 5.11. Thus, we get for the two-step scheme (5.4a)

$$\mathbf{q}_n = \cos(n\mathbf{\Phi}_\star)\mathbf{q}_0 + \boldsymbol{\mathcal{S}}_{n,\star}(\mathbf{q}_1 - \cos\mathbf{\Phi}_\star\, \mathbf{q}_0) + \tau^2 \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\widehat{\Psi}(\tau^2\mathbf{LR})\, \mathbf{g}_\ell \tag{5.24}$$

with $\cos\mathbf{\Phi}_\star$, $\sin\mathbf{\Phi}_\star$, and $\boldsymbol{\mathcal{S}}_{k,\star}$ defined as in (5.23). Inserting the definition of the starting value (5.4b) leads with $\cos\mathbf{\Phi}_\star = \mathbf{I}_m - \frac{1}{2}\tau^2\mathbf{L}_{\Psi,\tau}$ to (5.23a). □

Next, we provide bounds for the occurring (trigonometric) matrix functions in (5.23a). As in Section 3.3 for the general two-step schemes, we have for $\tau \leq \tau_{\mathrm{SSR}}$ given in (5.16b) that $\|\cos(s\mathbf{\Phi})\mathbf{q}\| \leq \|\mathbf{q}\|$ and $\|\sin(s\mathbf{\Phi})\mathbf{q}\| \leq \|\mathbf{q}\|$ for all $s \in \mathbb{R}$ and $\mathbf{q} \in \mathbb{R}^m$.

Lemma 5.16. (a) *Let $\tau \leq \tau_{\mathrm{SSR}}$. Then we have for all $\mathbf{q} \in \mathbb{R}^m$ and $n \in \mathbb{N}$*

$$\|\boldsymbol{\mathcal{S}}_{n,\star}\mathbf{q}\| \leq n\|\mathbf{q}\|. \tag{5.25a}$$

(b) *Let $\vartheta \in (0,1)$, $\tau \leq \tau_{\mathrm{SSR}}(\vartheta)$, and $\mathbf{L}$ be positive definite. Then $\sin\mathbf{\Phi}_\star$ is nonsingular for $\tau > 0$ and we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$\tau\,\|(\sin\mathbf{\Phi}_\star)^{-1}\mathbf{q}\| \leq c^\star_{\mathrm{s},1}(\vartheta)\|\mathbf{q}\| \qquad \textit{with } c^\star_{\mathrm{s},1}(\vartheta) = \Big(\frac{c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3}{m_1(1-\vartheta^2)}\Big)^{1/2}. \tag{5.25b}$$

For $\vartheta = 1$ or $\mathbf{L}$ positive semidefinite we formally set $c^\star_{\mathrm{s},1}(\vartheta) = \infty$.

Proof. We proceed similarly to the proof of Lemma 3.23 because of the symmetry of $\mathbf{L}_{\Psi,\tau}$. In particular, it is sufficient to show the bounds for the eigenvalues $\lambda_{\Psi,\tau} \in [0,4]$ and $\phi_\star \in [0,\pi]$ of $\tau^2\mathbf{L}_{\Psi,\tau}$ and $\mathbf{\Phi}_\star$, respectively, belonging to the same eigenvector.

(a) The estimate follows as in part (a) of the proof of Lemma 3.23.

(b) Let $\vartheta \in (0,1)$. From (5.22) and (5.18) we obtain that

$$\tau^2\big(c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3\big)^{-1} \le \lambda_{\Psi,\tau} \le 4\big(1 - m_1(1-\vartheta^2)\big).$$

Hence, we get with (5.23b)

$$|\sin\phi_\star| = \Big|\lambda_{\Psi,\tau}^{1/2}\big(1 - \tfrac{1}{4}\lambda_{\Psi,\tau}\big)^{1/2}\Big| \ge \tau\big(c_{\mathrm{inv}}^2 + \tau^2\widetilde{m}_3\big)^{-1/2}\big(m_1(1-\vartheta^2)\big)^{1/2},$$

which shows the nonsingularity of $\sin\mathbf{\Phi}_\star$ for $\tau > 0$. Taking the inverse completes the proof. □

We emphasize that the constant in the bound for $\|(\sin\mathbf{\Phi}_\star)^{-1}\|$ can be improved in the same way as the bound for $\|(\sin\mathbf{\Phi})^{-1}\|$ in (3.36b) if one employs the ideas mentioned in Remark 3.24. Additionally to these bounds we also use the following bound

Lemma 5.17. *Let $\vartheta \in (0,1)$ and $\tau \le \tau_{\mathrm{SSR}}(\vartheta)$. If additionally $\mathbf{L}$ is positive definite, we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$\tau\,\|(\sin\mathbf{\Phi}_\star)^{-1}\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}\| \le \big(m_1(1-\vartheta^2)\big)^{-1/2}\|\mathbf{q}\|_{\mathbf{L}^{-1}} \le c_{\mathrm{s},2}^\star(\vartheta)\|\mathbf{q}\| \tag{5.26}$$

with $c_{\mathrm{s},2}^\star(\vartheta) = c_{\mathrm{inv}}\big(m_1(1-\vartheta^2)\big)^{-1/2}$.

For $\vartheta = 1$ or $\mathbf{L}$ positive semidefinite we formally set $c_{\mathrm{s},2}^\star(\vartheta) = \infty$.

Proof. With (5.23b) and the upper bound in (5.18) we get in the same way as in the previous proof

$$\tau\,\|(\sin\mathbf{\Phi}_\star)^{-1}\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}\| \le \big(m_1(1-\vartheta^2)\big)^{-1/2}\|\mathbf{L}_{\Psi,\tau}^{-1/2}\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}\|.$$

However, to estimate this further we cannot proceed as before by going back to the eigenvalues because $\widehat{\Psi}(\tau^2\mathbf{LR})$ and $\mathbf{L}_{\Psi,\tau}$ are not simultaneously diagonalizable in general. Nevertheless, we have with the symmetry of $\mathbf{L}_{\Psi,\tau}$, decomposition (5.15) for $\mathbf{L}$, and (B.23)

$$\begin{aligned}
\|\mathbf{L}_{\Psi,\tau}^{-1/2}\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}\|^2 &= (\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}, \mathbf{L}_{\Psi,\tau}^{-1}\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}) = (\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{q}, \mathbf{L}^{-1}\mathbf{q}) \\
&= (\mathbf{C}_\mathbf{S}^{-1}\widehat{\Psi}(\tau^2\mathbf{C}_\mathbf{S}\mathbf{C}_\mathbf{S}^T\mathbf{R})\mathbf{q}, \mathbf{C}_\mathbf{S}^{-1}\mathbf{q}) \\
&= (\widehat{\Psi}(\tau^2\mathbf{C}_\mathbf{S}^T\mathbf{R}\mathbf{C}_\mathbf{S})\mathbf{C}_\mathbf{S}^{-1}\mathbf{q}, \mathbf{C}_\mathbf{S}^{-1}\mathbf{q}) \\
&\le (\mathbf{C}_\mathbf{S}^{-1}\mathbf{q}, \mathbf{C}_\mathbf{S}^{-1}\mathbf{q}) = \|\mathbf{q}\|_{\mathbf{L}^{-1}}^2,
\end{aligned}$$

where we used $\mathbf{C}_\mathbf{S}^T\mathbf{R}\mathbf{C}_\mathbf{S} = \widetilde{\mathbf{S}}$ and (3.19) together with the step-size restriction (5.17a) for the inequality. Combining both estimates yields the first inequality in (5.26). The second one simply follows from (2.13). □

With these lemmas we are able to state a stability result in the standard norm. As mentioned at the beginning of this section, we only show a stability result for the scheme (5.4) applied to the linear problem (2.12).

Theorem 5.18. *Let Assumptions 3.2, 3.16, and 5.1 hold. Further, let $\tau \le \tau_{\mathrm{SSR}}(\vartheta)$ for a fixed $\vartheta \in (0,1]$. Then the approximation $\mathbf{q}_n$ of the scheme (5.4) applied to the linear problem (2.12) satisfy for $n = 0, 1, 2, \ldots$*

$$\|\mathbf{q}_n\| \le \|\mathbf{q}_0\| + \min\{t_n, c_{\mathrm{s},1}^\star(\vartheta)\}\|\dot{\mathbf{q}}_0\| + \min\{t_n\widehat{c}_\Psi, c_{\mathrm{s},2}^\star(\vartheta)\}\tau\sum_{\ell=0}^{n-1}\|\mathbf{g}_\ell\|. \tag{5.27}$$

If $\mathbf{L}$ is singular, we have that the minima are attained for t_n and $t_n\widehat{c}_\Psi$, since $c^\star_{\mathrm{s},1}(\vartheta) = \infty$ and $c^\star_{\mathrm{s},2}(\vartheta) = \infty$, which is in accordance with (2.15a) for the exact solution. The same holds true for $\vartheta = 1$, which yields in contrast to (2.15a) for $\mathbf{L}$ positive definite constants which are not bounded uniformly in time.

Proof. From the representation formula (5.23a) we obtain with Lemmas 5.16 and 5.17

$$\begin{aligned}\|\mathbf{q}_n\| &\le \|\cos(n\mathbf{\Phi}_\star)\|\|\mathbf{q}_0\| + \tau\|\boldsymbol{\mathcal{S}}_{n,\star}\|\|\dot{\mathbf{q}}_0\| + \tau^2\sum_{\ell=0}^{n-1}\|\boldsymbol{\mathcal{S}}_{n-\ell,\star}\widehat{\Psi}(\tau^2\mathbf{LR})\|\|\mathbf{g}_\ell\| \\ &\le \|\mathbf{q}_0\| + \min\{t_n, c^\star_{\mathrm{s},1}(\vartheta)\}\|\dot{\mathbf{q}}_0\| + \tau\sum_{\ell=0}^{n-1}\min\{\tau(n-\ell)\widehat{c}_\Psi, c^\star_{\mathrm{s},2}(\vartheta)\}\|\mathbf{g}_\ell\|.\end{aligned}$$

Employing $\tau(n-\ell) \le t_n$ completes the proof □

5.4. Error analysis

After investigating the stability of the multirate scheme (5.4) in the last section, we next provide its error analysis. More precisely, we show that the multirate scheme converges with order two in the standard norm $\|\cdot\|$ for linear (2.12) as well as for semilinear problems (2.1). To do so, we mainly proceed in the same steps as for the error analysis of the general class of two-step schemes (3.1) in Section 3.4.

We shortly recall the notation from Section 3.4, which we adopt here. The error of the scheme (5.4) at time t_n is denoted by

$$\mathbf{e}_n = \widetilde{\mathbf{q}}_n - \mathbf{q}_n, \qquad \widetilde{\mathbf{q}}_n = \mathbf{q}(t_n) \tag{5.28}$$

and bounds on derivatives of $\mathbf{q}$ by

$$B_n^{(k)} = \max_{0\le t\le t_n}\|\mathbf{q}^{(k)}(t)\|, \qquad k = 1, 2, \dots . \tag{5.29}$$

Moreover, for $\mathbf{q} \in C^k([0,T])$, $k \in \mathbb{N}$, we write

$$\boldsymbol{\delta}^{(k)}_{n,\pm} = \tau^k\int_0^1 \kappa_\pm^{(k-1)}(\sigma)\,\mathbf{q}^{(k)}(t_n \pm \tau\sigma)\,\mathrm{d}\sigma, \qquad \kappa_\pm^{(\ell)}(\sigma) = \frac{(\pm 1)^{\ell+1}(1-\sigma)^\ell}{\ell!}, \tag{5.30}$$

for the remainder terms of the $(k-1)$st-order Taylor expansion of $\widetilde{\mathbf{q}}_{n\pm1}$ at t_n, which are bounded in Lemma 3.35.

5.4.1. Representation formula for errors

We start with deriving error recursions for (5.4). From these we then deduce the representation formula.

Lemma 5.19. *Let $\mathbf{q} \in C^4([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_n$ of the two-step scheme (5.4a) satisfies for $n \ge 1$ the recursion*

$$\mathbf{e}_{n+1} - 2\mathbf{e}_n + \mathbf{e}_{n-1} = \tau^2\widehat{\Psi}(\tau^2\mathbf{LR})(-\mathbf{L}\mathbf{e}_n + \mathbf{r}_n) + \mathbf{d}_n, \tag{5.31a}$$

where

$$\mathbf{d}_n = \boldsymbol{\Delta}_n + \boldsymbol{\delta}_n^{(4)}, \qquad \boldsymbol{\Delta}_n = -\tau^4\Upsilon(\tau^2\mathbf{LR})\mathbf{LR}\ddot{\mathbf{q}}(t_n), \tag{5.31b}$$

and $\boldsymbol{\delta}_n^{(4)}$, $\mathbf{r}_n$ are defined as in Lemma 3.36.

Proof. As in the proof of Lemma 3.36 we insert the exact solution $\widetilde{\mathbf{q}}_n$ into the scheme (5.4a) to obtain

$$\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1} = \tau^2 \widehat{\Psi}(\tau^2 \mathbf{LR})\big(-\mathbf{L}\widetilde{\mathbf{q}}_n + \mathbf{g}(t_n, \widetilde{\mathbf{q}}_n)\big) + \mathbf{d}_n \tag{5.32}$$

with a defect $\mathbf{d}_n$. This shows (5.31a) by subtracting the two-step scheme (5.4a) from (5.32). For the defect we have

$$\mathbf{d}_n = \tau^2 \ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_n^{(4)} - \tau^2 \widehat{\Psi}(\tau^2 \mathbf{LR}) \ddot{\mathbf{q}}(t_n) = \tau^2 (\mathbf{I}_m - \widehat{\Psi}(\tau^2 \mathbf{LR})) \ddot{\mathbf{q}}(t_n) + \boldsymbol{\delta}_n^{(4)},$$

where we employed (3.55) and the differential equation (2.1). Using the definition (3.14) of Υ completes the proof. □

We point out that we are able to write the defect as

$$\mathbf{d}_n = \boldsymbol{\Delta}_{n,*} + \widehat{\Psi}(\tau^2 \mathbf{LR}) \boldsymbol{\delta}_n^{(4)}, \qquad \boldsymbol{\Delta}_{n,*} = -\tau^2 \Upsilon(\tau^2 \mathbf{LR}) \mathbf{LR} (\widetilde{\mathbf{q}}_{n+1} - 2\widetilde{\mathbf{q}}_n + \widetilde{\mathbf{q}}_{n-1}), \tag{5.33}$$

similarly to the defect in Section 3.4; see Lemma 3.36. The subsequent error analysis could also be done with this representation of the defect leading to the same convergence order under the same assumptions.

In a next step towards a representation formula for $\mathbf{e}_n$ we investigate the error of the starting value (5.4b).

Lemma 5.20. *Let $\mathbf{q} \in C^3([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_1$ of the starting value (5.4b) satisfies*

$$\mathbf{e}_1 = \tfrac{1}{2} \boldsymbol{\Delta}_0 + \boldsymbol{\delta}_{0,+}^{(3)}, \tag{5.34}$$

with $\boldsymbol{\Delta}_0$ given in (5.31b) and $\boldsymbol{\delta}_{0,+}^{(3)}$ in (5.30).

Proof. We proceed as before. Inserting the exact solution into (5.4b) yields

$$\widetilde{\mathbf{q}}_1 = \widetilde{\mathbf{q}}_0 + \tau \dot{\mathbf{q}}(0) + \tfrac{1}{2} \tau^2 \widehat{\Psi}(\tau^2 \mathbf{LR})\big(-\mathbf{L}\widetilde{\mathbf{q}}_0 + \mathbf{g}(0, \widetilde{\mathbf{q}}_0)\big) + \mathbf{d}_0$$

with a defect $\mathbf{d}_0$. Since the initial values coincide with the exact values of the solution, we have $\mathbf{e}_1 = \mathbf{d}_0$. Further, (2.1) and Taylor expansion of $\widetilde{\mathbf{q}}_1$ leads to

$$\mathbf{d}_0 = \tfrac{1}{2} \tau^2 \ddot{\mathbf{q}}(0) + \boldsymbol{\delta}_{0,+}^{(3)} - \tfrac{1}{2} \tau^2 \widehat{\Psi}(\tau^2 \mathbf{LR}) \ddot{\mathbf{q}}(0) = \tfrac{1}{2} \tau^2 (\mathbf{I}_m - \widehat{\Psi}(\tau^2 \mathbf{LR})) \ddot{\mathbf{q}}(0) + \boldsymbol{\delta}_{0,+}^{(3)},$$

which shows (5.34) by using again (3.14). □

With these two lemmas we are able to derive a representation formula for the error $\mathbf{e}_n$.

Lemma 5.21. *Let $\tau \leq \tau_{\mathrm{SSR}}$ and let $\mathbf{q} \in C^4([0,T])$ be the exact solution of (2.1). The error $\mathbf{e}_n$, $n \in \mathbb{N}_0$, of the scheme (5.4) satisfies*

$$\mathbf{e}_n = \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star} \Big(\widehat{\Psi}(\tau^2 \mathbf{LR}) \mathbf{r}_\ell + \boldsymbol{\delta}_\ell^{(4)} \Big) + \boldsymbol{\mathcal{S}}_{n,\star} \boldsymbol{\delta}_{0,+}^{(3)} + \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star} \boldsymbol{\Delta}_\ell + \tfrac{1}{2} \boldsymbol{\mathcal{S}}_{n,\star} \boldsymbol{\Delta}_0 \tag{5.35}$$

with $\boldsymbol{\Delta}_\ell$ given in (5.31b).

Proof. As in equation (5.24) in the proof of Theorem 5.15 for the two-step scheme (5.4a) one has that the recursion (5.31a) satisfies

$$\mathbf{e}_n = \cos(n\boldsymbol{\Phi}_\star)\mathbf{e}_0 + \boldsymbol{\mathcal{S}}_{n,\star}(\mathbf{e}_1 - \cos \boldsymbol{\Phi}_\star \, \mathbf{e}_0) + \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star} \Big(\widehat{\Psi}(\tau^2 \mathbf{LR}) \, \mathbf{r}_\ell + \mathbf{d}_\ell \Big)$$

for $n \geq 0$. Employing $\mathbf{e}_0 = 0$ together with the defect (5.34) for $\mathbf{e}_1$ and the definition of $\mathbf{d}_n$ in (5.31b) completes the proof. □

5.4.2. Error bounds

In the following we first state the error bound for linear problems (2.12). Afterwards we turn towards error bounds for semilinear problems (2.1).

Before we state these bounds, there is one detail we have to be conscious about. For this, we observe that from the representation formula (5.35) we get for $\tau \leq \tau_{\mathrm{SSR}}$ with (5.25a)

$$\|\mathbf{e}_n\| \leq \sum_{\ell=1}^{n-1}(n-\ell)\|\boldsymbol{\delta}_\ell^{(4)}\| + n\|\boldsymbol{\delta}_{0,+}^{(3)}\| + \sum_{\ell=1}^{n-1}(n-\ell)\|\boldsymbol{\Delta}_\ell\| + \tfrac{1}{2}n\|\boldsymbol{\Delta}_0\|,$$

where we neglected the terms $\mathbf{r}_\ell$ for simplicity. Since we have $\|\boldsymbol{\delta}_\ell^{(4)}\| \leq C\tau^4$ and $\|\boldsymbol{\delta}_0^{(3)}\| \leq C\tau^3$ because of Lemma 3.35, the scheme would be of second order if $\|\boldsymbol{\Delta}_\ell\| \leq C\tau^4$ for $\boldsymbol{\Delta}_\ell$ defined in (5.31b). And in fact, having a closer look at $\boldsymbol{\Delta}_n$ leads to $\|\boldsymbol{\Delta}_n\| \leq \tau^4 c_\Upsilon \|\mathbf{LR}\,\ddot{\mathbf{q}}(t_n)\|$, since similarly to the proof of Lemma 5.13 one can show $\|\Upsilon(\tau^2\mathbf{LR})\| \leq c_\Upsilon$. However, if $\mathbf{L}$ is a discretized differential operator such as the Laplacian in Section 2.2.2, we want the bounds to depend only on derivatives of $\mathbf{q}$ or $\mathbf{Lq}$ to avoid a loss of consistency; cf. , e.g., [HS16, Lemma 2.8], where the problematic nature of the term $\|\mathbf{LR}\ddot{\mathbf{q}}(t_n)\|$ is shown for a discretized first-order differential operator.

A remedy to this problem is given in the subsequent lemma. The main idea consists in combining the defects $\boldsymbol{\Delta}_\ell$ of three successive time steps. We point out that a similar trick is used in the error analysis of one-step methods for (spatially discretized) partial differential equations; see, e.g., [BCT82] or [HV03, Lemma II.2.3]. In the context of locally implicit schemes this approach was also employed in [HS16, Ver11] for Maxwell's equation.

Lemma 5.22. *Let $\tau \leq \tau_{\mathrm{SSR}}$. Then we have for $\boldsymbol{\Delta}_\ell$ given in (5.31b)*

$$\begin{aligned}\tfrac{1}{2}\boldsymbol{\mathcal{S}}_{n,\star}\boldsymbol{\Delta}_0 + \sum_{\ell=1}^{n-1}\boldsymbol{\mathcal{S}}_{n-\ell,\star}\boldsymbol{\Delta}_\ell = \sum_{\ell=1}^{n-1}\boldsymbol{\mathcal{S}}_{n-\ell,\star}(\tilde{\boldsymbol{\Delta}}_{\ell+1} - 2\tilde{\boldsymbol{\Delta}}_\ell + \tilde{\boldsymbol{\Delta}}_{\ell-1}) \\ - \tilde{\boldsymbol{\Delta}}_n + \boldsymbol{\mathcal{S}}_{n,\star}(\tilde{\boldsymbol{\Delta}}_1 - \tilde{\boldsymbol{\Delta}}_0) + \cos(n\boldsymbol{\Phi}_\star)\tilde{\boldsymbol{\Delta}}_0\end{aligned} \tag{5.36}$$

with

$$\tilde{\boldsymbol{\Delta}}_n = \tau^2\mathbf{R}\,\widehat{\Psi}(\tau^2\widetilde{\mathbf{S}})^{-1}\Upsilon(\tau^2\widetilde{\mathbf{S}})\mathbf{R}\,\ddot{\mathbf{q}}(t_n). \tag{5.37}$$

Proof. Let $\ell \in \{0, 1, \ldots, n-1\}$. Under the step-size restriction (5.17a) we have for $\boldsymbol{\Delta}_\ell$ given in (5.31b) with Lemma 5.13, (5.12), and definition (5.11) for $\mathbf{L}_{\Psi,\tau}$

$$\begin{aligned}\boldsymbol{\Delta}_\ell &= -\tau^4\,\widehat{\Psi}(\tau^2\mathbf{LR})\,\widehat{\Psi}(\tau^2\mathbf{LR})^{-1}\Upsilon(\tau^2\mathbf{LR})\,\mathbf{LR}\,\ddot{\mathbf{q}}(t_\ell) \\ &= -\tau^4\,\widehat{\Psi}(\tau^2\mathbf{LR})\mathbf{LR}\,\widehat{\Psi}^{-1}(\tau^2\widetilde{\mathbf{S}})\Upsilon(\tau^2\widetilde{\mathbf{S}})\mathbf{R}\,\ddot{\mathbf{q}}(t_\ell) \\ &= -\tau^2\mathbf{L}_{\Psi,\tau}\tilde{\boldsymbol{\Delta}}_\ell.\end{aligned}$$

This yields for the term containing $\boldsymbol{\Delta}_0$ with (5.23b)

$$\tfrac{1}{2}\boldsymbol{\mathcal{S}}_{n,\star}\boldsymbol{\Delta}_0 = \boldsymbol{\mathcal{S}}_{n,\star}\tfrac{1}{2}\tau^2\mathbf{L}_{\Psi,\tau}\tilde{\boldsymbol{\Delta}}_0 = \boldsymbol{\mathcal{S}}_{n,\star}\cos(\boldsymbol{\Phi}_\star)\tilde{\boldsymbol{\Delta}}_0 - \boldsymbol{\mathcal{S}}_{n,\star}\tilde{\boldsymbol{\Delta}}_0. \tag{5.38}$$

Further, we obtain together with (5.23b) and the trigonometric identity (B.5a)

$$\boldsymbol{\mathcal{S}}_{n-\ell,\star}\boldsymbol{\Delta}_\ell = \boldsymbol{\mathcal{S}}_{n-\ell,\star}\,2\,(\cos(\boldsymbol{\Phi}_\star) - \mathbf{I}_m)\,\tilde{\boldsymbol{\Delta}}_\ell = (\boldsymbol{\mathcal{S}}_{n-\ell+1,\star} - 2\boldsymbol{\mathcal{S}}_{n-\ell,\star} + \boldsymbol{\mathcal{S}}_{n-\ell-1,\star})\tilde{\boldsymbol{\Delta}}_\ell,$$

which implies with $\boldsymbol{\mathcal{S}}_{0,\star} = 0$ and $\boldsymbol{\mathcal{S}}_{1,\star} = \mathbf{I}_m$

$$\begin{aligned}\sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\boldsymbol{\Delta}_\ell &= \sum_{\ell=0}^{n-2} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\tilde{\boldsymbol{\Delta}}_{\ell+1} - 2\sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\tilde{\boldsymbol{\Delta}}_\ell + \sum_{\ell=2}^{n} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\tilde{\boldsymbol{\Delta}}_{\ell-1} \\ &= \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}(\tilde{\boldsymbol{\Delta}}_{\ell+1} - 2\tilde{\boldsymbol{\Delta}}_\ell + \tilde{\boldsymbol{\Delta}}_{\ell-1}) - \tilde{\boldsymbol{\Delta}}_n + \boldsymbol{\mathcal{S}}_{n,\star}\tilde{\boldsymbol{\Delta}}_1 - \boldsymbol{\mathcal{S}}_{n-1,\star}\tilde{\boldsymbol{\Delta}}_0.\end{aligned}$$

Combining this equation with (5.38) and applying the angle addition formula for sine (B.1a) to $\boldsymbol{\mathcal{S}}_{n-1,\star}$ then yields (5.36). □

With this lemma we are now in the position to prove our error results. As for the error analysis of the general two-step scheme (3.1) in Section 3.4 we start with the linear case, i.e., we consider the application of the scheme (5.4) to the linear differential equation (2.12).

Theorem 5.23. *Let Assumptions 3.2, 3.16, and 5.1 hold and let $\vartheta \in (0,1]$. Further, assume that the solution $\mathbf{q}$ of (2.12) satisfies $\mathbf{q} \in C^4([0,T])$. Then, for $\tau \le \tau_{\mathrm{SSR}}(\vartheta)$ and $t_n \le T$ we have for the approximations $\mathbf{q}_n$ of the scheme (5.4)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \le (\min\{T, c^\star_{\mathrm{s},1}(\vartheta)\}(C_{\blacklozenge}T + C_{\bigstar}) + C_{\blacktriangle})\tau^2 \tag{5.39}$$

with

$$C_{\blacklozenge} = (\tfrac{1}{12} + \widetilde{m}_3)B_n^{(4)}, \quad C_{\bigstar} = (\tfrac{1}{6} + \widetilde{m}_3)B_1^{(3)}, \quad C_{\blacktriangle} = \widetilde{m}_3(\|\ddot{\mathbf{q}}(0)\| + \|\ddot{\mathbf{q}}(t_n)\|),$$

where $B_\tau^{(3)}$, $B_T^{(4)}$ are given by (5.29) and $\widetilde{m}_3$ in (3.17).

We emphasize that the error bound only depends on the function Ψ and the exact solution $\mathbf{q}$ and its derivatives but is independent of $\mathbf{L}$. Further, for $\vartheta = 1$ or $\mathbf{L}$ singular, the error bound grows quadratically in time, since $c^\star_{\mathrm{s},1}(\vartheta) = \infty$ in this case. Contrarily, we have for positive definite $\mathbf{L}$ and $\vartheta \in (0,1)$ that the error bound grows only linearly in time.

Proof. We first employ Lemma 5.22 to rewrite the representation formula (5.35) as

$$\begin{aligned}\mathbf{e}_n = &\sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}\Big(\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})\,\mathbf{r}_\ell + \boldsymbol{\delta}_\ell^{(4)}\Big) + \boldsymbol{\mathcal{S}}_{n,\star}\boldsymbol{\delta}_{0,+}^{(3)} \\ &+ \sum_{\ell=1}^{n-1} \boldsymbol{\mathcal{S}}_{n-\ell,\star}(\tilde{\boldsymbol{\Delta}}_{\ell+1} - 2\tilde{\boldsymbol{\Delta}}_\ell + \tilde{\boldsymbol{\Delta}}_{\ell-1}) - \tilde{\boldsymbol{\Delta}}_n + \boldsymbol{\mathcal{S}}_{n,\star}(\tilde{\boldsymbol{\Delta}}_1 - \tilde{\boldsymbol{\Delta}}_0) + \cos(n\boldsymbol{\Phi}_\star)\tilde{\boldsymbol{\Delta}}_0\,.\end{aligned} \tag{5.40}$$

Since $\mathbf{r}_\ell = 0$ by (3.53c) and the assumption on $\mathbf{g}$, we have with Lemma 5.16

$$\begin{aligned}\|\mathbf{e}_n\| \le{}& \min\{t_n, c^\star_{\mathrm{s},1}(\vartheta)\}\Big(\sum_{\ell=1}^{n-1} \tfrac{1}{\tau}\|\boldsymbol{\delta}_\ell^{(4)}\| + \tfrac{1}{\tau}\|\boldsymbol{\delta}_{0,+}^{(3)}\|\Big) \\ &+ \min\{t_n, c^\star_{\mathrm{s},1}(\vartheta)\}\Big(\sum_{\ell=1}^{n-1} \tfrac{1}{\tau}\|\tilde{\boldsymbol{\Delta}}_{\ell+1} - 2\tilde{\boldsymbol{\Delta}}_\ell + \tilde{\boldsymbol{\Delta}}_{\ell-1}\| + \tfrac{1}{\tau}\|\tilde{\boldsymbol{\Delta}}_1 - \tilde{\boldsymbol{\Delta}}_0\|\Big) + \|\tilde{\boldsymbol{\Delta}}_0\| + \|\tilde{\boldsymbol{\Delta}}_n\|.\end{aligned}$$

Hence, it remains to bound the defects.

(i) The defects $\tilde{\boldsymbol{\Delta}}_0$ and $\tilde{\boldsymbol{\Delta}}_n$ defined in (5.37) are bounded by

$$\|\tilde{\boldsymbol{\Delta}}_0\| + \|\tilde{\boldsymbol{\Delta}}_n\| \le \tau^2\widetilde{m}_3(\|\ddot{\mathbf{q}}(0)\| + \|\ddot{\mathbf{q}}(t_n)\|).$$

because of (3.17) and the step-size restriction (5.17a).

(ii) For the central difference quotient of $\tilde{\mathbf{\Delta}}_\ell$ we again employ (3.17) and (5.17a) together with Taylor expansion to obtain

$$\begin{aligned}\sum_{\ell=1}^{n-1} \tfrac{1}{\tau}\|\tilde{\mathbf{\Delta}}_{\ell+1} - 2\tilde{\mathbf{\Delta}}_\ell + \tilde{\mathbf{\Delta}}_{\ell-1}\| &\leq \sum_{\ell=1}^{n-1} \tau\widetilde{m}_3\|\ddot{\mathbf{q}}(t_{\ell+1}) - 2\ddot{\mathbf{q}}(t_\ell) + \ddot{\mathbf{q}}(t_{\ell-1})\| \\ &\leq \sum_{\ell=1}^{n-1} \tau^3\widetilde{m}_3 \max_{t_{\ell-1}\leq t\leq t_{\ell+1}} \|\mathbf{q}^{(4)}(t)\| \\ &\leq \tau^2\widetilde{m}_3\, t_n B_n^{(4)}.\end{aligned}$$

(iii) With the same arguments as in the previous part we get

$$\tfrac{1}{\tau}\|\tilde{\mathbf{\Delta}}_1 - \tilde{\mathbf{\Delta}}_0\| \leq \tau\widetilde{m}_3\|\tilde{\mathbf{\Delta}}_1 - \tilde{\mathbf{\Delta}}_0\| \leq \tau^2\widetilde{m}_3 B_1^{(3)}.$$

(iv) Finally, for the remainder terms $\boldsymbol{\delta}_\ell^{(4)}$ and $\boldsymbol{\delta}_{0,+}^{(3)}$ of the Taylor expansion we have with (3.52) as in Section 3.4

$$\sum_{\ell=1}^{n-1} \tfrac{1}{\tau}\|\boldsymbol{\delta}_\ell^{(4)}\| + \tfrac{1}{\tau}\|\boldsymbol{\delta}_{0,+}^{(3)}\| \leq \tau^2 t_n \tfrac{1}{12} B_n^{(4)} + \tau^2 \tfrac{1}{6} B_1^{(3)}.$$

Collecting and inserting the estimates into the above estimate completes the proof. □

As in Section 3.4 we are able to transfer the result to the semilinear case. We again distinguish between the two Lipschitz conditions (2.5) and (2.19) given in Assumptions 2.3 and 2.9, respectively.

Theorem 5.24. *Let $\vartheta \in (0,1]$ and let Assumptions 3.2, 3.16, 5.1, as well as Assumption 2.3 on $\mathbf{g}$ hold. Further, assume that for $T \in (0, t_*)$ the solution $\mathbf{q}$ of (2.1) satisfies $\mathbf{q} \in C^4([0,T])$. Then, there exists a $\tau_* > 0$ such that for $\tau \leq \min\{\tau_*, \tau_{\mathrm{SSR}}(\vartheta)\}$ and $t_n \leq T$ we have for the approximations $\mathbf{q}_n$ of the scheme (5.4)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \leq \big(\min\{T, c_{\mathrm{s},1}^\star(\vartheta)\}(C_{\blacklozenge} T + C_{\bigstar}) + C_{\blacktriangle}\big) \mathrm{e}^{(\mathcal{L}_{\mathbf{g}}\widehat{c}_\Psi)^{1/2} T} \tau^2, \tag{5.41}$$

with $C_{\blacklozenge}$, $C_{\bigstar}$, and $C_{\blacktriangle}$ defined as in Theorem 5.23.

Proof. As the proof closely follows the lines of the proof of Theorem 3.44, we only show the estimate for the error. From (5.40) we obtain with Lemmas 5.13 and 5.16 and the Lipschitz condition (2.5) of $\mathbf{g}$

$$\|\mathbf{e}_n\| \leq \mathcal{L}_{\mathbf{g}}\tau^2 \sum_{\ell=1}^{n-1} (n-\ell)\widehat{c}_\Psi \|\mathbf{e}_\ell\| + \big(\min\{T, c_{\mathrm{s},1}^\star(\vartheta)\}(C_{\blacklozenge} T + C_{\bigstar}) + C_{\blacktriangle}\big)\tau^2,$$

where the defects are bounded as in the previous proof. Thus, an application of the Gronwall Lemma B.19 completes the proof. □

For the weaker assumption on the local Lipschitz continuity (Assumption 2.9) we require that $\mathbf{L}$ is positive definite; see Section 2.2.2. Moreover, we need $\vartheta < 1$ to prove our error result.

Theorem 5.25. *Let $\vartheta \in (0,1)$, $\mathbf{L}$ be positive definite, and let Assumptions 3.2, 3.16, 5.1, as well as Assumption 2.9 on $\mathbf{g}$ hold. Further, assume that for $T \in (0, t_*)$ the solution $\mathbf{q}$ of (2.1) satisfies $\mathbf{q} \in C^4([0,T])$. Then, there exists a $\tau_* > 0$ such that for $\tau \le \min\{\tau_*, \tau_{\mathrm{SSR}}(\vartheta)\}$ and $t_n \le T$ we have for the approximations $\mathbf{q}_n$ of the scheme (5.4)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \le \big(\min\{T, c^\star_{\mathrm{s},1}(\vartheta)\}(C_{\blacklozenge} T + C_{\bigstar}) + C_{\blacktriangle}\big) \mathrm{e}^{(m_1(1-\vartheta^2))^{-1/2} \widehat{\mathcal{L}}_{\mathbf{g}} T} \tau^2 \tag{5.42}$$

with $C_{\blacklozenge}$, $C_{\bigstar}$, and $C_{\blacktriangle}$ defined as in Theorem 5.23.

Proof. As before, we only show the error bound. Again from (5.40) we have with Lemmas 5.16 and 5.17 and the Lipschitz condition (2.19) of $\mathbf{g}$

$$\|\mathbf{e}_n\| \le \widehat{\mathcal{L}}_{\mathbf{g}} (m_1(1-\vartheta^2))^{-1/2} \tau^2 \sum_{\ell=1}^{n-1} \|\mathbf{e}_\ell\| + \big(c^\star_{\mathrm{s},1}(\vartheta)(C_{\blacklozenge} T + C_{\bigstar}) + C_{\blacktriangle}\big)\tau^2.$$

Together with the Gronwall Lemma B.18 this yields the error bound (5.42). □

5.5. Improved results for θ-functions

In this section we show that the stability constants and also the error bounds obtained so far are suboptimal for θ-functions (3.4). A similar observation we have already made in Section 3.6 where it was shown that the constants for the general two-step scheme (3.1) can be improved if equipped with the θ-functions. Moreover, we see that for this functions we can allow for a weaker step-size restriction than (5.17) to achieve stability.

As in Section 3.6 we restrict ourselves to the case $\theta \ge \frac{1}{4}$ which refers to unconditional stable schemes due to $\beta_\Psi = \beta_\theta = \infty$; see (3.92). Nevertheless, the following results can easily be extended to $\theta \in [0, \frac{1}{4})$.

We start with an assumption on the step size, which is sufficient to show stability of the multirate scheme (5.4a) if combined with the θ-functions (3.4).

Assumption 5.26. *Let $\theta \ge \frac{1}{4}$. The step size $\tau > 0$ satisfies*

$$\tau^2 \|\mathbf{N}\| \le 4\vartheta^2 \tag{5.43}$$

for a fixed but arbitrary $\vartheta \in (0,1]$.

In comparison to the step-size restrictions (5.17), it is obvious that there is no step-size restriction for the matrix $\mathbf{S}$, since this part is treated implicitly. More astonishing is the fact that we do not have the factor γ in the step-size restriction for $\mathbf{N}$ compared to (5.17b). Thus, the strength of the coupling between $\mathbf{S}$ and $\mathbf{N}$ does not enter the stability analysis.

We first show that under this weaker step-size restriction the largest eigenvalue of $\tau^2 \mathbf{L}_{\Psi_\theta,\tau}$ is indeed bounded by 4; cf. Lemma 5.11 for the general case.

Lemma 5.27. *Let Assumption 5.26 on the step size τ hold. Then we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$0 \le \tau^2 (\mathbf{L}_{\Psi_\theta,\tau} \mathbf{q}, \mathbf{q}) \le \max\{\tfrac{1}{\theta}, 4\vartheta^2\} \|\mathbf{q}\|^2. \tag{5.44}$$

In particular, we have $\tau^2 \|\mathbf{L}_{\Psi_\theta,\tau}\| \le \max\{\frac{1}{\theta}, 4\vartheta^2\} \le 4$.

Proof. The lower bound follows as in the proof of Lemma 5.11, because we only require that $\widehat{\Psi}_\theta(z) \geq 0$ for all $z \geq 0$.

For the upper bound a crucial observation is that the functions Ψ_θ, $\widehat{\Psi}_\theta$, and Υ satisfy the relations

$$\Upsilon(z) = \widehat{\Psi}_\theta(z)\frac{1-\widehat{\Psi}_\theta(z)^{-1}}{z} = -\theta\widehat{\Psi}_\theta(z), \qquad \Psi_\theta(z) - \frac{1}{\theta} = \frac{\theta z - (1+\theta z)}{\theta(1+\theta z)} = -\frac{1}{\theta}\widehat{\Psi}_\theta(z), \quad (5.45)$$

which are implied by the corresponding definitions (3.4), (3.2), and (3.14). Using these relations we are able to rewrite $\mathbf{L}_{\Psi_\theta,\tau}$ by using the block formula (5.13) as

$$\begin{aligned}\tau^2\mathbf{L}_{\Psi_\theta,\tau} &= \begin{pmatrix} \frac{1}{\theta}\mathbf{I}_s & 0 \\ 0 & \tau^2\mathbf{N}\end{pmatrix} + \begin{pmatrix} \Psi_\theta(\tau^2\mathbf{S}) - \frac{1}{\theta}\mathbf{I}_s & \tau^2\widehat{\Psi}_\theta(\tau^2\mathbf{S})\mathbf{K}^T \\ \tau^2\mathbf{K}\widehat{\Psi}_\theta(\tau^2\mathbf{S}) & \tau^4\mathbf{K}\Upsilon(\tau^2\mathbf{S})\mathbf{K}^T\end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\theta}\mathbf{I}_s & 0 \\ 0 & \tau^2\mathbf{N}\end{pmatrix} - \begin{pmatrix} \frac{1}{\theta}\widehat{\Psi}_\theta(\tau^2\mathbf{S}) & -\tau^2\widehat{\Psi}_\theta(\tau^2\mathbf{S})\mathbf{K}^T \\ -\tau^2\mathbf{K}\widehat{\Psi}_\theta(\tau^2\mathbf{S}) & \tau^4\theta\mathbf{K}\widehat{\Psi}_\theta(\tau^2\mathbf{S})\mathbf{K}^T\end{pmatrix} = \mathbf{B}_1 - \mathbf{B}_2.\end{aligned}$$

Obviously, $\mathbf{B}_1$ and $\mathbf{B}_2$ are symmetric. Moreover, having a closer look at $\mathbf{B}_2$ reveals that it admits a block LDL decomposition

$$\mathbf{B}_2 = \begin{pmatrix} \mathbf{I}_s & 0 \\ -\tau^2\theta\mathbf{K} & \mathbf{I}_{m-s}\end{pmatrix}\begin{pmatrix} \frac{1}{\theta}\widehat{\Psi}_\theta(\tau^2\mathbf{S}) & 0 \\ 0 & 0\end{pmatrix}\begin{pmatrix} \mathbf{I}_s & -\tau^2\theta\mathbf{K}^T \\ 0 & \mathbf{I}_{m-s}\end{pmatrix}.$$

Hence, since $\widehat{\Psi}_\theta(z) > 0$ for all $z \geq 0$, we have with Sylvester's law of inertia that $\mathbf{B}_2$ is positive semidefinite. Adopting the notation (5.19) for a vector $\mathbf{q} \in \mathbb{R}^m$ then leads to

$$\tau^2(\mathbf{L}_{\Psi_\theta,\tau}\mathbf{q},\mathbf{q}) = (\mathbf{B}_1\mathbf{q},\mathbf{q}) - (\mathbf{B}_2\mathbf{q},\mathbf{q}) \leq \tfrac{1}{\theta}\|\mathbf{q}_S\|^2 + \tau^2\|\mathbf{N}\|\|\mathbf{q}_N\|^2 \leq \max\{\tfrac{1}{\theta},\tau^2\|\mathbf{N}\|\}\|\mathbf{q}\|^2,$$

where in the last step we used that $\|\mathbf{q}_S\|^2 + \|\mathbf{q}_N\|^2 = \|\mathbf{q}\|^2$. The step-size restriction (5.43) completes the proof. □

Clearly, for general functions Ψ the proof cannot be applied, since the relations (5.45) do not hold in general. Further, an estimate of the smallest eigenvalue of $\tau^2\mathbf{L}_{\Psi_\theta,\tau}$ in the case of $\mathbf{L}$ positive definite can be done as in Lemma 5.14 due to Lemma 3.62 and $\widehat{\Psi}_\theta(z) > 0$ for all $z \geq 0$. More precisely, we have with (3.93) and (5.22) that

$$\|\mathbf{L}_{\Psi_\theta,\tau}^{-1}\| \leq c_{\mathrm{inv}}^2 + \tau^2\theta. \quad (5.46)$$

Next, we show improved bounds in Lemmas 5.16 and 5.17 by employing the previous results for θ-functions. We focus on $\mathbf{L}$ positive definite because otherwise the bound (5.25a) is the best we can achieve for $\|\boldsymbol{\mathcal{S}}_{n,\star}\|$.

Lemma 5.28. *Let $\vartheta < 1$, $\mathbf{L}$ be positive definite, and let Assumption 5.26 on the step size τ hold. Then $\sin\boldsymbol{\Phi}_\star$ is nonsingular for $\tau > 0$. Moreover, we have the following.*

(a) *If $\theta > \frac{1}{4}$, we have for all $\mathbf{q} \in \mathbb{R}^m$*

$$\tau\,\|(\sin\boldsymbol{\Phi}_\star)^{-1}\mathbf{q}\| \leq \hat{c}_{\mathrm{s},1}^\star(\theta,\vartheta)\|\mathbf{q}\| \qquad \text{with} \quad \hat{c}_{\mathrm{s},1}^\star(\theta,\vartheta) = \Big(\frac{c_{\mathrm{inv}}^2 + \tau^2\theta}{1 - \max\{\frac{1}{4\theta},\vartheta^2\}}\Big)^{1/2}. \quad (5.47)$$

For $\vartheta = 1$, $\theta = \frac{1}{4}$, or $\mathbf{L}$ singular we formally set $\hat{c}_{\mathrm{s},1}^\star(\theta,\vartheta) = \infty$.

(b) *We have for all* $\mathbf{q} \in \mathbb{R}^m$

$$\tau \left\|(\sin \mathbf{\Phi}_\star)^{-1} \widehat{\Psi}_\theta(\tau^2 \mathbf{L}\mathbf{R})\mathbf{q}\right\| \le (1-\vartheta^2)^{-1/2} \|\mathbf{q}\|_{\mathbf{L}^{-1}} \le \hat{c}^\star_{\mathrm{s},2}(\vartheta)\|\mathbf{q}\| \tag{5.48}$$

with $\hat{c}^\star_{\mathrm{s},2}(\vartheta) = c_{\mathrm{inv}}(1-\vartheta^2)^{-1/2}$.
For $\vartheta = 1$ *or* $\mathbf{L}$ *singular we formally set* $\hat{c}^\star_{\mathrm{s},2}(\vartheta) = \infty$.

We emphasize that, although $\sin \mathbf{\Phi}_\star$ is nonsingular for $\vartheta < 1$ and $\mathbf{L}$ positive definite, we cannot give a uniform bound for $\|(\sin \mathbf{\Phi}_\star)^{-1}\|$ in the case of $\theta = \frac{1}{4}$; see the proof below for more details. In contrast to this, the bound in (5.48) holds for $\theta \ge \frac{1}{4}$ with constants independent of θ, similarly to the bounds in Lemma 3.63 for the modified θ-schemes. Moreover, they are optimal in the sense that the estimates would be the same if we had set $\mathbf{R} = 0$, i.e., we had the standard leapfrog scheme; cf. (3.36a) and Example 3.11.

For the proof of this lemma we need yet another block decomposition of $\mathbf{L}$. Similarly to (5.15) one can show that $\mathbf{L}$ admits the block decomposition

$$\mathbf{L} = \mathbf{C}_\mathbf{N} \mathbf{C}_\mathbf{N}^T, \qquad \mathbf{C}_\mathbf{N} = \begin{pmatrix} \mathbf{A}_\mathbf{N}^{\frac{1}{2}} & \mathbf{K}^T \mathbf{N}^+ \mathbf{N}^{\frac{1}{2}} \\ 0 & \mathbf{N}^{\frac{1}{2}} \end{pmatrix}, \tag{5.49}$$

where $\mathbf{A}_\mathbf{N} = \mathbf{S} - \mathbf{K}^T \mathbf{N}^+ \mathbf{K}$. As before we have that $\mathbf{A}_\mathbf{N}$ is symmetric and positive semidefinite and $\mathbf{K}^T \mathbf{N}^+ \mathbf{N} = \mathbf{K}^T$; see again [Alb69] or [HZ05, Theorems 1.19 and 1.20].

Proof of Lemma 5.28. From the definition of $\sin \mathbf{\Phi}_\star$ in (5.23b) we have that $\sin \mathbf{\Phi}_\star$ is nonsingular if and only if $\mathbf{L}_{\Psi_\theta,\tau}$ and $\mathbf{I}_m - \frac{1}{4}\tau^2 \mathbf{L}_{\Psi_\theta,\tau}$ are nonsingular. For $\mathbf{L}_{\Psi_\theta,\tau}$ this is shown in Lemma 5.14. For $\mathbf{I}_m - \frac{1}{4}\tau^2 \mathbf{L}_{\Psi_\theta,\tau}$ we obtain from (5.44) that it is nonsingular if $\theta > \frac{1}{4}$. With the relations in (5.45) one can further show that the matrix $\mathbf{I}_m - \frac{1}{4}\tau^2 \mathbf{L}_{\Psi_\theta,\tau}$ is positive definite by proceeding similarly as in the proof of Lemma 5.27. However, the smallest eigenvalue tends to zero if the largest eigenvalue of $\mathbf{S}$ tends to ∞. Thus, a uniform lower bound does not exist for $\theta = \frac{1}{4}$ in accordance with Lemma 5.27.

(a) The bound follows as in the proof of Lemma 5.16 by replacing the lower and upper bound for the smallest and largest eigenvalue of $\tau^2 \mathbf{L}_{\Psi_\theta,\tau}$ with the one from (5.46) and (5.44), respectively.

(b) With the definition of $\sin \mathbf{\Phi}_\star$ in (5.23b), the symmetry of $\mathbf{L}_{\Psi_\theta,\tau}$, and $\widehat{\Psi}_\theta(z)^{-1} = 1 + \theta z$ we obtain

$$\begin{aligned} \tau^2 \left\|(\sin \mathbf{\Phi}_\star)^{-1} \widehat{\Psi}_\theta(\tau^2 \mathbf{L}\mathbf{R})\mathbf{q}\right\|^2 &= ((\mathbf{I}_m - \tfrac{1}{4}\tau^2 \mathbf{L}_{\Psi_\theta,\tau})^{-1} \widehat{\Psi}_\theta(\tau^2 \mathbf{L}\mathbf{R})\mathbf{q}, \mathbf{L}_{\Psi_\theta,\tau}^{-1} \widehat{\Psi}_\theta(\tau^2 \mathbf{L}\mathbf{R})\mathbf{q}) \\ &= ((\mathbf{I}_m + \theta\tau^2 \mathbf{L}\mathbf{R} - \tfrac{1}{4}\tau^2 \mathbf{L})^{-1}\mathbf{q}, \mathbf{L}^{-1}\mathbf{q}) \\ &= ((\mathbf{I}_m + \theta\tau^2 \mathbf{C}_\mathbf{N}^T \mathbf{R}\mathbf{C}_\mathbf{N} - \tfrac{1}{4}\tau^2 \mathbf{C}_\mathbf{N}^T \mathbf{C}_\mathbf{N})^{-1} \mathbf{C}_\mathbf{N}^{-1}\mathbf{q}, \mathbf{C}_\mathbf{N}^{-1}\mathbf{q}), \end{aligned}$$

where we used the decomposition (5.49) in the last step. A simple calculation shows

$$\widetilde{\mathbf{B}} = \mathbf{I}_m + \theta\tau^2 \mathbf{C}_\mathbf{N}^T \mathbf{R}\mathbf{C}_\mathbf{N} - \tfrac{1}{4}\tau^2 \mathbf{C}_\mathbf{N}^T \mathbf{C}_\mathbf{N} = \begin{pmatrix} \mathbf{I}_s & 0 \\ 0 & \mathbf{I}_{m-s} - \frac{1}{4}\tau^2 \mathbf{N} \end{pmatrix} + (\theta - \tfrac{1}{4})\tau^2 \mathbf{C}_\mathbf{N}^T \mathbf{R}\mathbf{C}_\mathbf{N}.$$

Hence, with the step-size restriction (5.43) and $\theta \ge \frac{1}{4}$ we have that $(\widetilde{\mathbf{B}}\mathbf{q}, \mathbf{q}) \ge (1-\vartheta^2)\|\mathbf{q}\|^2$. Altogether this leads to

$$\tau^2 \left\|(\sin \mathbf{\Phi}_\star)^{-1} \widehat{\Psi}_\theta(\tau^2 \mathbf{L}\mathbf{R})\mathbf{q}\right\|^2 \le (1-\vartheta^2)^{-1} (\mathbf{C}_\mathbf{N}^{-1}\mathbf{q}, \mathbf{C}_\mathbf{N}^{-1}\mathbf{q}) = (1-\vartheta^2)^{-1} \|\mathbf{q}\|_{\mathbf{L}^{-1}}^2,$$

which is the first estimate in (5.48). The second one follows from (2.13). □

From these two lemmas we then can derive stability and error bounds analogously to the general case in Sections 5.3 and 5.4. In particular, the stability and error results in Theorems 5.18, 5.23, and 5.24 hold as before with the weaker step-size restriction (5.43) instead of (5.17) and $c^\star_{\mathrm{s},1}(\vartheta)$, $c^\star_{\mathrm{s},2}(\vartheta)$ replaced with$\hat{c}^\star_{\mathrm{s},1}(\theta,\vartheta)$, $\hat{c}^\star_{\mathrm{s},2}(\vartheta)$, respectively. For the error bound of the semilinear problem (2.1) with the weaker Lipschitz condition from Assumption 2.9 we obtain the following improved result compared to Theorem 5.25.

Corollary 5.29. *Let $\vartheta \in (0,1)$, $\mathbf{L}$ be positive definite, and let Assumptions 3.2, 3.16, 5.1, as well as Assumption 2.9 on $\mathbf{g}$ hold. Further, assume that for $T \in (0, t_*)$ the solution $\mathbf{q}$ of (2.1) satisfies $\mathbf{q} \in C^4([0,T])$. Then, there exists a $\tau_* > 0$ such that for $\tau \le \min\{\tau_*, 2\vartheta/\|\mathbf{N}\|^{1/2}\}$ and $t_n \le T$ we have for the approximations $\mathbf{q}_n$ of the scheme (5.4) equipped with θ-functions (3.4)*

$$\|\mathbf{q}(t_n) - \mathbf{q}_n\| \le \big(\min\{T, \hat{c}^\star_{\mathrm{s},1}(\theta,\vartheta)\}(C_{\blacklozenge} T + C_{\bigstar}) + C_{\blacktriangle}\big) \mathrm{e}^{(1-\vartheta^2)^{-1/2}\widehat{\mathcal{L}}_{\mathbf{g}} T} \tau^2 \tag{5.50}$$

with $C_{\blacklozenge}$, $C_{\bigstar}$, and $C_{\blacktriangle}$ defined as in Theorem 5.23.

Proof. The proof follows as in Theorem 5.25 by replacing the bounds from Lemmas 5.16 and 5.17 with the ones in Lemma 5.28, which hold under the weaker step-size restriction $\tau^2 \le 4\vartheta^2/\|\mathbf{N}\|$; cf. Assumption 5.26. □

Last, we point out that for $\theta = \frac{1}{4}$ we always have $\min\{T, \hat{c}^\star_{\mathrm{s},1}(\theta,\vartheta)\} = T$ in contrast to $\theta > \frac{1}{4}$ because of the definition of $\hat{c}^\star_{\mathrm{s},1}(\theta,\vartheta)$ in (5.47) above. Thus, we get an additional factor T in the stability and error bounds. However, we are convinced that this factor T can be replaced with a constant by a modification of the starting value and a more involved error analysis; cf. the results for the closely related locally implicit scheme in [HS16], where the error bound grows only linearly in time. From Lemma 5.28 we see that this would be the case if we had the factor $\widehat{\Psi}_\theta(\tau^2\mathbf{LR})$ everywhere where the matrix $(\sin \mathbf{\Phi}_\star)^{-1}$ appears, since the bound (5.48) admits a uniform constant also for $\theta = \frac{1}{4}$.

5.6. Implementation and efficiency for two specific functions

The aim of this section is to present efficient implementations of the multirate scheme (5.4) equipped with the LFC polynomials (4.1b) (*sLFC scheme*) and with the θ-functions (3.4) (*split-θ-scheme*). Moreover, we show that in the situation of Assumption 5.1 described at the beginning of this chapter these multirate schemes are beneficial in terms of computational effort compared to the standard leapfrog scheme (2.20).

Recall that we have shown in the last chapter that the LFC polynomials satisfy Assumptions 3.2 and 3.16. Further, values for all occurring constants are stated in Sections 4.2 and 4.2.2. In particular, values for the constants $\widehat{\beta}_\Psi$ and m_1 are given in Theorem 4.9 and for the special choice $\nu = \nu_{p,\eta}$ defined in (4.3) in Lemma 4.12. For the θ-functions we use the improved results from the last section. Further results for these functions are shown in Section 3.6, e.g., that they satisfy Assumptions 3.2 and 3.16.

5.6.1. Implementation

For the sake of readability we focus on the implementation of the two-step schemes, since for the implementation of the corresponding one-step formulations similar strategies can be applied. In particular, the dominating parts of the computational costs coincide for both the two-step

and the one-step variants. For similar reasons we also omit the implementation of the starting value (5.4b). Moreover, as before we present the implementation of the multirate schemes for the more general problem (2.2) with a general mass matrix $\mathbf{M}$; see Remark 5.3 for changes in the general scheme. Hence, we consider the two-step scheme (5.6a) instead of (5.4).

We point out that for both schemes the efficient implementation is based on a block representation of the matrix $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{C}_{\mathbf{M}}^{-T}\mathbf{R}\mathbf{C}_{\mathbf{M}}^{-1})$ occurring in (5.6a); cf. the block representation (5.14) for $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$ if $\mathbf{M} = \mathbf{I}_m$. For a more concise presentation of the subsequent algorithms, we use the notation in (5.19), i.e., for $\mathbf{v} \in \mathbb{R}^m$ we denote by $\mathbf{v}_S \in \mathbb{R}^s$ and $\mathbf{v}_N \in \mathbb{R}^{m-s}$ the subvectors of $\mathbf{v}$ belonging to the stiff and the nonstiff part of the differential equation.

Implementation of the sLFC scheme

We start with the implementation of the sLFC scheme, i.e., we consider the two-step scheme (5.6a) equipped with the LFC polynomials (4.1b). For this, we first have a closer look at the matrix function $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{C}_{\mathbf{M}}^{-T}\mathbf{R}\mathbf{C}_{\mathbf{M}}^{-1})$. Proceeding similarly to the calculation of $\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{R})$ in (5.14) yields with $\mathbf{M}$ defined in (5.7)

$$\begin{aligned}\widehat{\Psi}(\tau^2\mathbf{L}\mathbf{C}_{\mathbf{M}}^{-T}\mathbf{R}\mathbf{C}_{\mathbf{M}}^{-1}) &= \mathbf{I}_m + \tau^2 \begin{pmatrix} \mathbf{S}\mathbf{M}_S^{-1} & 0 \\ \mathbf{K}\mathbf{M}_S^{-1} & 0 \end{pmatrix} \begin{pmatrix} \Upsilon(\tau^2\mathbf{S}\mathbf{M}_S^{-1}) & 0 \\ 0 & 0 \end{pmatrix} \\ &= \mathbf{I}_m + \tau^2 \begin{pmatrix} \mathbf{S} & 0 \\ \mathbf{K} & 0 \end{pmatrix} \begin{pmatrix} \Upsilon(\tau^2\mathbf{M}_S^{-1}\mathbf{S})\mathbf{M}_S^{-1} & 0 \\ 0 & 0 \end{pmatrix},\end{aligned} \tag{5.51}$$

where we adopted the notation from Remark 5.3. Employing this block structure for the LFC polynomials P_p yields Algorithm 5.1, which contains the computation of one time step of the two-step sLFC scheme. We state the algorithm only for $p \geq 2$, since for $p = 1$ the scheme reduces to the leapfrog scheme; see (4.5).

Algorithm 5.1.: Computation of nth time step of two-step sLFC scheme (5.6a) with LFC polynomials (4.1b) for $p \geq 2$ (indices of vectors as in (5.19)).

1: Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
2: $\mathbf{v} = -\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n$
3: Solve $\mathbf{M}_S\check{\mathbf{v}}_S = \mathbf{v}_S$
4: Compute $\widetilde{\mathbf{v}}_S = \Upsilon(\tau^2\mathbf{M}_S^{-1}\mathbf{S})\check{\mathbf{v}}_S$ by Algorithm 5.2 below
5: $\bar{\mathbf{v}}_S = \mathbf{v}_S + \tau^2\mathbf{S}\widetilde{\mathbf{v}}_S$
6: $\bar{\mathbf{v}}_N = \mathbf{v}_N + \tau^2\mathbf{K}\widetilde{\mathbf{v}}_S$
7: Solve $\mathbf{M}\widehat{\mathbf{v}} = \bar{\mathbf{v}}$
8: $\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n-1} + \tau^2\widehat{\mathbf{v}}$

For the computation of $\Upsilon(\tau^2\mathbf{S})\check{\mathbf{v}}_S$ the same observations hold as for $\widehat{P}_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\widetilde{\mathbf{v}}$ in Algorithm 4.1; see the comments below this algorithm. In particular, the computation via a linear three-term recurrence relation is advantageous in terms of stability over a computation via Horner's method.

Lemma 5.30. *Let $p \in \mathbb{N}$, $k \in \mathbb{N}$. The polynomials $\Upsilon_{k,p}\colon \mathbb{R} \to \mathbb{R}$, defined by*

$$\Upsilon_{k,p}(z) = \frac{1}{z^2}\left(2 - \frac{2}{T_k(\nu)}T_k\Big(\nu - \frac{z}{\alpha_p}\Big) - \frac{\alpha_k}{\alpha_p}z\right),$$

satisfy the linear recurrence relation

$$\Upsilon_{1,p}(z) = 0, \qquad \Upsilon_{2,p}(z) = -\frac{4}{\alpha_p^2 T_2(\nu)},$$
$$T_{k+1}(\nu)\Upsilon_{k+1,p}(z) = 2\nu T_k(\nu)\Upsilon_{k,p}(z) - 2\frac{T_k(\nu)}{\alpha_p}\Big(\frac{\alpha_k}{\alpha_p} + z\Upsilon_{k,p}(z)\Big) - T_{k-1}(\nu)\Upsilon_{k-1,p}(z),$$

for $k \geq 2$.

Proof. It is sufficient to show a recurrence relation for

$$R_{k,p}(z) = z^2\Upsilon_{k,p}(z)$$

because the recurrence relation for $\Upsilon_{k,p}$ simply follows by division of z^2. For $k = 1$ and $k = 2$ the statements are easily verified by direct calculations.

For $k \geq 2$ we obtain by using Lemma 4.15

$$\begin{aligned} T_{k+1}(\nu)R_{k+1,p}(z) &= T_{k+1}(\nu)P_{k+1,p}(z) - T_{k+1}(\nu)\frac{\alpha_{k+1}}{\alpha_p}z \\ &= 2\Big(\nu - \frac{z}{\alpha_p}\Big)T_k(\nu)P_{k,p}(z) - T_{k-1}(\nu)P_{k-1,p}(z) + \frac{4}{\alpha_p}T_k(\nu)z - T_{k+1}(\nu)\frac{\alpha_{k+1}}{\alpha_p}z. \end{aligned}$$

For the last term on the right-hand side we take the derivative of the Chebyshev recurrence relation (B.9)

$$T'_{k+1}(\nu) = 2\nu T'_k(\nu) + 2T_k(\nu) - T'_{k-1}(\nu)$$

in order to get

$$\begin{aligned} -T_{k+1}(\nu)\frac{\alpha_{k+1}}{\alpha_p}z = -\frac{2T'_{k+1}(\nu)}{\alpha_p}z &= -\frac{2}{\alpha_p}z\big(2\nu T'_k(\nu) - T'_{k-1}(\nu) + 2T_k(\nu)\big) \\ &= -2\nu T_k(\nu)\frac{\alpha_k}{\alpha_p}z + T_{k-1}(\nu)\frac{\alpha_{k-1}}{\alpha_p}z - \frac{4}{\alpha_p}T_k(\nu)z. \end{aligned}$$

Inserting this into the above equation yields

$$\begin{aligned} T_{k+1}(\nu)R_{k+1,p}(z) &= 2\Big(\nu - \frac{z}{\alpha_p}\Big)T_k(\nu)P_{k,p}(z) - 2\nu T_k(\nu)\frac{\alpha_k}{\alpha_p}z - T_{k-1}(\nu)\Big(P_{k-1,p}(z) - \frac{\alpha_{k-1}}{\alpha_p}z\Big) \\ &= 2\Big(\nu - \frac{z}{\alpha_p}\Big)T_k(\nu)R_{k,p}(z) - 2T_k(\nu)\frac{\alpha_k}{\alpha_p^2}z^2 - T_{k-1}(\nu)R_{k-1,p}(z), \end{aligned}$$

which completes the proof. □

The fact that $\Upsilon_{k,p}$ are polynomials can be seen from the recurrence relation because only constant terms and terms multiplied with z occur in the recurrence relation. More precisely, $\Upsilon_{k,p}$ are polynomials of degree $k-1$. We further point out that because of the definition of the LFC polynomials (4.1b) and Definition 3.12 of Υ we have

$$\Upsilon_{p,p}(z) = \frac{P_p(z) - z}{z^2} = \Upsilon(z) \qquad \text{for all } p \in \mathbb{N}.$$

Hence, we have derived a linear three-term recurrence relation for Υ if $\Psi = P_p$.

In Algorithm 5.2 we present the details for the computation of $\Upsilon(\tau^2\mathbf{S})\check{\mathbf{v}}_S$ via the above recurrence relation, where $\mathbf{w}_k = \Upsilon_{k,p}(\tau^2\mathbf{M}_S^{-1}\mathbf{S})\check{\mathbf{v}}_S$. As for the computation of $P_p(\tau^2\mathbf{M}^{-1}\mathbf{L})\tilde{\mathbf{v}}$ in Algorithm 4.2 the values of the parameters $T_0(\nu), \ldots, T_p(\nu)$, and α_p for a fixed $\nu \geq 1$ have to be computed only once by means of the Chebyshev recurrence relations (B.9) and (B.18).

Algorithm 5.2.: Computation of $\Upsilon(\tau^2\mathbf{M}_S^{-1}\mathbf{S})\check{\mathbf{v}}_S$ in Algorithm 5.1 for LFC polynomials (4.1b).

$\mathbf{w}_1 = 0$, $\mathbf{w}_2 = -\frac{4}{\alpha_p^2 T_2(\nu)}\check{\mathbf{v}}_S$
for $k = 3, \dots, p$ **do**
  Solve $\mathbf{M}_S\widetilde{\mathbf{w}}_{k-1} = \mathbf{S}\mathbf{w}_{k-1}$
  $\mathbf{w}_k = 2\nu\frac{T_{k-1}(\nu)}{T_k(\nu)}\mathbf{w}_{k-1} - \frac{2}{\alpha_p}\frac{T_{k-1}(\nu)}{T_k(\nu)}\left(\frac{\alpha_k}{\alpha_p}\check{\mathbf{v}}_S + \tau^2\widetilde{\mathbf{w}}_{k-1}\right) - \frac{T_{k-2}(\nu)}{T_k(\nu)}\mathbf{w}_{k-2}$
end for
$\widetilde{\mathbf{v}}_S = \mathbf{w}_p$

We conclude this section about the implementation of the sLFC scheme by a closer investigation of the matrix-vector product $\mathbf{g}_n^* = \mathbf{M}\mathbf{g}_n$ in the first step of Algorithm 5.1. Compared to Algorithm 4.1 we have an additional matrix-vector multiplication with $\mathbf{M}$ in Algorithm 5.2 which we want to eliminate (although it is often cheap in applications). Clearly, this is not a problem at all if $\mathbf{g}_n^*$ is given explicitly, cf. the comments above Algorithm 4.1.

Nevertheless, we still can save the matrix-vector multiplication of $\mathbf{M}$ with $\mathbf{g}_n$ if the mass matrix $\mathbf{M}$ defined as in (5.7) is block diagonal with blocks corresponding to those of $\mathbf{L}$ given in (5.1), i.e., $\mathbf{M}_K = 0$. This implies that the mass matrix $\mathbf{M}$ and the restriction matrix $\mathbf{R}$ commute and the two-step scheme (5.6a) can be written as

$$(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1}) = \tau^2\widehat{\Psi}(\tau^2\mathbf{M}^{-1}\mathbf{L}\mathbf{R})(-\mathbf{M}^{-1}\mathbf{L}\mathbf{q}_n + \mathbf{g}_n),$$

where

$$\widehat{\Psi}(\tau^2\mathbf{M}^{-1}\mathbf{L}\mathbf{R}) = \mathbf{I}_m + \tau^2\begin{pmatrix}\mathbf{M}_S^{-1}\mathbf{S} & 0\\ \mathbf{M}_N^{-1}\mathbf{K} & 0\end{pmatrix}\begin{pmatrix}\Upsilon(\tau^2\mathbf{M}_S^{-1}\mathbf{S}) & 0\\ 0 & 0\end{pmatrix}.$$

Algorithm 5.3 states the details for $\widehat{\Psi} = \widehat{P}_p$.

Algorithm 5.3.: Computation of nth time step of two-step sLFC scheme (5.6a) with LFC polynomials (4.1b) for $p \geq 2$ if $\mathbf{M}_K = 0$ defined in (5.7) (indices of vectors as in (5.19)).

Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
$\mathbf{v}_1 = -\mathbf{L}\mathbf{q}_n$, $\mathbf{v}_2 = \mathbf{g}_n$
Solve $\mathbf{M}_S\check{\mathbf{v}}_{1,S} = \mathbf{v}_{1,S}$
$\check{\mathbf{v}}_S = \check{\mathbf{v}}_{1,S} + \mathbf{v}_{2,S}$
Compute $\widetilde{\mathbf{v}}_S = \Upsilon(\tau^2\mathbf{M}_S^{-1}\mathbf{S})\check{\mathbf{v}}_S$; by Algorithm 5.2
Solve $\mathbf{M}_S\bar{\mathbf{v}}_S = \mathbf{S}\widetilde{\mathbf{v}}_S$
$\widehat{\mathbf{v}}_S = \check{\mathbf{v}}_S + \tau^2\bar{\mathbf{v}}_S$
Solve $\mathbf{M}_N\bar{\mathbf{v}}_S = \tau^2\mathbf{K}\widetilde{\mathbf{v}}_S + \mathbf{v}_{1,N}$
$\widehat{\mathbf{v}}_N = \mathbf{v}_{2,N} + \bar{\mathbf{v}}_S$
$\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n-1} + \tau^2\widehat{\mathbf{v}}$

Implementation of the split-θ-scheme

Next, we present an efficient implementation of the split-θ-scheme, i.e., we consider the two-step scheme (5.6a) equipped with the rational functions (3.4). To state an algorithm, we first observe that for $\widehat{\Psi} = \widehat{\Psi}_\theta$ the two-step scheme (5.6a) with $\mathbf{M}$ given as in (5.7) is equivalent to

$$\widehat{\Psi}_\theta(\tau^2\mathbf{L}\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1})^{-1}\mathbf{M}(\mathbf{q}_{n+1} - 2\mathbf{q}_n + \mathbf{q}_{n-1}) = \tau^2(-\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n), \tag{5.52}$$

where

$$\widehat{\Psi}_\theta(\tau^2\mathbf{L}\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1})^{-1} = \mathbf{I}_m + \theta\tau^2 \begin{pmatrix} \mathbf{S} & 0 \\ \mathbf{K} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{M}_S^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{M}_S + \theta\tau^2\mathbf{S} & 0 \\ \theta\tau^2\mathbf{K} & \mathbf{I}_{m-s} \end{pmatrix} \begin{pmatrix} \mathbf{M}_S^{-1} & 0 \\ 0 & \mathbf{I}_{m-s} \end{pmatrix}$$

because of $\widehat{\Psi}_\theta^{-1}(z) = 1 + \theta z$. The details of the implementation are given in Algorithm 5.4. We observe that, if the computation of the solution of the linear system with $\mathbf{M}$ is cheap, we only have to solve a small linear system of size $s \ll m$ in contrast to the modified θ-schemes, where a system of size m has to be solved; cf. Algorithm 3.1.

Algorithm 5.4.: Computation of nth time step of two-step split-θ-scheme (5.6a), (3.4) (indices of vectors as in (5.19)).

1: Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
2: $\mathbf{v} = -\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n$
3: Solve $(\mathbf{M}_S + \theta\tau^2\mathbf{S})\widetilde{\mathbf{v}}_S = \mathbf{v}_S$
4: $\bar{\mathbf{v}}_N = \mathbf{v}_N - \theta\tau^2\mathbf{K}\widetilde{\mathbf{v}}_S$
5: $\bar{\mathbf{v}}_S = \mathbf{M}_S\widetilde{\mathbf{v}}_S$
6: Solve $\mathbf{M}\widehat{\mathbf{v}} = \bar{\mathbf{v}}$
7: $\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n+1} + \tau^2\widehat{\mathbf{v}}$

Similarly as before, we can save some computational costs if we assume that $\mathbf{M}_K = 0$ in (5.7). We then have in (5.52)

$$\widehat{\Psi}_\theta(\tau^2\mathbf{L}\mathbf{C}_\mathbf{M}^{-T}\mathbf{R}\mathbf{C}_\mathbf{M}^{-1})^{-1}\mathbf{M} = \begin{pmatrix} \mathbf{M}_S + \theta\tau^2\mathbf{S} & 0 \\ \theta\tau^2\mathbf{K} & \mathbf{M}_N \end{pmatrix}.$$

The details are stated in Algorithm 5.5. In comparison to Algorithm 5.4 we save with this implementation a matrix-vector multiplication with $\mathbf{M}_S$. In addition, we have to solve only with $\mathbf{M}_N$ instead of $\mathbf{M}$.

Algorithm 5.5.: Computation of nth time step of two-step split-θ-scheme (5.6a), (3.4) if $\mathbf{M}_K = 0$ defined in (5.7) (indices of vectors as in (5.19)).

1: Evaluate $\mathbf{g}_n = \mathbf{g}(t_n, \mathbf{q}_n)$
2: $\mathbf{v} = -\mathbf{L}\mathbf{q}_n + \mathbf{M}\mathbf{g}_n$
3: Solve $(\mathbf{M}_S + \theta\tau^2\mathbf{S})\widehat{\mathbf{v}}_S = \mathbf{v}_S$
4: $\widetilde{\mathbf{v}}_N = \mathbf{v}_N - \theta\tau^2\mathbf{K}\widehat{\mathbf{v}}_S$
5: Solve $\mathbf{M}_N\widehat{\mathbf{v}}_N = \widetilde{\mathbf{v}}_N$
6: $\mathbf{q}_{n+1} = 2\mathbf{q}_n - \mathbf{q}_{n+1} + \tau^2\widehat{\mathbf{v}}$

5.6.2. Costs and efficiency

We now compare the efficiency of the standard leapfrog scheme (2.20) with the sLFC scheme and the split-θ-scheme. As in Section 4.4.2 about the costs and efficiency of the LFC scheme, we here focus only on the comparison between the maximum step sizes, for which the schemes are stable, and the required computational cost, neglecting the influence of the step size to the accuracy of the approximations; cf. Remark 4.18.

As in the last chapter we first look at the main effort of these schemes per time step in terms of matrix-vector multiplications, evaluations of $\mathbf{g}$, and solving of linear systems. Since we always

Table 5.1.: Comparison of main costs per time step of leapfrog scheme, sLFC scheme (with polynomial of degree p), and split-θ-scheme in terms of matrix-vector multiplications (MVM), solutions of linear systems, and evaluations of $\mathbf{g}$, if implemented as in Algorithms 4.1, 5.3, and 5.5 for the case $\mathbf{M}_K = 0$.

leapfrog scheme (2.20a)	sLFC scheme (5.4a), (4.1b)	split-θ-scheme (5.4a), (3.4)
1 evaluation of $\mathbf{g}$	1 evaluation of $\mathbf{g}$	1 evaluation of $\mathbf{g}$
1 MVM with $\mathbf{L}$	1 MVM with $\mathbf{L}$	1 MVM with $\mathbf{L}$ and $\mathbf{M}$
	$p-1$ MVMs with $\mathbf{S}$	1 linear system with $\mathbf{M}_S + \theta\tau^2\mathbf{S}$
	p linear systems with $\mathbf{M}_S$	
	1 MVM with $\mathbf{K}$	1 MVM with $\mathbf{K}$
1 linear system with $\mathbf{M}$	1 linear system with $\mathbf{M}_N$	1 linear system with $\mathbf{M}_N$

have $\mathbf{M}_K = 0$ in our applications, we state the effort only for the corresponding algorithms. In Table 5.1 these costs are given if implemented as in Algorithms 4.1, 5.3, and 5.5. We observe that for the sLFC scheme and for the split-θ-scheme the additional costs compared to the leapfrog scheme are comparatively cheap – possibly with the exception of the multiplication with $\mathbf{M}$ – because of $s \ll m$. For Algorithms 5.1 and 5.4 with general $\mathbf{M}_K$ the computational effort is only slightly larger, if multiplications and linear systems with $\mathbf{M}$ are cheap.

In order to obtain the total cost, we now relate the effort per time step to the maximum step sizes for which the schemes are stable. We neglect the influence of the starting value, since it has to be computed only once. Moreover, the ratio of the effort of the starting values for the different schemes is similar as for the two-step schemes. To simplify the presentation we omit the mass matrices in the following and set $\vartheta = 1$.

We shortly recall Assumption 5.1, where we postulated that the matrix $\mathbf{S}$ is small compared to the whole matrix $\mathbf{L}$ but determines the norm of the matrix $\mathbf{L}$, whereas the matrices $\mathbf{N}$ is of moderate norm but of large size. For the coupling matrix $\mathbf{K}$ we assumed a rather weak coupling $\kappa \ll r^{1/2}$ where one typically has $\kappa \approx 1$ or even $\kappa < 1$.

With Assumption 5.1 the step-size restriction for the leapfrog scheme is given by

$$\tau^2 \leq \tau^2_{\mathrm{SSR,LF}} = \frac{4}{\|\mathbf{L}\|} \approx \frac{4}{\|\mathbf{S}\|} = \frac{4}{r\|\mathbf{N}\|}.$$

To state the step-size restriction of the sLFC scheme we recall (5.16a) under Assumption 5.1

$$\tau^2 \leq \tau^2_{\mathrm{SSR}} = \min\left\{\frac{\widehat{\beta}^2_\Psi}{\|\mathbf{S}\|}, \frac{4\gamma}{\|\mathbf{N}\|}\right\} = \frac{1}{\|\mathbf{N}\|}\min\left\{\frac{\widehat{\beta}^2_\Psi}{r}, 4\gamma\right\}, \quad \gamma = \frac{2}{1+(1+4\kappa^2 m_1^{-1})^{1/2}}. \tag{5.53}$$

Lemma 5.31. *Let $p \in \mathbb{N}$, $\eta > 0$, and let Assumption 5.1 be satisfied. For the LFC polynomials P_p defined in (4.1b) with the special choice of $\nu = \nu_{p,\eta}$ given in (4.3) the step-size restriction (5.53) is satisfied for $\tau_{\mathrm{SSR}} = \tau_{\mathrm{SSR,sLFC}}$, where*

$$\mathcal{L}^{p,r}_{\eta,\kappa} \leq \tau^2_{\mathrm{SSR,sLFC}}\|\mathbf{N}\| = \min\left\{\frac{\widehat{\beta}^2_{p,\nu}}{r}, 4\gamma\right\} \leq \mathcal{U}^{p,r}_\kappa$$

with

$$\mathcal{L}^{p,r}_{\eta,\kappa} = \min\left\{\frac{4p^2}{r}\frac{(1+\frac{1}{4}\eta^2)^{1/2}}{1+\frac{1}{2}\eta^2}, 4\gamma^{\eta,\kappa}_L\right\}, \qquad \gamma^{\eta,\kappa}_L = 2\left(1+\left(1+8\kappa^2\tfrac{2+\eta^2}{\eta^2}\right)^{1/2}\right)^{-1},$$

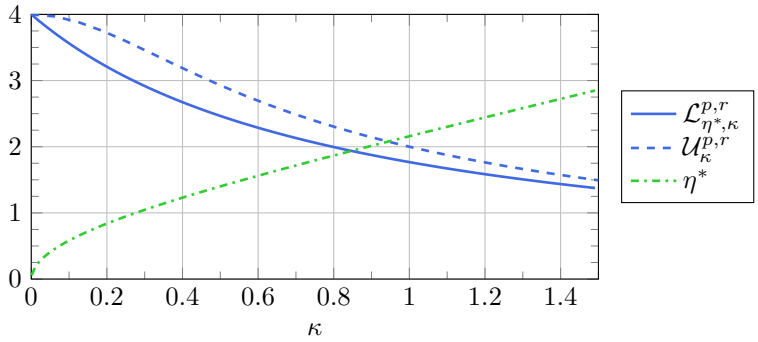


Figure 5.2.: Theoretical lower and upper bounds of $\tau^2_{\mathrm{SSR,sLFC}}\|\mathbf{N}\|$ plotted against the coupling parameter κ for $p^2 = r$. The values η^* are determined such that $\mathcal{L}^{p,r}_{\eta,\kappa}$ is maximal in dependency of κ.

and

$$\mathcal{U}^{p,r}_{\kappa} = \min\Big\{\frac{4p^2}{r}, 4\gamma^{\kappa}_{U}\Big\}, \qquad \gamma^{\kappa}_{U} = 2\Big(1 + (1+8\kappa^2)^{1/2}\Big)^{-1}.$$

Proof. By definition we have $\tau_{\mathrm{SSR}} = \tau_{\mathrm{SSR,sLFC}}$. The upper bound follows from the fact that $\widehat{\beta}^2_{p,\nu} \le 4p^2$ and $m^{p,\nu}_1 \le \frac{1}{2}$; see Lemma A.7. For the lower bound we employ Lemma 4.12; see also Lemma A.7. □

The lemma implies that, if we choose the polynomial degree $p \in \mathbb{N}$ of the LFC polynomial such that $p^2 \approx r$, the step-size restriction $\tau \le \tau_{\mathrm{SSR,sLFC}}$ of the sLFC scheme only depends on η, the submatrix $\mathbf{N}$, and the factor κ, but is independent of $\mathbf{S}$ (and the polynomial degree p). In Figure 5.2 the bounds $\mathcal{L}^{p,r}_{\eta,\kappa}$ and $\mathcal{U}^{p,r}_{\kappa}$ for $\tau^2_{\mathrm{SSR,sLFC}}\|\mathbf{N}\|$ are plotted against the coupling parameter κ for $p^2 = r$, where $\eta = \eta^*(\kappa)$ is chosen such that $\mathcal{L}^{p,r}_{\eta^*,\kappa}$ is maximal. As indicated in the comments after Definition 5.10 we get for $\kappa = 0$ the same step-size restriction as for the leapfrog scheme applied to the nonstiff problem (2.1) with $\mathbf{S} = \mathbf{K} = 0$. Clearly, as shown for the LFC scheme (4.1), the stabilization parameter should in general be chosen greater zero to avoid (linear) instabilities; cf. Chapters 3 and 4. We further observe in accordance with the definition (5.16a) of τ_{SSR} that with increasing κ the optimal value for η increases and the step-size restriction becomes stronger. In Section 5.7.1 we confirm this dependency on κ by a simple numerical experiment.

We emphasize that in our applications we often observe a weaker step-size restriction than predicted by our theory. Besides the fact that κ is often smaller than 1, taking only the value κ to model the coupling between "stiff" and "nonstiff" components turns out to be rather pessimistic. Depending on the exact structure of the coupling matrix $\mathbf{K}$, one can observe numerically a larger value for $\tau_{\mathrm{SSR,sLFC}}$ if the stabilization parameter η is chosen appropriately; see Remark 5.32 below for appropriate choices as well as the numerical experiments in Sections 5.7.2 and 5.7.3.

Remark 5.32 (Choice of η). Our numerical experiments indicate that the choice $\eta \in [0.4, 1]$ is sufficient for $\kappa \le 1$. For smaller values of η instabilities can occur for certain step sizes; cf. the numerical examples in Sections 5.7.1 and 5.7.3. If η is chosen too large, the value for $\widehat{\beta}_{p,\nu_{p,\eta}}$ deteriorates rapidly; see Lemma 4.12 and also Figures 4.3 and 4.4. Additionally, for semilinear problems one requires η sufficiently large to compensate small to moderate instabilities occurring from $\mathbf{g}$; cf. Section 3.3.3. As initial guess we thus suggest to use $\eta = 0.5$ as for the LFC schemes; see Remark 4.19. ◇

The step-size restriction of the split-θ-scheme for $\theta \geq \frac{1}{4}$ is stated in (5.43). Hence, if $\theta \geq \frac{1}{4}$, we can allow for a $r^{1/2}$ times larger step size than the standard leapfrog scheme. Thus, if r is large enough, the (small) additional effort per time step is fully compensated by the larger step sizes we can allow for leading to a more efficient scheme. Clearly, the improvement in the efficiency compared to the leapfrog scheme strongly depends on the size s of the "stiff" submatrix $\mathbf{S}$ and the factor r.

We conclude this section with the observation that the larger the ratio r between the norms of $\mathbf{S}$ and $\mathbf{N}$, the larger the polynomial degree p can be chosen in the sLFC scheme. Hence, the computational effort increases for the sLFC scheme, whereas it stays constant for the split-θ-scheme. Thus, there is a threshold which determines the minimum value of r where the split-θ-scheme becomes more efficient than the sLFC scheme.

5.7. Analytical and numerical examples

We conclude this chapter with some examples confirming our theoretical findings. The first example is of theoretical nature in which we verify the necessity of the upper bound in (3.12) in Definition 3.9(b) for the function Ψ to obtain stability of the multirate scheme (5.4). Afterwards we turn towards numerical experiments in Sections 5.7.2 and 5.7.3 by considering again specific situations of the more realistic examples from Section 2.2.

The examples in Sections 5.7.1 and 5.7.2 as well as the second one in Section 5.7.3 are modifications and extensions of the ones in [CH21, Section 6]. The codes for reproducing the numerical results are available on https://doi.org/10.5445/IR/1000147744.

5.7.1. A two-dimensional problem

With this analytical example we show that it is in general not sufficient for stability of the scheme that the function Ψ satisfies the upper bound in (3.11). Moreover, we show that in general the step-size restriction (5.17) indeed depends on the coupling parameter κ defined in Assumption 5.1. To verify these statements, we consider the simple two-dimensional linear problem

$$\ddot{\mathbf{q}}(t) = -\mathbf{L}\mathbf{q}(t), \qquad \mathbf{L} = \begin{pmatrix} r & \kappa \\ \kappa & 1 \end{pmatrix}, \tag{5.54}$$

with $r > 1$ and $|\kappa| \leq r^{1/2}$. The assumption on κ guarantees the positive semidefiniteness of $\mathbf{L}$.

Obviously, this problem fits into the setting of Assumption 5.1 if we set $\mathbf{S} = r$, $\mathbf{N} = 1$, and $\mathbf{K} = \kappa$ in (5.1). From Lemma 5.8 we then obtain for $\tau^2\mathbf{L}_{\Psi,\tau}$ with the Definitions 3.3 and 3.12 of Ψ and Υ

$$\tau^2\mathbf{L}_{\Psi,\tau} = \begin{pmatrix} \Psi(\tau^2 r) & \tau^2\widehat{\Psi}(\tau^2 r)\kappa \\ \tau^2\widehat{\Psi}(\tau^2 r)\kappa & \tau^2 1 + \tau^4\Upsilon(\tau^2 r)\kappa^2 \end{pmatrix} = \begin{pmatrix} \Psi(\tau^2 r) & \Psi(\tau^2 r)\rho \\ \Psi(\tau^2 r)\rho & \tau^2(1-\kappa\rho) + \Psi(\tau^2 r)\rho^2 \end{pmatrix}, \tag{5.55}$$

where $\rho = \kappa r^{-1} \in [0, r^{-1/2}]$.

Recall that for stability we need as minimum requirement that the spectrum of $\tau^2\mathbf{L}_{\Psi,\tau}$ is contained in $[0, 4]$. The eigenvalues of $\tau^2\mathbf{L}_{\Psi,\tau}$ are given by

$$\begin{aligned} \lambda_\pm = \lambda_\pm(\tau^2\mathbf{L}_{\Psi,\tau}) &= \tfrac{1}{2}\Psi(\tau^2 r)(1+\rho^2) + \tfrac{1}{2}\tau^2(1-\kappa\rho) \\ &\quad \pm \tfrac{1}{2}\Big(\Psi(\tau^2 r)^2(1+\rho^2)^2 - 2\Psi(\tau^2 r)\tau^2(1-\rho^2)(1-\kappa\rho) + \tau^4(1-\kappa\rho)^2\Big)^{1/2}. \end{aligned}$$

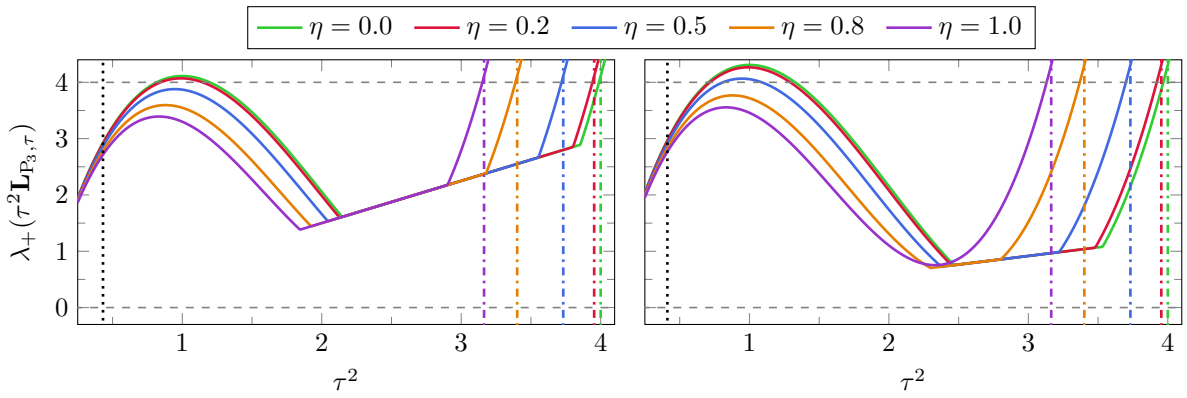


Figure 5.3.: Larger eigenvalue λ_+ of (5.55) with $r = 9$, $\kappa = 1.5$ (left) and $\kappa = 2.5$ (right), and $\Psi = P_p$ (leapfrog-Chebyshev polynomials) plotted over step sizes τ^2. For P_p we use polynomial degree $p = r^{1/2} = 3$ and stabilization parameters $\eta = 0$, $\eta = 0.2$, $\eta = 0.5$, $\eta = 0.8$, and $\eta = 1$. The dash-dotted lines indicate, where the polynomials $\tau^2 \mapsto P_3(\tau^2 r)$ leave the interval $[0, 4]$, the black dotted line indicates the maximum step size, for which the leapfrog scheme applied to (5.54) is stable.

Using $1 - \rho^2 \leq 1 + \rho^2$ yields for the larger eigenvalue λ_+

$$\lambda_+ \geq \tfrac{1}{2}\Psi(\tau^2 r)(1+\rho^2) + \tfrac{1}{2}\tau^2(1-\kappa\rho) + \tfrac{1}{2}\big|\Psi(\tau^2 r)(1+\rho^2) - \tau^2(1-\kappa\rho)\big|.$$

Hence, we have

$$\lambda_+ \geq \Psi(\tau^2 r)(1+\rho^2) \qquad \text{if } \tau^2 \leq \Psi(\tau^2 r)\frac{1+\rho^2}{1-\kappa\rho}. \tag{5.56}$$

From the estimate of λ_+ we see that the weaker step-size restriction $\tau^2 \leq \min\{\beta_\Psi^2/r, 4\}$ is in general not sufficient to guarantee $\lambda_+ \leq 4$, since from the definition Definition 3.9(a) of β_Ψ we only have $\Psi(z) \leq 4$ for all $z \in [0, \beta_\Psi^2]$. Hence, condition (3.12) with $m_1 > 0$ is indeed necessary to ensure $\lambda_+ \leq 4$. Note that the restriction on τ in (5.56) is only required to distinguish which term in the absolute value is larger. In particular, this is satisfied for $\Psi(\tau^2 r)$ near 4, since $\tau^2 \leq 4$ even under the weaker step-size restriction.

We further observe that for a fixed r the parameter $\rho = \kappa r^{-1}$ increases with increasing κ. Thus, if we also fix the step size τ, we see that the stronger the coupling, the greater λ_+ can become because of the estimate (5.56).

In Figure 5.3 these theoretical results are illustrated for the LFC polynomials (4.1b) by plotting the larger eigenvalue λ_+ of (5.55) for different stabilization parameters $\eta \geq 0$. We choose $r = 9$, and, hence, $p = 3$ as polynomial degree for the leapfrog-Chebyshev polynomial; cf. Lemma 5.31 and the comments below. For the coupling parameter κ we use two rather larger values for a better visualization of its influence.

One observes that the eigenvalues λ_+ with the unstabilized polynomial P_3 are clearly larger than 4 if the polynomial is equal or too close to 4. With a sufficiently large stabilization, i.e., $\eta > 0$ large enough, the eigenvalues are bounded away from 4. The price to pay is a (slightly) smaller value for $\widehat{\beta}_\Psi$, yielding a (slightly) stronger step-size restriction. Moreover, we see that for the larger value of κ we need larger values of η to guarantee that $\lambda_+ \leq 4$ for all $\tau^2 \leq \widehat{\beta}_\Psi^2/r$.

5.7.2. Modified Fermi–Pasta–Ulam–Tsingou problem

As second example we consider the modification of the FPUT β-problem introduced in Section 2.2.1. Here, we choose a chain of $m+2 = 82$ mass points, where we again set $\mu_i = 1$ for all $i = 1, \dots, m$, leading to $\mathbf{M} = \mathbf{I}_m$. We further set

$$k_i = 108^2, \quad i = 1, 2, 3, 4, \qquad \text{and} \qquad k_i = 25^2, \quad i = 5, \dots, m+1,$$

as well as $\beta_i^* = 3$ for all springs. The starting values are given by $\mathbf{q}_0 = (q_{0,1}, \dots, q_{0,m})^T$ and $\dot{\mathbf{q}}_0 = (\dot{q}_{0,1}, \dots, \dot{q}_{0,m})^T$ with

$$q_{0,i} = \begin{cases} 1, & i = 6, \\ 0, & \text{else}, \end{cases} \qquad \text{and} \qquad \dot{q}_{0,i} = \begin{cases} 0.5, & i = 6, \\ 0, & \text{else}. \end{cases}$$

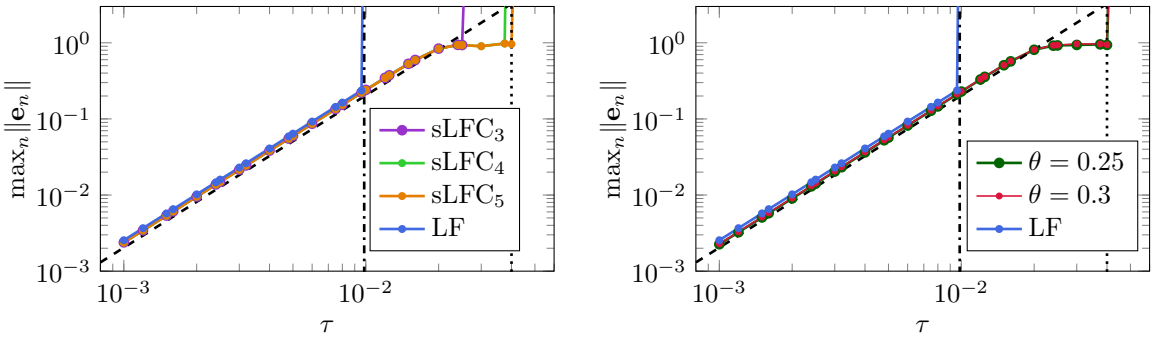
With the above choice of the spring constants k_i we set $\mathbf{S} = (\mathbf{L}_{i,j})_{i,j=1}^4$ and accordingly $\mathbf{N}$, $\mathbf{K}$ in (5.1). For these matrices we have $\|\mathbf{S}\| \approx 41231.51$, $\|\mathbf{N}\| \approx 2498.96$, and $\|\mathbf{K}\| = 625$. Note that $\|\mathbf{S}\| \approx \|\mathbf{L}\| \approx 41232.04$. In particular, we obtain $r \approx 16.50$ and $\kappa \approx 0.25$ for the constants in Assumption 5.1.

In Figure 5.4 we apply the leapfrog scheme (2.20) as well as the multirate scheme (5.4) to this FPUT β-problem with final time $T = 1.2$. We use the sLFC scheme (5.4), (4.1b) with $\eta = 0.5$ and $p = 3, 4, 5$, as well the split-θ-scheme (5.4), (3.4) to the FPUT β-problem. The reference solution for calculating errors is computed with the leapfrog scheme with step size $\tau = 10^{-5}$. In the error plot on the left we observe that with the sLFC scheme the maximum step size for which the scheme is stable is approximately p times larger than for the leapfrog scheme until $p = 4$. A further increase of the polynomial degree has almost no positive effect on the step size, since then the step-size restriction (5.17b) is the restricting one because of $r^{1/2} \approx 4.06$; cf. Section 5.6.2. Similarly, we see on the right plot that the split-θ-scheme allows for step sizes which are approximately a factor $r^{1/2}$ larger than the leapfrog scheme in accordance with our theoretical findings in Sections 5.5 and 5.6.2. Moreover, there is no visible difference in the error for these two choices of θ.

In Figure 5.5 we plot the relative error $\mathbf{err}_{\mathcal{H}}(n)$ of the Hamiltonian (2.11) with U given in (2.17), for the FPUT β-problem with the same data as above over time until $T = 100$ for two different step sizes. The relative error $\mathbf{err}_{\mathcal{H}}(n)$ is computed as in (4.22). Here, we apply the one-step formulations of the leapfrog scheme, the sLFC scheme (5.4), (4.1b), and its variant (5.5), (4.1b) with $p = 4$ and $\eta = 0.5$. Recall that in Section 5.2.2 we have shown that the one-step formulation of the multirate scheme (5.4) is not symplectic, in contrast to the one-step scheme belonging to (5.5). We observe in Figure 5.5a that the relative error of all three schemes is of the same magnitude over time. In particular, although not symplectic, the one-step scheme to the multirate scheme (5.4) nearly preserves the Hamiltonian for long times without having a visible drift. For the larger step size in Figure 5.5b we observe that the relative error of the two multirate schemes is larger compared to the ones for the smaller step size. This would also be the case for the leapfrog scheme, however, the step size already violates the step-size restriction of the leapfrog scheme; cf. Figure 5.4.

5.7.3. Spatially discretized acoustic wave equation

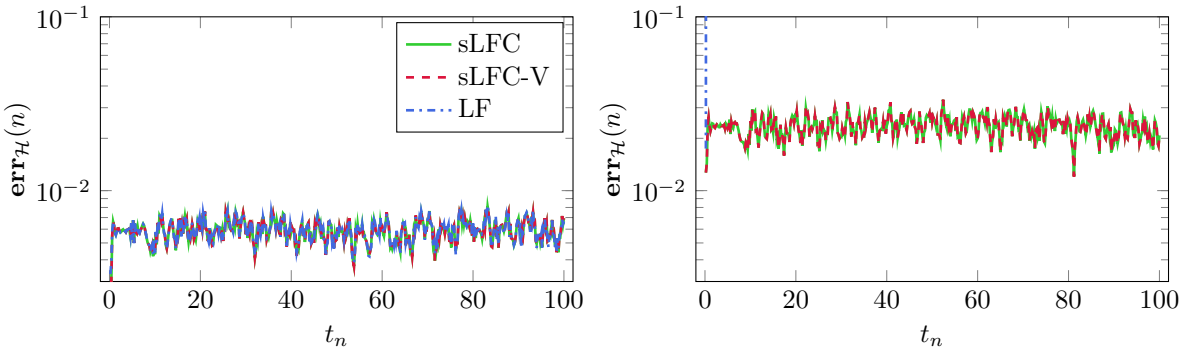
We conclude the numerical examples by considering the spatially discretized wave equation introduced in Section 2.2.2. Here, we focus on two examples for the linear, inhomogeneous



(a) Errors for sLFC schemes with $\eta = 0.5$.

(b) Errors for split-θ-schemes.

Figure 5.4.: Error for the numerical solution of the FPUT β-problem in Section 2.2.1 computed with the multirate scheme (5.4) up to $T = 1.2$. In the left plot we use $\Psi = P_p$ defined in (4.1b) with polynomial degree $p = 3$, $p = 4$, $p = 5$ and stabilization parameter $\eta = 0.5$. The blue line represents the leapfrog scheme. In the right plot we equip the scheme (5.4) with the θ-functions Ψ_θ defined in (3.4) with $\theta = 0.25$ and $\theta = 0.3$. The dashed lines indicate order two, the dash-dotted and dotted lines correspond to the maximum step sizes for which the leapfrog scheme (2.20) applied to the stiff system (2.1) and to the nonstiff problem (2.1) with $\mathbf{S} = \mathbf{K} = 0$, respectively, is stable.



(a) $\tau = 0.005$.

(b) $\tau = 0.01$.

Figure 5.5.: Relative error in Hamiltonian for the numerical solution of the FPUT β-problem computed with one-step formulations of the leapfrog scheme (blue, dash-dotted), the sLFC scheme (5.4), (4.1b) (green, solid), and its variant sLFC-V (5.5), (4.1b) (red, dashed) for two different step sizes τ. For the LFC polynomial P_p we use polynomial degree $p = 4$ and $\eta = 0.5$. The relative error of the Hamiltonian is only plotted at times $t = 0.2k$, $k = 1, \ldots, 500$, for the sake of clarity.

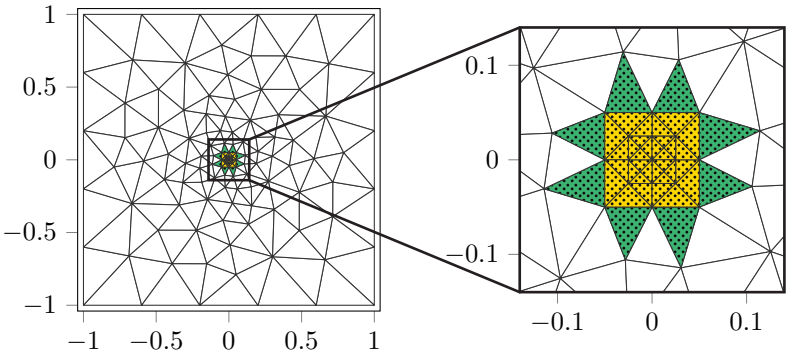


Figure 5.6.: Locally refined triangulation of $\Omega = [-1, 1]^2$.

problem (2.18), i.e., we have $g(t, x, q(t, x)) = g(t, x)$ for all $t > 0$ and $x \in \Omega$. In the first example a locally refined mesh is used and in the second one a heterogeneous material is considered.

Recall that we employ for the space discretization a (symmetric interior penalty) dG-FEM; see Section 2.2.2. We thus obtain the general problem (2.2) with a block-diagonal mass matrix $\mathbf{M}$, where each block belongs to a single mesh element. Hence, we are in the setting of Remark 5.3 and the error is measured in the weighted norm $\|\cdot\|_{\mathbf{M}}$. As in the examples in Section 4.5.3 we compute the error via (4.25) for a more representative illustration, because it contains not only errors of the time integration but also errors from the space discretization.

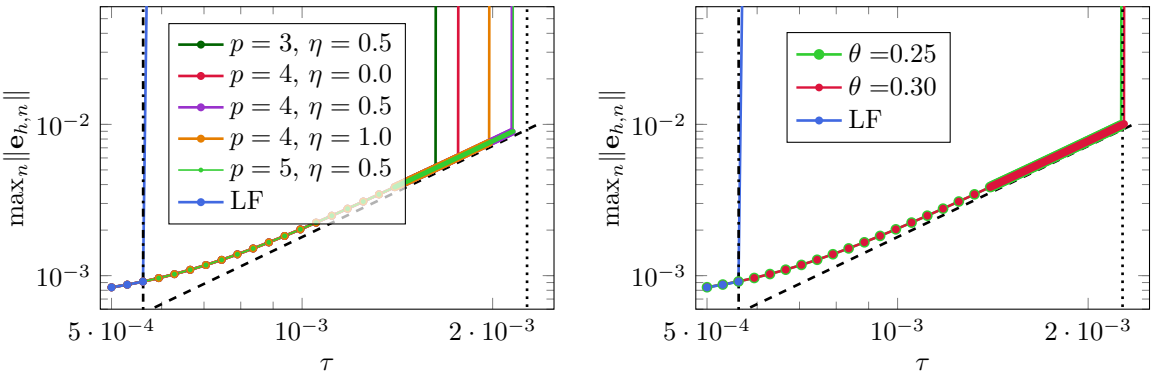
Locally refined mesh

For the first example we consider the problem (2.18) on the domain $\Omega = [-1, 1]^2$ with material parameter $c \equiv 1$. As exact solution of (2.18) we use (4.23) with $\delta = 2000$, from which the initial values and the inhomogeneity g are derived. For the dG-FEM we employ polynomials of degree four. Together with the large value of δ this ensures that the errors from the space discretization are dominated by errors of the time integration for large step sizes.

The triangulation of this simple domain is constructed in such a way that it contains a locally refined part; see the yellow area in Figure 5.6. Since $\|\mathbf{M}^{-1}\mathbf{L}\| \sim h_{\min}$, where $h_{\min}$ denotes the minimum diameter of all mesh elements, the main stiffness of the differential equation is induced from those degrees of freedom (dofs) belonging to the yellow mesh elements. However, also the adjacent mesh elements sharing an edge with yellow elements (the green ones in Figure 5.6) have to be taken into the stiff part, because the coupling between mesh elements in the dG-FEM is done via *flux terms*; see [DPE12] and [HW08] for more insight into the dG-FEM. Hence, after a possible reordering, $\mathbf{S}$ corresponds to the part of $\mathbf{L}$ with the dofs belonging to the dotted area in Figure 5.6. A numerical computation of the largest eigenvalues yields $r \approx 16.32$ and $\kappa \approx 0.513$ in Assumption 5.1.

In Figure 5.7 we apply the leapfrog scheme (2.20), sLFC scheme (5.4), (4.1b) and the split-θ-scheme (5.4), (3.4) to this problem up to $T = 5$. Because of $r^{1/2} \approx 4$ we use the parameters $p = 3, 4, 5$ with $\eta = 0.5$ and additionally $p = 4$ with $\eta = 0, 1$ for the sLFC scheme. For the split-θ-scheme we take the values $\theta = 0.25$ and $\theta = 0.3$.

We observe that all schemes show second-order convergence until the error of the space discretization is visible. Moreover, the errors for the leapfrog, sLFC, and split-θ-schemes are almost identical if the methods are stable. In accordance with Section 5.6.2, we further see that for the sLFC scheme (with sufficient stabilization) the polynomial degree $p = 4 \approx r^{1/2}$ is



(a) Errors for sLFC schemes.

(b) Errors for split-θ-schemes.

Figure 5.7.: Error for the numerical solution of the (spatially discretized) wave equation (2.18) with exact solution (4.23) ($\delta = 2000$) plotted against the step size for multirate leapfrog-type schemes (5.4) up to $T = 5$. In the left plot we use $\Psi = P_p$ defined in (4.1b)with values $(p, \eta) = (3, 0.5)$, $(4, 0)$ $(4, 0.5)$, $(4, 1)$, and $(5, 0.5)$. In the right plot we use $\Psi = \Psi_\theta$ defined in (3.4) with $\theta = 0.25$ and $\theta = 0.3$. The blue lines represents the leapfrog scheme. The dashed line indicates order two, the dash-dotted and dotted lines correspond to the maximum step sizes for which the leapfrog scheme (2.20) applied to the stiff system (2.2) and to the nonstiff problem (2.2) with $\mathbf{S} = \mathbf{K} = 0$, respectively, is stable.

optimal in terms of efficiency since a further increase of p barely enlarges the maximum step size, where the scheme is stable. For $p = 4$ we additionally observe that without ($\eta = 0$) or too much stabilization ($\eta = 1$) the largest step size for which the scheme is stable is significantly smaller than for $\eta = 0.5$. In contrast, the split-θ-schemes have (almost) the same step-size restrictions as the leapfrog scheme if applied to the nonstiff problem (2.2) with $\mathbf{S} = \mathbf{K} = 0$, confirming the theoretical results from Section 5.5.

Heterogeneous medium

In contrast to the previous examples, we consider this time the wave equation (2.18) on a domain with heterogeneous material, i.e., the material parameter c in (2.18) differs in value on the domain. More precisely, we set

$$c(x) = \begin{cases} 8.5, & x \in [0, 0.25] \times [0, 1], \\ 0.78, & x \in [0.25, 1] \times [0, 1]. \end{cases} \tag{5.57a}$$

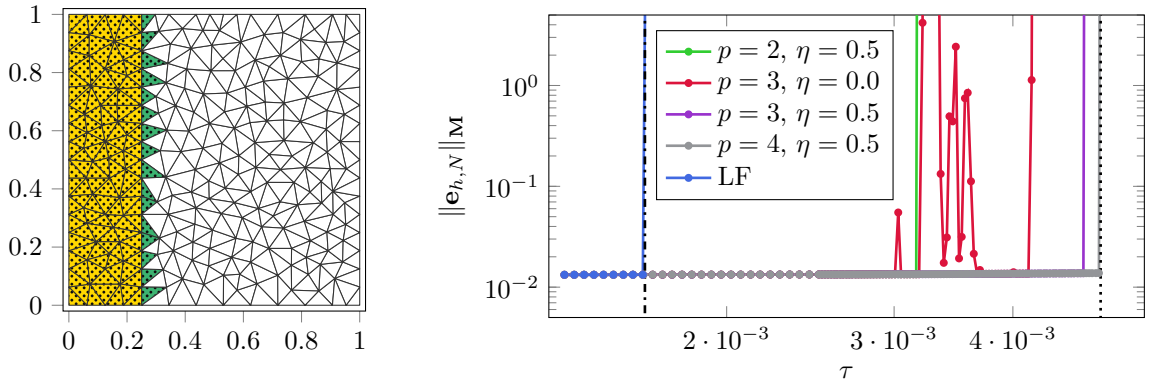
For the initial data and the inhomogeneity we choose the smooth functions

$$q_0(x) = h(x; 2, 0.25, \left(\begin{smallmatrix} 0.6 \\ 0.6 \end{smallmatrix}\right)), \quad \dot{q}_0(x) = 0, \qquad g(t, x) = h(x; 1, 0.1, \left(\begin{smallmatrix} 0.8 \\ 0.8 \end{smallmatrix}\right)) \cos(t), \tag{5.57b}$$

where

$$h(x; a, r_0, x_0) = \begin{cases} a \exp\Big(-(1 - \|x - x_0\|^2 / r_0^2)^{-1}\Big), & \|x - x_0\| \le r_0, \\ 0, & \text{otherwise.} \end{cases}$$

For the space discretization we use a mesh which matches the discontinuity of the material parameter c, i.e., the boundary of the subdomains for the different values of c are on edges of the mesh; cf. Figure 5.8a. For the dG-FEM we employ polynomials of degree two.



(a) Triangulation of $\Omega = [0,1]^2$.

(b) Error for $t = 2.7$.

Figure 5.8.: Mesh used for space discretization and error of the numerical solution of the (spatially discretized) wave equation (2.18), (5.57) computed with the sLFC scheme and the leapfrog scheme. For the polynomial P_p in the sLFC scheme we use as polynomial degrees p and stabilization parameters η the values $(p, \eta) = (2, 0.5)$, $(3, 0.0)$, $(3, 0.5)$, and $(4, 0.5)$. The blue lines represents the leapfrog scheme. The dash-dotted and dotted lines correspond to the maximum step sizes for which the leapfrog scheme (2.20) applied to the stiff system (2.2) and to the nonstiff problem (2.2) with $\mathbf{S} = \mathbf{K} = 0$, respectively, is stable.

Similar to the previous example, the stiffness of the differential equation is induced by those dofs belonging to the mesh elements, where c is large (yellow mesh elements in Figure 5.8a), and its adjacent elements, which share at least one edge (the green ones). Thus, after a possible reordering, $\mathbf{S}$ again corresponds to the part of $\mathbf{L}$ with the dofs belonging to the dotted area in Figure 5.8a. A numerical computation of the largest eigenvalues yields $r \approx 9.093$ and $\kappa \approx 0.269$ in Assumption 5.1.

In Figure 5.8b we apply the leapfrog scheme (2.20) and the sLFC scheme (5.4), (4.1b) with $p = 2, 3, 4$, $\eta = 0.5$ as well as $p = 3$, $\eta = 0.1$ to this equation. As final time we set $T = 2.7$. Since we do not know an exact solution to the above problem, a reference solution is computed on a twice refined mesh with the leapfrog scheme with step size $\tau = 5 \cdot 10^{-5}$. The errors are measured on the refined mesh at the final time.

We observe that the errors for the leapfrog scheme and the sLFC schemes are again almost identical if the methods are stable. Here, we cannot observe the second-order convergence in time because of the large errors of the space discretization. Moreover, we see that for $\eta = 0.5$ an increase of the polynomial degree p from 2 to 3 results in significantly larger step sizes, for which the sLFC scheme is stable, For $p = 4$ the gain is rather small in accordance with our theory in Section 5.6.2 which states that $p \approx r^{1/2} \approx 3$ is optimal in terms of efficiency. In addition, we once again see – in this example for $p = 3$ – that without stabilization ($\eta = 0$) the maximum step size for which the sLFC scheme is stable is drastically reduced compared to the stabilized variant.

Appendix A

Further calculations and properties of leapfrog-Chebyshev polynomials

In this appendix we present postponed but also further properties of the LFC polynomials (4.1b). The postponed results from Chapter 4 are shown in Sections A.1 and A.2, where some of the results are stated in a more general form than required. Section A.3 contains general formulae for the expanded forms of the stabilized and unstabilized LFC polynomials. The results in this appendix have been not published so far anywhere else, if not directly stated otherwise.

To prove the results in this appendix, we require some auxiliary results, which we show first.

Lemma A.1. *Let the function $h\colon [0,\infty) \to \mathbb{R}$ be convex and $h(0) = 0$. Then we have for all $y \geq 0$ and $p \in \mathbb{N}$*

$$ph(y) \leq h(py).$$

If h is concave instead of convex, the converse inequality holds true.

Proof. The inequalities are a direct consequence of the definition of convexity and concavity. □

With this lemma we immediately obtain that

$$p \sinh y \leq \sinh(py) \qquad \text{for all } p \in \mathbb{N},\ y \geq 0 \tag{A.1a}$$

and

$$p \tanh y \geq \tanh(py) \qquad \text{for all } p \in \mathbb{N},\ y \geq 0 \tag{A.1b}$$

because of the convexity of sinh and the concavity of tanh, respectively, for nonnegative numbers. Moreover, we have the following.

Lemma A.2. *The function $h\colon [0,\infty) \to \mathbb{R}$, defined by $h(y) = \frac{\sinh y}{\cosh(y)^{1/3}}$, is convex.*

Proof. The derivatives of h are given by

$$h'(y) = \frac{1 + \frac{2}{3}\sinh(y)^2}{\cosh(y)^{4/3}} \qquad \text{and} \qquad h''(y) = \frac{4}{9}\frac{\sinh(y)^3}{\cosh(y)^{7/3}}.$$

Since $h''(y) \geq 0$ for $y \geq 0$, h is convex. □

Combining both lemmas yields

$$p\frac{\sinh y}{\cosh(y)^{1/3}} \leq \frac{\sinh(py)}{\cosh(py)^{1/3}} \qquad \text{for all } p \in \mathbb{N},\ y \geq 0. \tag{A.2}$$

A.1. Calculations for general ν

We start by presenting further properties of the LFC polynomials (4.1b) in the case of the general stabilization parameter ν. We first prove the monotonicity of $\widetilde{\Upsilon}$ defined in (4.16) in a fixed interval which is required in the proof of Theorem 4.11. After that we show the monotonicity of the constants $\beta_{p,\nu}$, $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ in $\nu \geq 1$. To prove these results we make frequently use of the transformation

$$\varphi = \operatorname{arcosh} \nu \geq 0,\ \nu \geq 1 \qquad \Longleftrightarrow \qquad \nu = \cosh \varphi \geq 1,\ \varphi \geq 0. \tag{A.3}$$

We start with the auxiliary lemma for the proof of Theorem 4.11.

Lemma A.3. *Let $p \geq 2$ and $\nu > 1$. The rational functions $\widetilde{\Upsilon}$ defined in (4.16) are monotonically decreasing for $z \in [0, \sigma_{p,\nu}]$, where $\sigma_{p,\nu} = \alpha_p(\nu - 1)$.*

Proof. We show the monotonicity by proving that $\widetilde{\Upsilon}'(z) \leq 0$ for $z \in (0, \sigma_{p,\nu})$. Because of the continuity of the function it is then monotone also on the closed interval $[0, \sigma_{p,\nu}]$.

For $z > 0$ we have with the definition of P_p, α_p in (4.1b) and transformation (4.8)

$$\widetilde{\Upsilon}'(z) = \left(\widehat{\Psi}(z)^{-1}\Upsilon(z)\right)' = \frac{P_p'(z)}{P_p(z)^2} - \frac{1}{z^2} = \frac{T_p'(\nu)T_p'(x)}{\alpha_p^2\left(T_p(\nu) - T_p(x)\right)^2} - \frac{1}{\alpha_p^2(\nu - x)^2}. \tag{A.4}$$

Since $\nu > 1$ and $x \in (1, \nu)$ for $z \in (0, \sigma_{p,\nu})$, we can employ the transformations (A.3) (yielding $\varphi > 0$) and $x = \cosh \psi$ with $\psi \in (0, \varphi)$. Inserting this into (A.4) yields with (B.11), (B.12), and (B.7), (B.8)

$$\begin{aligned}
\widetilde{\Upsilon}'(z) &= \frac{1}{\alpha_p^2}\left(p^2 \frac{\sinh(p\varphi)\sinh(p\psi)}{\sinh\varphi\,\sinh\psi\left(\cosh(p\varphi) - \cosh(p\psi)\right)^2} - \frac{1}{\left(\cosh\varphi - \cosh\psi\right)^2}\right) \\
&= \frac{1}{\alpha_p^2 \sinh\varphi \sinh\psi}\left(p^2 \frac{\sinh(pa)^2 - \sinh(pb)^2}{4\sinh(pa)^2\sinh(pb)^2} - \frac{\sinh(a)^2 - \sinh(b)^2}{4\sinh(a)^2\sinh(b)^2}\right) \\
&= \frac{1}{4\alpha_p^2 \sinh\varphi \sinh\psi}\left(\frac{p^2}{\sinh(pb)^2} - \frac{1}{\sinh(b)^2} - \left(\frac{p^2}{\sinh(pa)^2} - \frac{1}{\sinh(a)^2}\right)\right),
\end{aligned}$$

where we abbreviate $a = \frac{1}{2}(\varphi + \psi)$ and $b = \frac{1}{2}(\varphi - \psi)$.

Since $\psi \in (0, \varphi)$, we have $a \in (\frac{\varphi}{2}, \varphi)$ and $b \in (0, \frac{\varphi}{2})$, hence $b < a$. Moreover, showing $\widetilde{\Upsilon}'(z) \leq 0$ for $z \in (0, \sigma_{p,\nu})$ is equivalent to proving that the function

$$f_p(y)\colon (0, \infty) \to \mathbb{R}, \qquad f_p(y) = \frac{p^2}{\sinh^2(py)} - \frac{1}{\sinh^2(y)},$$

is monotonically increasing (note that $\lim_{y\to 0} f_p(y) = \frac{1-p^2}{3}$). Taking the derivative of f_p yields together with (A.2)

$$f_p'(y) = 2\frac{\cosh y}{\sinh(y)^3} - 2p^3\frac{\cosh(py)}{\sinh(py)^3} \geq 2p^3\frac{\cosh(py)}{\sinh(py)^3} - 2p^3\frac{\cosh(py)}{\sinh(py)^3} = 0.$$

Thus, f_p is monotonically increasing, which concludes the proof. □

Next, we show that the stability constants $\beta_{p,\nu}$, $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ given in (4.7) and (4.14) are monotone functions in $\nu \geq 1$ (independent of the polynomial degree).

Lemma A.4. *Let $p \in \mathbb{N}$ and $\nu \geq 1$. Then $\beta_{p,\nu}$ defined in (4.7) and $\widehat{\beta}_{p,\nu}$ defined in (4.14b) are monotonically decreasing in ν.*

Proof. Using the transformation (A.3) we obtain with the definition of α_p in (4.1b) and the formula (B.11), (B.12) for T_p and its derivative that

$$\widehat{\beta}_{p,\nu}^2 = 2\frac{T_p'(\nu)}{T_p(\nu)}(\nu+1) = 2p\,\frac{\sinh(p\varphi)}{\sinh\varphi\cosh(p\varphi)}(\cosh\varphi+1) = 2p\left(\frac{\tanh(p\varphi)}{\tanh\varphi}+\frac{\tanh(p\varphi)}{\sinh\varphi}\right), \quad \text{(A.5)}$$

and, similarly,

$$\beta_{p,\nu}^2 = 4\frac{T_p'(\nu)}{T_p(\nu)}\nu = 4p\,\frac{\tanh(p\varphi)}{\tanh\varphi}.$$

To show that $\nu \mapsto \widehat{\beta}_{p,\nu}^2$ and $\nu \mapsto \beta_{p,\nu}^2$ are monotonically decreasing for $\nu \geq 1$, we prove that both of the functions $f, \tilde{f}\colon [0,\infty) \to \mathbb{R}$, defined by

$$f(\varphi) = \frac{\tanh(p\varphi)}{\tanh\varphi}, \qquad \tilde{f}(\varphi) = \frac{\tanh(p\varphi)}{\sinh\varphi}, \quad \text{(A.6)}$$

are monotonically decreasing. Differentiating the first function yields

$$f'(\varphi) = \frac{p\sinh\varphi\,\cosh\varphi - \sinh(p\varphi)\cosh(p\varphi)}{\tanh(\varphi)^2\cosh(\varphi)^2\cosh(p\varphi)^2}.$$

Using (A.1a) leads with the monotonicity of cosh then to $f'(\varphi) \leq 0$ for all $\varphi \geq 0$, showing the monotonicity of f. Analogously one can show that $\tilde{f}$ is monotonically decreasing. □

Lemma A.5. *Let $p \in \mathbb{N}$ and $\nu > 1$. Then $m_1^{p,\nu}$ and $\widetilde{m}_2^{p,\nu}$, given in (4.14a), are monotonically increasing and decreasing, respectively, in ν.*

Proof. The monotonicity of $m_1^{p,\nu}$ directly follows from the monotonicity of T_p for $x \geq 1$; see Lemma B.5. For $\widetilde{m}_2^{p,\nu}$ we have again with (A.3), (B.11), (B.12), and the identity (B.6)

$$\widetilde{m}_2^{p,\nu} = \frac{T_p(\nu)-1}{T_p'(\nu)(\nu-1)} = \frac{(\cosh(p\varphi)-1)\sinh\varphi}{p\sinh(p\varphi)(\cosh\varphi-1)} = \frac{\sinh(\frac{1}{2}p\varphi)^2\sinh\varphi}{p\sinh(p\varphi)\sinh(\frac{1}{2}\varphi)^2} = \frac{\tanh(\frac{1}{2}p\varphi)}{p\tanh(\frac{1}{2}\varphi)}. \quad \text{(A.7)}$$

This yields the monotonicity of $\widetilde{m}_2^{p,\nu}$, since we have shown in the previous proof that the function f given in (A.6) is monotonically decreasing. □

A.2. Calculations for special choice of ν

In this section we prove the bounds stated in Lemma 4.12 for the special choice of the stabilization parameter $\nu = \nu_{p,\eta}$ given in (4.3). To do so, we require lower and upper bounds on $T_p^{(k)}(\nu_{p,\eta})$, $k \in \mathbb{N}_0$, which we show first. A proof of the lower bound can also be found in [GMS21, Lemma A.1].

Lemma A.6. *Let $p \in \mathbb{N}$, $\eta \geq 0$, and $\nu_{p,\eta} = 1 + \eta^2/(2p^2)$. Then we have for $k \in \mathbb{N}_0$ and $x \geq 0$*

$$T_p^{(k)}(\nu_{p,\eta}) \geq T_p^{(k)}(1) + T_p^{(k+1)}(1)(\nu_{p,\eta} - 1) \tag{A.8}$$

and

$$\frac{1}{(2p^2)^k} T_p^{(k)}\Big(1 + \frac{x}{2p^2}\Big) = \frac{\mathrm{d}^k}{\mathrm{d}x^k} T_p\Big(1 + \frac{x}{2p^2}\Big) \leq \frac{\mathrm{d}^k}{\mathrm{d}x^k} \cosh(\sqrt{x}). \tag{A.9}$$

Proof. The first estimate follows from Taylor expansion and the monotonicity of the Chebyshev polynomials for $x \geq 1$; see Lemma B.5.

For the second estimate we have on the one hand for $x \geq 0$

$$\cosh(\sqrt{x}) = \sum_{j=0}^{\infty} \frac{1}{(2j)!} x^j.$$

On the other hand, we obtain from Lemma B.9

$$T_p\Big(1 + \frac{x}{2p^2}\Big) = \sum_{j=0}^{p} 2^j \, \frac{p\,(p+j-1)!}{(p-j)!} \frac{1}{(2j)!} \Big(\frac{x}{2p^2}\Big)^j = \sum_{j=0}^{p} \frac{p\,(p+j-1)!}{p^{2j}(p-j)!} \frac{1}{(2j)!} \, x^j = \sum_{j=0}^{p} a_j \, x^j,$$

where

$$a_j = \frac{b_{j,p}}{(2j)!}, \quad b_{j,p} = \frac{p\,(p+j-1)!}{p^{2j}\,(p-j)!} \qquad \text{for } j \in \mathbb{N}_0, j \leq p.$$

Since $b_{0,p} = 1$ for all $p \in \mathbb{N}$ and

$$0 \leq \frac{b_{j,p}}{b_{j-1,p}} = \frac{p\,(p+j-1)!}{p^{2j}\,(p-j)!} \frac{p^{2j-2}\,(p-j+1)!}{p\,(p+j-2)!} = \frac{(p+j-1)(p-j+1)}{p^2} = 1 - \frac{(j-1)^2}{p^2} \leq 1$$

for $1 \leq j \leq p$, we have that $0 \leq b_{j,p} \leq 1$ for all $j \leq p$, $p \in \mathbb{N}$. Hence, we obtain $0 < a_j \leq \frac{1}{(2j)!}$ for all $j \leq p$, which implies the second estimate. □

The bounds in Lemma 4.12 are proven in the subsequent lemma. Besides the lower bounds for $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ the following lemma also contains upper bounds for the constants. The lower bound for $m_1^{p,\nu}$ is also proven in [GMS21, Lemma A.4]; cf. Remark 4.13.

Lemma A.7. *Let $p \in \mathbb{N}$ and $\eta > 0$. For $\nu = \nu_{p,\eta} = 1 + \eta^2/(2p^2)$ we have*

$$\frac{(1 + \frac{1}{4}\eta^2)^{1/2}}{1 + \frac{1}{2}\eta^2} \leq \frac{\widehat{\beta}_{p,\nu}^2}{4p^2} \leq \frac{\tanh \eta}{\eta} (1 + \tfrac{1}{4}\eta^2)^{1/2} \leq 1 \tag{A.10a}$$

and

$$\frac{\eta^2}{4 + 2\eta^2} \leq m_1^{p,\nu} \leq \frac{1}{2} \frac{\cosh \eta - 1}{\cosh \eta}, \qquad (1 + \tfrac{1}{4}\eta^2)^{-1/2} \leq \widetilde{m}_2^{p,\nu} \leq 1, \tag{A.10b}$$

with $\widehat{\beta}_{p,\nu}$, $m_1^{p,\nu}$, and $\widetilde{m}_2^{p,\nu}$ defined in (4.14).

Proof. (i) We start with the bounds for $m_1^{p,\nu}$. From (A.8) with $k = 0$ we obtain the estimate $T_p(\nu_{p,\eta}) \geq 1 + \frac{1}{2}\eta^2$, which leads to the lower bound for $m_1^{p,\nu}$. For the upper bound we employ (A.9) with $k = 0$.

(ii) Next, we show the bounds for $\widehat{\beta}_{p,\nu}$. Inserting $\nu = \nu_{p,\eta}$ into the definition (4.14b) of $\widehat{\beta}_{p,\nu}$ yields with the transformation (A.3) similarly as in (A.5)

$$\widehat{\beta}_{p,\nu}^2 = 2\frac{T_p'(\nu)}{T_p(\nu)}(\nu+1) = 2p\Big(2+\frac{\eta^2}{2p^2}\Big)\frac{\tanh(p\varphi)}{\sinh\varphi} = 4p^2\Big(1+\frac{\eta^2}{4p^2}\Big)^{1/2}\frac{1}{\eta}\tanh(p\varphi), \tag{A.11}$$

where we used in the last step that for $\nu_{p,\eta} > 1$

$$\sinh\varphi = \sinh\big(\operatorname{arcosh}(\nu_{p,\eta})\big) = (\nu_{p,\eta}^2-1)^{1/2} = \frac{\eta}{p}\Big(1+\frac{\eta^2}{4p^2}\Big)^{1/2} \tag{A.12}$$

because of $\sinh(\operatorname{arcosh} y) = (y^2-1)^{1/2}$ for $y > 1$.

Further, since

$$2\operatorname{arsinh}(y) = \operatorname{arcosh}(2y^2+1) \qquad \text{for all } y \geq 0 \tag{A.13}$$

due to identity (B.6b), we obtain with the transformation (A.3)

$$p\varphi = p\operatorname{arcosh}(\nu_{p,\eta}) = 2p\operatorname{arsinh}\Big(\tfrac{\eta}{2p}\Big) = \eta\frac{\operatorname{arsinh}\big(\frac{\eta}{2p}\big)}{\frac{\eta}{2p}}. \tag{A.14}$$

The function $y \mapsto \operatorname{arsinh}(y)/y$ is for $y \geq 0$ monotonically increasing due to the monotonicity and concavity of arsinh for nonnegative numbers. Hence, we have by employing the monotonicity of tanh, again (A.13), and (A.12) with $p = 1$

$$\tanh(p\varphi) \geq \tanh\bigg(\eta\frac{\operatorname{arsinh}\big(\frac{1}{2}\eta\big)}{\frac{1}{2}\eta}\bigg) = \tanh\Big(\operatorname{arcosh}\big(1+\tfrac{1}{2}\eta^2\big)\Big) = \eta\frac{\big(1+\frac{1}{4}\eta^2\big)^{1/2}}{1+\frac{1}{2}\eta^2}.$$

Using this in (A.11) then leads to

$$\widehat{\beta}_{p,\nu}^2 \geq 4p^2\Big(1+\frac{\eta^2}{4p^2}\Big)^{1/2}\frac{\big(1+\frac{1}{4}\eta^2\big)^{1/2}}{1+\frac{1}{2}\eta^2} \geq 4p^2\frac{\big(1+\frac{1}{4}\eta^2\big)^{1/2}}{1+\frac{1}{2}\eta^2},$$

which shows the lower bound for $\widehat{\beta}_{p,\nu}^2$.

For the upper bound we get from (A.14) with $\lim_{y\to 0}\operatorname{arsinh}(y)/y = 1$ that

$$\tanh(p\varphi) \leq \tanh\eta,$$

which yields

$$\widehat{\beta}_{p,\nu}^2 \leq 4p^2\Big(1+\frac{\eta^2}{4p^2}\Big)^{1/2}\frac{\tanh\eta}{\eta} \leq 4p^2\big(1+\tfrac{1}{4}\eta^2\big)^{1/2}\frac{\tanh\eta}{\eta}.$$

It remains to show that for all $\eta \geq 0$

$$\frac{\tanh\eta}{\eta}\big(1+\tfrac{1}{4}\eta^2\big)^{1/2} \leq 1 \qquad \text{or, equivalently,} \qquad \tanh\eta \leq \frac{\eta}{\big(1+\frac{1}{4}\eta^2\big)^{1/2}}.$$

Obviously, for $\eta = 0$ the second bound is true. Further, by differentiating the functions $f_1\colon y \mapsto \tanh y$ and $f_2\colon y \mapsto y(1+\frac{1}{4}y^2)^{-1/2}$ one easily concludes that $0 \leq f_1'(y) \leq f_2'(y)$ for $y \geq 0$. Thus, the second bound holds for all $\eta \geq 0$.

(iii) Last, we turn towards the bounds for $\widetilde{m}_2^{p,\nu}$. The upper bound follows from (A.7) and (A.1b). For the lower bound we first observe that with (A.13) and $\cosh(\operatorname{arsinh} y) = (1+y^2)^{1/2}$ we have

$$\tanh(\tfrac{1}{2}\varphi) = \tanh(\operatorname{arsinh}(\tfrac{\eta}{2p})) = \tfrac{\eta}{2p}(1+\tfrac{\eta^2}{4p^2})^{-1/2}.$$

Moreover, similarly to the previous part of the proof we obtain

$$\tanh(\tfrac{1}{2}p\varphi) \geq \tanh(\operatorname{arsinh}(\tfrac{1}{2}\eta)) = \tfrac{1}{2}\eta(1+\tfrac{1}{4}\eta^2)^{-1/2}.$$

Hence, inserting these into formula (A.7) leads to

$$\widetilde{m}_2^{p,\nu} = \tanh(\tfrac{1}{2}p\varphi)\frac{2}{\eta}(1+\tfrac{\eta^2}{4p^2})^{1/2} \geq (1+\tfrac{1}{4}\eta^2)^{-1/2}(1+\tfrac{\eta^2}{4p^2})^{1/2} \geq (1+\tfrac{1}{4}\eta^2)^{-1/2},$$

which is the stated lower bound for $\widetilde{m}_2^{p,\nu}$. □

Numerical observations suggest that the bounds for $\widehat{\beta}_{p,\nu}/p^2$ and $\widetilde{m}_2^{p,\nu}$ in the previous lemma can be (slightly) improved, since the constants seem to be monotonically decreasing in p for every fixed $\eta > 0$; cf. Figure 4.4. Similar observations can be made for the constants $m_3^{p,\nu}$ and $m_4^{p,\nu}$ if we use $\nu = \nu_{p,\eta}$. Employing the (conjectured) monotonicity then leads to the following bounds.

Conjecture A.8. *Let $p \in \mathbb{N}$, $\eta \geq 0$, and $\nu = \nu_{p,\eta} = 1 + \eta^2/(2p^2)$.*
(a) *For $\eta > 0$ we have*

$$\frac{\tanh\eta}{\eta} \leq \frac{\widehat{\beta}_{p,\nu}^2}{4p^2} \leq \frac{1+\frac{1}{4}\eta^2}{1+\frac{1}{2}\eta^2}, \qquad \text{and} \qquad \frac{\tanh(\frac{1}{2}\eta)}{\frac{1}{2}\eta} \leq \widetilde{m}_2^{p,\nu} \leq 1,$$

where $\widehat{\beta}_{p,\nu}$ and $\widetilde{m}_2^{p,\nu}$ are defined in (4.14).
(b) *For $\eta \geq 0$ we have*

$$m_3^{p,\nu} \leq \frac{\cosh\eta\,(\cosh\eta - \sinh\eta/\eta)}{2\sinh(\eta)^2} \tag{A.15}$$

with $m_3^{p,\nu}$ defined in (4.9a).

Note that for $\eta = 0$ the singularity in the bound (A.15) is removable, since the function $\eta \mapsto \cosh\eta - \sinh\eta/\eta$ has a double root at $\eta = 0$. More precisely, we have

$$\lim_{\eta\to 0} \frac{\cosh\eta\,(\cosh\eta - \sinh\eta/\eta)}{2\sinh(\eta)^2} = \frac{1}{6},$$

which coincides with the limit of $m_3^{p,1}$ for $p \to \infty$; see (4.17).

A.3. Expanded forms of the leapfrog-Chebyshev polynomials

Concluding, we state formulae for the expanded forms of the LFC polynomials (4.1b) for $\nu \geq 1$ and $\nu = 1$, respectively. In the general case $\nu \geq 1$ the following holds.

Lemma A.9. *For every $p \in \mathbb{N}$ and $\nu \geq 1$ the LFC polynomials P_p defined in (4.1b) satisfy*

$$P_p(z) = \sum_{k=1}^{p} \frac{(-1)^{k+1}}{k!}\frac{T_p^{(k)}(\nu)}{T_p'(\nu)\alpha_p^{k-1}}\,z^k = z + \sum_{k=2}^{p} \frac{(-1)^{k+1}}{k!}\frac{T_p^{(k)}(\nu)}{T_p'(\nu)\alpha_p^{k-1}}\,z^k. \tag{A.16}$$

Proof. Taylor expansion of T_p at ν yields

$$T_p\Big(\nu - \frac{z}{\alpha_p}\Big) = \sum_{k=0}^{p} \frac{1}{k!} T_p^{(k)}(\nu)\Big(-\frac{z}{\alpha_p}\Big)^k = T_p(\nu) - \sum_{k=1}^{p} \frac{(-1)^{k+1}}{k!} \frac{T_p^{(k)}(\nu)}{\alpha_p^k} z^k.$$

Inserting this into the definition (4.1b) of the LFC polynomials then shows (A.16). □

For the unstabilized case, i.e., for $\nu = 1$, the expanded form of the LFC polynomials P_p can be expressed in a more concise formula which is implicitly already shown in [GMS18, Appendix A] (first equation on page 1019).

Lemma A.10. *Let $\nu = 1$. For every $p \in \mathbb{N}$ the LFC polynomials P_p defined in (4.1b) satisfy*

$$P_p(z) = \sum_{k=1}^{p} (-1)^{k+1} \frac{2}{(2k)!} \frac{(p+k-1)!}{(p-k)!\, p^{2k-1}} z^k = z + \sum_{k=2}^{p} (-1)^{k+1} \frac{2}{(2k)!} \frac{(p+k-1)!}{(p-k)!\, p^{2k-1}} z^k. \quad \text{(A.17)}$$

Proof. The claim directly follows from (A.16) together with (B.17) (note that $T_p'(1) = p^2$ and $\alpha_p = 2p^2$ for $\nu = 1$). Alternatively, one could employ (4.4) and (B.16) to obtain

$$P_p(z) = 2 - 2T_p\Big(1 - \frac{z}{2p^2}\Big) = 2 - 2\sum_{k=0}^{p} (-2)^k \frac{p(p+k-1)!}{(p-k)!} \frac{1}{(2k)!} \Big(\frac{z}{2p^2}\Big)^k,$$

which also shows (A.17). □

Using these lemmas we are able to explicitly write down the first LFC polynomials in the monomial basis. For instance, for the general case $\nu \geq 1$ the first four LFC polynomials are given by

$$\begin{aligned}
p = 1: &\quad P_1(z) = z, \\
p = 2: &\quad P_2(z) = z - \frac{2\nu^2 - 1}{16\nu^2} z^2, \\
p = 3: &\quad P_3(z) = z - \frac{2\nu(4\nu^3 - 3\nu)}{3(4\nu^2 - 1)^2} z^2 + \frac{(4\nu^3 - 3\nu)^2}{27(4\nu^2 - 1)^3} z^3, \\
p = 4: &\quad P_4(z) = z - \frac{(6\nu^2 - 1)(8\nu^4 - 8\nu^2 + 1)}{64(2\nu^3 - \nu)^2} z^2 + \frac{\nu(8\nu^4 - 8\nu^2 + 1)^2}{512(2\nu^3 - \nu)^3} z^3 - \frac{(8\nu^4 - 8\nu^2 + 1)^3}{65536(2\nu^3 - \nu)^4} z^4.
\end{aligned}$$

For $\nu = 1$ these simplify to

$$\begin{aligned}
p = 1: &\quad P_1(z) = z, \\
p = 2: &\quad P_2(z) = z - \tfrac{1}{16} z^2, \\
p = 3: &\quad P_3(z) = z - \tfrac{2}{27} z^2 + \tfrac{1}{729} z^3, \\
p = 4: &\quad P_4(z) = z - \tfrac{5}{64} z^2 + \tfrac{1}{512} z^3 - \tfrac{1}{65536} z^4.
\end{aligned}$$

Appendix B

Some basic and auxiliary results

In the following we present some basic, mainly known results which are used throughout this thesis. After stating various trigonometric and hyperbolic identities in the first section, we recall in Section B.2 the definition and some properties of the Chebyshev polynomials of the first and second kind, which we need especially in Chapter 4. We continue with a few basic properties of matrix functions which are required for the definition and the analysis of the general class of schemes in Chapter 3 and especially for the multirate schemes in Chapter 5. Concluding, we present in Section B.4 two discrete Gronwall-type lemmas, which are used in the error analysis of the semilinear problems, and their continuous counterparts.

B.1. Trigonometric and hyperbolic identities

In the following let $x, y \in \mathbb{R}$ unless explicitly stated otherwise. The (angle) addition formulae for sine and cosine are given by

$$\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y), \tag{B.1a}$$
$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y). \tag{B.1b}$$

A direct consequence of these addition formulae are the double-angle formulae

$$\sin(2x) = 2\sin(x)\cos(x), \tag{B.2a}$$
$$\cos(2x) = 1 - 2\sin(x)^2 = 2\cos(x)^2 - 1, \tag{B.2b}$$

from which one concludes the half-angle formulae

$$\sin(\tfrac{1}{2}x) = \Big(\tfrac{1}{2}\big(1 - \cos(x)\big)\Big)^{1/2}, \qquad x \in [0, 2\pi], \tag{B.3a}$$
$$\cos(\tfrac{1}{2}x) = \Big(\tfrac{1}{2}\big(1 + \cos(x)\big)\Big)^{1/2}, \qquad x \in [-\pi, \pi]. \tag{B.3b}$$

Further, from the addition formula (B.1) one deduces the sum-to-product formulae

$$\sin(x) \pm \sin(y) = \quad 2\sin\Big(\tfrac{1}{2}(x \pm y)\Big)\cos\Big(\tfrac{1}{2}(x \mp y)\Big), \tag{B.4a}$$
$$\cos(x) + \cos(y) = \quad 2\cos\Big(\tfrac{1}{2}(x + y)\Big)\cos\Big(\tfrac{1}{2}(x - y)\Big), \tag{B.4b}$$
$$\cos(x) - \cos(y) = -2\sin\Big(\tfrac{1}{2}(x + y)\Big)\sin\Big(\tfrac{1}{2}(x - y)\Big). \tag{B.4c}$$

In particular, (B.4a),(B.4b) imply by replacing x with $(n+1)x$ and y with $(n-1)y$ that

$$\sin\big((n+1)x\big) = 2\sin(nx)\cos(x) - \sin\big((n-1)x\big), \qquad n \in \mathbb{N}, \tag{B.5a}$$
$$\cos\big((n+1)x\big) = 2\cos(nx)\cos(x) - \cos\big((n-1)x\big), \qquad n \in \mathbb{N}. \tag{B.5b}$$

Similar formulae hold for the hyperbolic functions sinh and cosh. In particular, we have

$$\sinh(2x) = 2\sinh(x)\cosh(x), \tag{B.6a}$$
$$\cosh(2x) = 2\sinh(x)^2 + 1, \tag{B.6b}$$

and the sum-to-product formula for cosh

$$\cosh(x) - \cosh(y) = 2\sinh\Big(\tfrac{1}{2}(x+y)\Big)\sinh\Big(\tfrac{1}{2}(x-y)\Big). \tag{B.7}$$

From these two one deduces the following formula involving only sinh

$$\sinh x \sinh y = \sinh\Big(\tfrac{1}{2}(x+y)\Big)^2 - \sinh\Big(\tfrac{1}{2}(x-y)\Big)^2. \tag{B.8}$$

B.2. Chebyshev polynomials of the first and second kind

Chebyshev polynomials are named after *Pafnuty Lvovich Chebyshev* who introduced this class of polynomials in 1854 [Che54]. For further insight into Chebyshev polynomials we refer to the monographs [FP68, Riv90, MH03], in which all of the following results can be found.

Chebyshev polynomials of the first kind

We start with the Chebyshev polynomials of the first kind and state some of their properties. There exist several equivalent definitions of these polynomials, among others via a linear three-term recurrence relation, which we use here.

Definition B.1. *The Chebyshev polynomials of the first kind $T_p\colon \mathbb{R} \to \mathbb{R}$ are defined via the linear three-term recurrence relation*

$$T_{p+1}(x) = 2xT_p(x) - T_{p-1}(x), \qquad p \in \mathbb{N}, \tag{B.9}$$

where $T_1(x) = x$ and $T_0(x) = 1$.

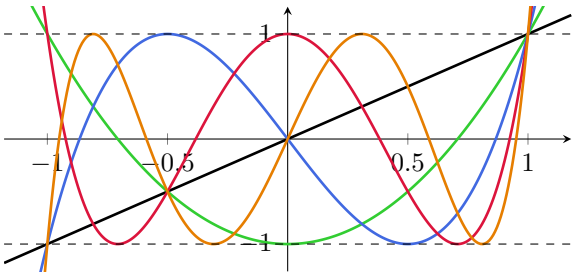
In Figure B.1 the Chebyshev polynomials T_p, $p = 1, \dots, 5$, are plotted and listed in their expanded form. From the definition one immediately obtains the following symmetry properties by induction.

Lemma B.2. *The Chebyshev polynomials T_p, $p \in \mathbb{N}_0$, are either even or odd according to the parity of p.*

Further basic properties of the polynomials can be seen more easily by another representation.

Lemma B.3. *The Chebyshev polynomials T_p, $p \in \mathbb{N}_0$, satisfy $T_p(\cos\psi) = \cos(p\psi)$ for $\psi \in \mathbb{R}$.*

Proof. The claim follows by induction together with the trigonometric identity (B.5b). □



$$T_1(x) = x$$
$$T_2(x) = 2x^2 - 1$$
$$T_3(x) = 4x^3 - 3x$$
$$T_4(x) = 8x^4 - 8x^2 + 1$$
$$T_5(x) = 16x^5 - 20x^3 + 5x$$

Figure B.1.: Chebyshev polynomials T_p of the first kind for polynomial degree $p = 1, \dots, 5$.

The lemma states that for $x \in [-1, 1]$ the Chebyshev polynomials are given by

$$T_p(x) = \cos(p \arccos(x)). \tag{B.10}$$

Often the Chebyshev polynomials of the first kind are defined via this formula. Similarly, one obtains for $x \geq 1$ that $T_p(x) = \cosh(p \operatorname{arcosh}(x))$ or, equivalently, with the transformation $x = \cosh(\varphi)$, $\varphi \geq 0$,

$$T_p(\cosh \varphi) = \cosh(p\varphi). \tag{B.11}$$

From these formulae we obtain for the derivatives of T_p

$$T_p'(\cos \psi) = p \frac{\sin(p\psi)}{\sin \psi} \quad \text{for } \psi \in \mathbb{R} \qquad \text{and} \qquad T_p'(\cosh \varphi) = p \frac{\sinh(p\varphi)}{\sinh \varphi} \quad \text{for } \varphi \geq 0. \tag{B.12}$$

Note that the singularities are removable.

With these representations of T_p one easily concludes many properties of the Chebyshev polynomials.

Lemma B.4. *Let $p \in \mathbb{N}$.*

(a) *For $x \in [-1, 1]$ we have $T_p(x) \in [-1, 1]$. Moreover, $T_p(1) = 1$.*

(b) *The roots of T_p are contained in $[-1, 1]$ and given by $x_i = \cos(\frac{\pi}{p}(i + \frac{1}{2}))$, $i = 0, \dots, p-1$.*

(c) *The (local) extrema of T_p are contained in $[-1, 1]$ and given by $x_i = \cos(\frac{\pi}{p} i)$, $i = 1, \dots, p-1$.*

Proof. The statements are a direct consequence of the representation (B.10). □

Lemma B.5. *The Chebyshev polynomials T_p, $p \in \mathbb{N}$, and its derivatives $T_p^{(k)}$, $k \in \mathbb{N}_0$, are monotonically increasing for $x \geq 1$ and, depending on the parity of $p + k$, either monotonically increasing or decreasing for $x \leq -1$. For $k \leq p - 1$ the monotonicity is strict.*

Proof. From part (b) in Lemma B.4 we get that T_p can be factorized in p linear factors. Since all roots are contained in $(-1, 1)$, this implies the result. □

Lemma B.6. *The Chebyshev polynomials T_p, $p \in \mathbb{N}_0$, satisfy for all $x \in \mathbb{R}$*

$$T_{pk}(x) = T_p(T_k(x)) = T_k(T_p(x)). \tag{B.13}$$

Proof. It suffices to show the result for $x \in [-1, 1]$. By using $x = \cos \psi$ we have with Lemma B.3 and (B.10)

$$T_p(T_k(x)) = T_p(\cos(k\psi)) = \cos(p \arccos(\cos(k\psi))) = \cos(pk\psi) = T_{pk}(x),$$

and analogously for $T_k(T_p(x))$. □

The great importance of Chebyshev polynomials in numerical analysis (interpolation, quadrature, best approximations to name some) relies on their remarkable properties to be extremal in some senses; see, for instance, [Riv90, Theorem 2.1] or [MH03, Corollary 3.4B] for the so-called *minimax* property. The following lemma represents a variant of this. Note that we have $T_p'(1) = p^2$ because of (B.12); see also (B.15) below.

Lemma B.7. *For all polynomials $P \not\equiv T_p$ of degree $p \in \mathbb{N}$ with $P(1) = 1$ and $P'(1) = p^2$ we have*

$$\max_{x\in[-1,1]} |P(x)| > 1 = \max_{x\in[-1,1]} |T_p(x)|.$$

Proof. Assume that there exists a polynomial $P_* \not\equiv T_p$ of degree $p \in \mathbb{N}$ with $P_*(1) = 1$ and $P_*'(0) = p^2$ satisfying $\max_{x\in[-1,1]}|P_*(x)| \le 1$. The polynomial $d = P_* - T_p$ is then again of degree p which has a double root at $x = 1$. Moreover, since T_p alternates p times between ± 1 in $[-1, 1]$, d has in each of the $p - 1$ intervals

$$\Big[\cos(\tfrac{k+1}{p}\pi), \cos(\tfrac{k}{p}\pi)\Big], \qquad k = 1, \dots, p-1,$$

at least one root. If one of these roots is at the boundary of such an interval (except of $x = -1$), it is a double root, because otherwise the condition $|P_*(x)| \le 1$ would be violated. Hence, counted with multiplicities there are $p - 1$ roots in $[-1, 1)$. Together with the double root at $x = 1$ this yields that d has $p + 1$ roots. Thus, $d \equiv 0$ which is in contradiction to $P_* \not\equiv T_p$. □

The next theorem is stated, for instance, in [GM99, Theorem 2.1 and 2.2] and plays an important role in approximation theory.

Theorem B.8 (Markov brothers' inequality)**.** *Let P be a polynomial of degree $p \in \mathbb{N}$. Then we have for $k \in \mathbb{N}_0$*

$$\max_{x\in[-1,1]} |P^{(k)}(x)| \le \max_{x\in[-1,1]} |T_p^{(k)}(x)| \max_{x\in[-1,1]} |P(x)|, \tag{B.14}$$

where equality only holds for $P = \pm T_p$. Moreover, it holds for all $k \in \mathbb{N}_0$

$$\max_{x\in[-1,1]} |T_p^{(k)}(x)| = T_p^{(k)}(1) = \prod_{j=0}^{k-1} \frac{p^2 - j^2}{2j+1}. \tag{B.15}$$

A proof of this theorem (in a slightly more general variant) is given in [Riv90, Theorem 2.24]; see also the original work [Mar90] for the case $k = 1$, and the German translation of the original work [MG16] for the general case $k \ge 1$.

Concluding we state another known explicit formula for the Chebyshev polynomials, which is a direct consequence of the previous theorem and the Taylor series of T_p at $x = 1$.

Lemma B.9. *The Chebyshev polynomials T_p satisfy for $p \in \mathbb{N}$*

$$T_p(x) = \sum_{k=0}^{p} (-2)^k \frac{p\,(p+k-1)!}{(p-k)!} \frac{1}{(2k)!} (1-x)^k. \tag{B.16}$$

Proof. From (B.15) we obtain for $k \le p$

$$T_p^{(k)}(1) = \prod_{j=0}^{k-1} \frac{(p+j)(p-j)}{2j+1} = \frac{p\,(p+k-1)!}{(p-k)!} \frac{2^k k!}{(2k)!}. \tag{B.17}$$

Thus, Taylor expansion of T_p at $x = 1$ yields

$$T_p(x) = \sum_{k=0}^{p} \frac{1}{k!} T_p^{(k)}(1)(x-1)^k = \sum_{k=0}^{p} (-1)^k \frac{p\,(p+k-1)!}{(p-k)!} \frac{2^k}{(2k)!} (1-x)^k,$$

which completes the proof. □

Chebyshev polynomials of the second kind

We now shortly present the Chebyshev polynomials of the second kind $U_p \colon \mathbb{R} \to \mathbb{R}$, $p \in \mathbb{N}_0$. Additionally, we consider some relations to T_p.

The Chebyshev polynomials of the second kind are defined via the same three-term recurrence relation as T_p but differ for U_1.

Definition B.10. *The Chebyshev polynomials of the second kind $U_p \colon \mathbb{R} \to \mathbb{R}$ are defined via the linear three-term recurrence relation*

$$U_{p+1}(x) = 2xU_p(x) - U_{p-1}(x), \qquad p \in \mathbb{N}, \tag{B.18}$$

where $U_1(x) = 2x$ and $U_0(x) = 1$.

As for the Chebyshev polynomials of the first kind there exist also a trigonometric representation for U_p. More precisely, for $p \in \mathbb{N}_0$ we have

$$U_p(\cos\psi)\sin\psi = \sin\big((p+1)\psi\big) \qquad \text{for all } \psi \in \mathbb{R}, \tag{B.19}$$

which can be shown similarly as (B.10) by using (B.5a) instead of (B.5b). In particular, with (B.12) this yields

$$T_p'(x) = p\,U_{p-1}(x) \qquad \text{for all } x \in \mathbb{R}, p \in \mathbb{N}. \tag{B.20}$$

Moreover, we have the following remarkable relation between T_p and U_p; see, e.g., [Riv90, equation (2.20)].

Lemma B.11. *The Chebyshev polynomials T_p, $p \in \mathbb{N}$, and U_p, $p \in \mathbb{N}_0$, satisfy for all $x \in \mathbb{R}$*

$$T_p(x)^2 - (x^2-1)U_{p-1}(x)^2 = 1 \qquad \textit{(Pell's equation).}$$

Proof. As before it is sufficient to show the equation for $x \in (-1, 1)$. Using $x = \cos\psi$ for $\psi \in (0, \pi)$ yields with Lemma B.3 and (B.19)

$$T_p(x)^2 - (x^2-1)U_{p-1}(x)^2 = \cos(p\psi)^2 + \sin(\psi)^2 \left(\frac{\sin(p\psi)}{\sin\psi}\right)^2 = \cos(p\psi)^2 + \sin(p\psi)^2 = 1,$$

which shows the claim. □

From this lemma we obtain together with (B.20)

$$p^2(1 - T_p(x)^2) = (1-x^2)T_p'(x)^2 \qquad \text{for all } x \in \mathbb{R}, p \in \mathbb{N}, \tag{B.21}$$

showing a relation between T_p and its derivative.

B.3. Basic properties of matrix functions

In this section we state some basic facts for matrix functions. For completeness we also recall one of several possible definition of matrix functions. More information about matrix functions can be found, e.g., in [Hig08], from which the following definitions and results are taken.

For the definition we first have to introduce some notation. We denote the Jordan normal form of a matrix $\mathbf{A} \in \mathbb{C}^{m\times m}$ by

$$\mathbf{J} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \operatorname{diag}(\mathbf{J}_1, \dots, \mathbf{J}_p), \qquad \mathbf{J}_k = \mathbf{J}(\lambda_k) = \begin{pmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{pmatrix} \in \mathbb{C}^{m_k\times m_k}, \quad \text{(B.22)}$$

where $\sum_{k=1}^{p} m_k = m$ and λ_k, $k = 1, \dots, p$, are the (not necessarily distinct) eigenvalues of $\mathbf{A}$. With this, a matrix function can be defined as follows; cf. [Hig08, Definitions 1.1 and 1.2].

Definition B.12. *Let $\mathbf{A} \in \mathbb{C}^{m\times m}$. The scalar function f is said to* be defined on the spectrum of the matrix $\mathbf{A}$ *if the values*

$$f^{(j)}(\lambda_k), \quad j = 0, \dots, m_k - 1,\ k = 1, \dots, p,$$

exist.

Definition B.13. *Let $\mathbf{A} \in \mathbb{C}^{m\times m}$ and the scalar function f be defined on the spectrum of $\mathbf{A}$. If $\mathbf{A}$ has the Jordan normal form (B.22), we define*

$$f(\mathbf{A}) = \mathbf{X} f(\mathbf{J}) \mathbf{X}^{-1} = \mathbf{X} \operatorname{diag}(f(\mathbf{J}_1), \dots, f(\mathbf{J}_p)) \mathbf{X}^{-1},$$

where

$$f(\mathbf{J}_k) = \begin{pmatrix} f(\lambda_k) & f'(\lambda_k) & \dots & \frac{1}{(m_k-1)!} f^{(m_k-1)}(\lambda_k) \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{pmatrix} \in \mathbb{C}^{m_k\times m_k}.$$

Clearly, if the matrix is diagonalizable, $f(\mathbf{J})$ is a diagonal matrix, since all Jordan blocks $\mathbf{J}_k$ are one-dimensional. Moreover, if f is real-valued and $\mathbf{A} \in \mathbb{R}^{m\times m}$ a real matrix, then $f(\mathbf{A})$ is in general a real matrix only if the eigenvalues and the (generalized) eigenvectors are real. This is, for instance, the case for real, symmetric matrices.

In the following we collect same basic results about matrix functions, which are used at some point in this thesis.

Lemma B.14 ([Hig08, Theorem 1.13])**.** *Let $\mathbf{A} \in \mathbb{C}^{m\times m}$ and let the scalar functions f be defined on the spectrum of $\mathbf{A}$.*

(a) *We have $f(\mathbf{A})^T = f(\mathbf{A}^T)$.*

(b) *If $\mathbf{B}$ commutes with $\mathbf{A}$, then $\mathbf{B}$ commutes with $f(\mathbf{A})$.*

(c) *If $\mathbf{A} = \operatorname{diag}(\mathbf{A}_{11}, \mathbf{A}_{22}, \dots, \mathbf{A}_{kk})$ is block diagonal ($k \le m$), we have*

$$f(\mathbf{A}) = \operatorname{diag}(f(\mathbf{A}_{11}), f(\mathbf{A}_{11}), \dots, f(\mathbf{A}_{kk})).$$

Lemma B.15 ([Hig08, Theorem 1.15]). *Let $\mathbf{A} \in \mathbb{R}^{m\times m}$ and let the scalar functions f and g be defined on the spectrum of $\mathbf{A}$.*
(a) *If $h(x) = f(x) + g(x)$ for all x in the spectrum of $\mathbf{A}$, we have $h(\mathbf{A}) = f(\mathbf{A}) + g(\mathbf{A})$.*
(b) *If $h(x) = f(x)g(x)$ for all x in the spectrum of $\mathbf{A}$, we have $h(\mathbf{A}) = f(\mathbf{A})g(\mathbf{A})$.*

Lemma B.16 ([Hig08, Corollary 1.34]). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$ and let the scalar function f be defined on the spectrum of both $\mathbf{AB}$ and $\mathbf{BA}$. Then we have*

$$\mathbf{A}f(\mathbf{BA}) = f(\mathbf{AB})\mathbf{A} \qquad \text{and} \qquad \mathbf{B}f(\mathbf{AB}) = f(\mathbf{BA})\mathbf{B}. \tag{B.23}$$

B.4. Gronwall-type lemmas

Last, we state discrete Gronwall-type lemmas (and their continuous counterparts), which are required in the error analysis of the semilinear problems in this thesis. We start with the following variant of the classical *Gronwall–Bellmann inequality* [Bel58].

Lemma B.17 (Gronwall-Bellmann inequality). *Let $T > 0$ and $\kappa, \varepsilon, \gamma\colon [0,T] \to \mathbb{R}$ continuous and nonnegative functions. If additionally κ is monotonically increasing, then*

$$\varepsilon(t) \le \kappa(t) + \int_0^t \gamma(s)\varepsilon(s)\,\mathrm{d}s \qquad \text{for all } t \in [0,T]$$

implies

$$\varepsilon(t) \le \kappa(t)\,\exp\!\left(\int_0^t \gamma(s)\,\mathrm{d}s\right) \qquad \text{for all } t \in [0,T].$$

A proof of this lemma can be found, e.g., in [Qin16, Theorem 1.1.4], [Pac98, Theorem 1.3.1], or the original work [Bel58]. A discrete, slightly more restrictive variant to this lemma is stated next; see, e.g., [Qin16, Theorem 2.1.2] or [Lee59].

Lemma B.18. *Let $\tau, \gamma \ge 0$. Further, let $\{\kappa_n\}_{n\ge 0}$ be a nonnegative, monotonically increasing sequence. If the nonnegative sequence $\{\varepsilon_n\}_{n\ge 0}$ satisfies*

$$\varepsilon_n \le \kappa_n + \gamma\tau \sum_{\ell=1}^{n-1} \varepsilon_\ell \qquad \text{for } n \ge 1,$$

then

$$\varepsilon_n \le \kappa_n\,\mathrm{e}^{\gamma\tau n} \qquad \text{for } n \ge 1.$$

The next lemma is shown in [CHS20, Lemma 3.8]. Since we are not aware of a proof of this result so far in the literature, we present it here in detail. More general cases but with worse estimates are, for instance, given in [Qin16, Theorem 2.1.12 and Theorem 2.1.44], or in [DM84].

Lemma B.19. *Let $\tau, \kappa, \gamma \ge 0$. If the nonnegative sequence $\{\varepsilon_n\}_{n\ge 0}$ satisfies*

$$\varepsilon_n \le \kappa + (\gamma\tau)^2 \sum_{\ell=0}^{n-1} (n-\ell)\varepsilon_\ell \qquad \text{for } n \ge 0,$$

then

$$\varepsilon_n \le \kappa\,\mathrm{e}^{\gamma\tau n} \qquad \text{for } n \ge 0.$$

We emphasize that with the discrete Gronwall Lemma B.18 we could also derive a bound for ε_n. Estimating $\tau(n-\ell) \le \mathrm{e}^{\tau(n-\ell)}$ and applying Lemma B.18 to $\mathrm{e}^{-\tau n}\varepsilon_n$ instead of ε_n yields $\varepsilon_n \le \kappa\,\mathrm{e}^{(\gamma^2+1)\tau n}$. However, as for the estimates in [Qin16, Theorem 2.1.12 and Theorem 2.1.44] we get a worse bound than with Lemma B.19.

Proof. For $\tau = 0$ or $\gamma = 0$ the result follows immediately. Hence, we assume $\tau, \gamma > 0$ in the following. Let $\{\rho_n\}_{n\ge 0}$ be defined by

$$\rho_n = \kappa + (\gamma\tau)^2 \sum_{\ell=0}^{n-1} (n-\ell)\rho_\ell.$$

By induction we obviously obtain $\varepsilon_n \le \rho_n$ for all $n \ge 0$. Further, we obtain for $n \ge 1$

$$\begin{aligned}
\rho_{n+1} - 2\rho_n + \rho_{n-1} &= (\gamma\tau)^2 \Bigg(\sum_{\ell=0}^{n} (n+1-\ell)\rho_\ell - 2\sum_{\ell=0}^{n-1} (n-\ell)\rho_\ell + \sum_{\ell=0}^{n-2} (n-1-\ell)\rho_\ell \Bigg) \\
&= (\gamma\tau)^2 \Bigg((\rho_n + 2\rho_{n-1}) - 2\rho_{n-1} + \sum_{\ell=0}^{n-2} (n+1-\ell-2(n-\ell)+n-1-\ell)\rho_\ell \Bigg) \\
&= (\gamma\tau)^2 \rho_n.
\end{aligned}$$

Thus, ρ_n satisfies for $n \ge 1$ the linear recurrence relation

$$\rho_{n+1} - (2 + (\gamma\tau)^2)\rho_n + \rho_{n-1} = 0.$$

Resolving this relation yields for $n \ge 0$

$$\rho_n = c_+ \eta_+^n + c_- \eta_-^n, \qquad \eta_\pm = f_\pm(\gamma\tau), \qquad f_\pm(x) = 1 + \tfrac{1}{2}x^2 \pm \tfrac{1}{2}x(4+x^2)^{1/2}$$

with $c_\pm \in \mathbb{R}$ (note that we used here that $\tau, \gamma > 0$).

In order to determine $c_\pm$ we observe that

$$c_+ + c_- = \rho_0 = \kappa \qquad \text{and} \qquad c_+\eta_+ + c_-\eta_- = \rho_1 = (1 + (\gamma\tau)^2)\kappa.$$

A simple calculation shows that

$$c_\pm = \frac{1}{2}\Bigg(1 \pm \frac{\gamma\tau}{(4+(\gamma\tau)^2)^{1/2}}\Bigg)\kappa.$$

Since $x(4+x^2)^{-1/2} < 1$ for all $x \ge 0$, we have that $c_\pm$ are both nonnegative.

Next, we bound $\eta_\pm^n$ for $n \in \mathbb{N}$. By employing $(4+y)^{1/2} \le 2 + \frac{1}{4}y$ for $y \ge 0$ we obtain on the one hand for $x \ge 0$

$$0 < f_+(x) \le 1 + \tfrac{1}{2}x^2 + \tfrac{1}{2}x(2 + \tfrac{1}{4}x^2) = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{8}x^3 \le \mathrm{e}^x$$

and on the other hand

$$f_-(x) \le 1 + \tfrac{1}{2}x^2 - \tfrac{1}{2}x^2 = 1 \qquad \text{and} \qquad f_-(x) = \frac{f_-(x)f_+(x)}{f_+(x)} = \frac{1}{f_+(x)} > 0.$$

Thus, we have $\eta_+^n \le \mathrm{e}^{\gamma\tau}$ and $0 < \eta_-^n \le 1$.

Altogether, this implies

$$\rho_n = c_+\eta_+^n + c_-\eta_-^n \le c_+\mathrm{e}^{\gamma t_n} + c_- \le \kappa \mathrm{e}^{\gamma t_n}.$$

Using $\varepsilon_n \le \rho_n$ for all $n \ge 0$ completes the proof. □

The previous lemma can be easily extended to the case of a non-decreasing, positive sequence $\{\kappa_n\}_{n\geq 0}$ instead of a constant κ.

Corollary B.20. *Let $\tau, \gamma \geq 0$. Further, let $\{\kappa_n\}_{n\geq 0}$ be a positive, monotonically increasing sequence. If the nonnegative sequence $\{\varepsilon_n\}_{n\geq 0}$ satisfies*

$$\varepsilon_n \leq \kappa_n + (\gamma\tau)^2 \sum_{\ell=0}^{n-1} (n-\ell)\varepsilon_\ell \qquad \text{for } n \geq 0,$$

then

$$\varepsilon_n \leq \kappa_n \, \mathrm{e}^{\gamma\tau n} \qquad \text{for } n \geq 0.$$

Proof. If we define $a_n = \varepsilon_n/\kappa_n$, $n \in \mathbb{N}$, we obtain for $n \geq 0$ with the monotonicity of $\{\kappa_n\}_{n\geq 0}$

$$a_n \leq 1 + (\gamma\tau)^2 \sum_{\ell=0}^{n-1} (n-\ell) a_\ell \frac{\kappa_\ell}{\kappa_n} \leq 1 + (\gamma\tau)^2 \sum_{\ell=0}^{n-1} (n-\ell) a_\ell.$$

An application of the previous lemma completes the proof. □

For the sake of completeness we also present the continuous counterpart to Lemma B.19. As for the discrete case we are not aware of a proof of this theorem in the literature. A more general case are the Volterra-type integral inequalities given by [NS87], which, however, yield worse estimates for our special case; see also [Qin16, Theorem 1.2.39] or [Pac98, Theorem 1.4.2] for a slightly more general variant.

Lemma B.21. *Let $\kappa, \gamma \geq 0$ and $T > 0$. If the nonnegative, continuous function $\varepsilon\colon [0,T] \to \mathbb{R}$ satisfies*

$$\varepsilon(t) \leq \kappa + \gamma^2 \int_0^t (t-s)\, \varepsilon(s)\, \mathrm{d}s \qquad \text{for all } t \in [0,T],$$

then

$$\varepsilon(t) \leq \kappa \, \mathrm{e}^{\gamma t} \qquad \text{for all } t \in [0,T].$$

Proof. We assume $\gamma > 0$, since for $\gamma = 0$ the statement follows immediately. Let $\delta > 0$. We first observe that the function $\varphi\colon [0,T] \to \mathbb{R}$ which solves the integral equation

$$\varphi(t) = (\kappa + \delta) + \gamma^2 \int_0^t (t-s)\, \varphi(s)\, \mathrm{d}s, \qquad t \in [0,T]$$

is given by $\varphi(t) = \frac{1}{2}(\kappa+\delta)(\mathrm{e}^{\gamma t} + \mathrm{e}^{-\gamma t})$. This can be seen, for instance, by differentiating the integral equation twice which then yields the initial value problem

$$\varphi''(t) = \gamma^2 \varphi(t), \qquad \varphi(0) = \kappa, \quad \varphi'(0) = 0.$$

We further know for all $t \in [0,T]$ that

$$\varphi(t) - \varepsilon(t) \geq \delta + \gamma^2 \int_0^t (t-s)\, (\varphi(s) - \varepsilon(s))\, \mathrm{d}s > 0,$$

since $\varphi(0) - \varepsilon(0) \geq \delta > 0$. Hence, we obtain

$$\varepsilon(t) < \varphi(t) = \tfrac{1}{2}(\kappa+\delta)(\mathrm{e}^{\gamma t} + \mathrm{e}^{-\gamma t}) \leq (\kappa+\delta)\mathrm{e}^{\gamma t}.$$

Since $\delta > 0$ is arbitrary, we get the result by taking the limit $\delta \to 0$. □

With the same procedure as in Corollary B.20 one can extend this lemma to the case of a monotonically increasing, positive function $\kappa\colon [0,T] \to \mathbb{R}$ instead of a constant κ. We omit the details.

Bibliography

[AGS21] A. Abdulle, M.J. Grote, and G. Rosilho de Souza. Explicit stabilized multirate method for stiff differential equations, 2021, arXiv:2006.00744 [math.NA]. http://arxiv.org/abs/2006.00744.

[Alb69] A. Albert. Conditions for positive and nonnegative definiteness in terms of pseudoinverses. *SIAM J. Appl. Math.*, 1969, vol. 17, pp. 434–440. ISSN 0036-1399. http://dx.doi.org/10.1137/0117041.

[And79] J.F. Andrus. Numerical solution of systems of ordinary differential equations separated into subsystems. *SIAM J. Numer. Anal.*, 1979, vol. 16(4), pp. 605–611. ISSN 0036-1429. http://dx.doi.org/10.1137/0716045.

[Arn82] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 1982, vol. 19(4), pp. 742–760. ISSN 0036-1429. http://dx.doi.org/10.1137/0719052.

[BCT82] P. Brenner, M. Crouzeix, and V. Thomée. Single-step methods for inhomogeneous linear differential equations in Banach space. *RAIRO Anal. Numér.*, 1982, vol. 16(1), pp. 5–26. ISSN 0399-0516. www.numdam.org/item/M2AN_1982__16_1_5_0/.

[BDH21] S. Buchholz, B. Dörich, and M. Hochbruck. On averaged exponential integrators for semilinear wave equations with solutions of low-regularity. *Partial Differ. Equ. Appl.*, 2021, vol. 2(2), pp. Paper No. 23, 27. ISSN 2662-2963. http://dx.doi.org/10.1007/s42985-020-00045-9.

[Bel58] R. Bellman. Asymptotic series for the solutions of linear differential-difference equations. *Rend. Circ. Mat. Palermo (2)*, 1958, vol. 7, pp. 261–269. ISSN 0009-725X. http://dx.doi.org/10.1007/BF02849324.

[BGG+18] S. Buchholz, L. Gauckler, V. Grimm, M. Hochbruck, and T. Jahnke. Closing the gap between trigonometric integrators and splitting methods for highly oscillatory differential equations. *IMA J. Numer. Anal.*, 2018, vol. 38(1), pp. 57–74. ISSN 0272-4979. http://dx.doi.org/10.1093/imanum/drx007.

[BS93] J.J. Biesiadecki and R.D. Skeel. Dangers of multiple time step methods. *J. Comput. Phys.*, 1993, vol. 109(2), pp. 318–328. ISSN 0021-9991. http://dx.doi.org/10.1006/jcph.1993.1220.

[BV09] M.A. Botchev and J.G. Verwer. Numerical integration of damped Maxwell equations. *SIAM J. Sci. Comput.*, 2008/09, vol. 31(2), pp. 1322–1346. ISSN 1064-8275. http://dx.doi.org/10.1137/08072108X.

[CDW96] L.C. Cowsar, T.F. Dupont, and M.F. Wheeler. A priori estimates for mixed finite element approximations of second-order hyperbolic equations with absorbing boundary conditions. *SIAM J. Numer. Anal.*, 1996, vol. 33(2), pp. 492–504. ISSN 0036-1429. http://dx.doi.org/10.1137/0733026.

[CH21] C. Carle and M. Hochbruck. Error analysis of multirate leapfrog-type methods for second-order semilinear ODEs. CRC 1173 Preprint 2021/26, Karlsruhe Institute of Technology, 2021. Revised version accepted for publication in SIAM J. Numer. Anal. http://dx.doi.org/10.5445/IR/1000133957.

[Che54] P.L. Chebyshev. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mém. Acad. Sci. Pétersb.*, 1854, vol. 7, pp. 539–568. https://hat.net.technion.ac.il/files/2021/02/cheb11.pdf.

[CHS20] C. Carle, M. Hochbruck, and A. Sturm. On leapfrog-Chebyshev schemes. *SIAM J. Numer. Anal.*, 2020, vol. 58(4), pp. 2404–2433. ISSN 0036-1429. http://dx.doi.org/10.1137/18M1209453.

[CI17] J. Chabassier and S. Imperiale. Space/time convergence analysis of a class of conservative schemes for linear wave equations. *C. R. Math. Acad. Sci. Paris*, 2017, vol. 355(3), pp. 282–289. ISSN 1631-073X. http://dx.doi.org/10.1016/j.crma.2016.12.009.

[CS13] E.M. Constantinescu and A. Sandu. Extrapolated multirate methods for differential equations with multiple time scales. *J. Sci. Comput.*, 2013, vol. 56(1), pp. 28–44. ISSN 0885-7474. http://dx.doi.org/10.1007/s10915-012-9662-z.

[DG09] J. Diaz and M.J. Grote. Energy conserving explicit local time stepping for second-order wave equations. *SIAM J. Sci. Comput.*, 2009, vol. 31(3), pp. 1985–2014. ISSN 1064-8275. http://dx.doi.org/10.1137/070709414.

[DM84] J. Dixon and S. McKee. Repeated integral inequalities. *IMA J. Numer. Anal.*, 1984, vol. 4(1), pp. 99–107. ISSN 0272-4979. http://dx.doi.org/10.1093/imanum/4.1.99.

[DPE12] D.A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, *Mathématiques & Applications (Berlin) [Mathematics & Applications]*, vol. 69. Springer, Heidelberg, 2012, pp. xviii+384. ISBN 978-3-642-22979-4. http://dx.doi.org/10.1007/978-3-642-22980-0.

[Dry03] M. Dryja. On discontinuous Galerkin methods for elliptic problems with discontinuous coefficients. *Comput. Methods Appl. Math.*, 2003, vol. 3(1), pp. 76–85. ISSN 1609-4840. Dedicated to Raytcho Lazarov, http://dx.doi.org/10.2478/cmam-2003-0007.

[DV56] R. De Vogelaere. Methods of integration which preserve the contact transformation property of the Hamilton equations. Technical Report 4, Department of Mathematics, University of Notre Dame, Notre Dame, Indiana, 1956. https://curate.nd.edu/downloads/6682x34921d.

[EL97] C. Engstler and C. Lubich. Multirate extrapolation methods for differential equations with different time scales. *Computing*, 1997, vol. 58(2), pp. 173–185. ISSN 0010-485X. http://dx.doi.org/10.1007/BF02684438.

[FP68] L. Fox and I.B. Parker. *Chebyshev polynomials in numerical analysis*. Oxford University Press, London-New York-Toronto, Ont., 1968, pp. ix+205.

[FPU+55] E. Fermi, J. Pasta, S. Ulam, and M. Tsingou. Studies of the nonlinear problems. Tech. Rep. LA-1940, Los Alamos Scientific Lab., N. Mex., 1955. Later published in *E. Fermi: Collected Papers*, Vol. II, E. Segrè (Ed.), University of Chicago Press (1965), pp. 978-988; and *Nonlinear wave motion*, A. C. Newell (Ed.), Lect. Appl. Math. 15, AMS, Providence, RI (1974), pp. 143-156. http://dx.doi.org/10.2172/4376203.

[Gau61] W. Gautschi. Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math.*, 1961, vol. 3, pp. 381–397. ISSN 0029-599X. http://dx.doi.org/10.1007/BF01386037.

[Gau15] L. Gauckler. Error analysis of trigonometric integrators for semilinear wave equations. *SIAM J. Numer. Anal.*, 2015, vol. 53(2), pp. 1082–1106. ISSN 0036-1429. http://dx.doi.org/10.1137/140977217.

[GCC+05] I. Grattan-Guinness, R. Cooke, L. Corry, P. Crépel, and N. Guicciardini (Eds.). *Landmark writings in western mathematics 1640–1940*. Elsevier Science, Amsterdam, 2005, first ed., pp. xviii+1022. ISBN 0-444-50871-6. http://dx.doi.org/10.1016/B978-0-444-50871-3.X5080-3.

[GHW+91] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Molecular Simulation*, 1991, vol. 6(1-3), pp. 121–142. http://dx.doi.org/10.1080/08927029108022142.

[GJ08] J.C. Gilbert and P. Joly. Higher order time stepping for second order hyperbolic problems and optimal CFL conditions. In: *Partial differential equations*, *Comput. Methods Appl. Sci.*, vol. 16. Springer, Dordrecht, 2008, pp. 67–93. http://dx.doi.org/10.1007/978-1-4020-8758-5_4.

[GKR01] M. Günther, A. Kværnø, and P. Rentrop. Multirate partitioned Runge-Kutta methods. *BIT*, 2001, vol. 41(3), pp. 504–514. ISSN 0006-3835. http://dx.doi.org/10.1023/A:1021967112503.

[GM99] N.K. Govil and R.N. Mohapatra. Markov and Bernstein type inequalities for polynomials. *J. Inequal. Appl.*, 1999, vol. 3(4), pp. 349–387. ISSN 1025-5834. https://www.emis.de/journals/HOA/JIA/Volume3_4/156027.pdf.

[GMS18] M.J. Grote, M. Mehlin, and S.A. Sauter. Convergence analysis of energy conserving explicit local time-stepping methods for the wave equation. *SIAM J. Numer. Anal.*, 2018, vol. 56(2), pp. 994–1021. ISSN 0036-1429. http://dx.doi.org/10.1137/17M1121925.

[GMS21] M. Grote, S. Michel, and S. Sauter. Stabilized leapfrog based local time-stepping method for the wave equation. *Math. Comp.*, 2021, vol. 90(332), pp. 2603–2643. ISSN 0025-5718. http://dx.doi.org/10.1090/mcom/3650.

[GR93] M. Günther and P. Rentrop. Multirate ROW methods and latency of electric circuits. *Appl. Numer. Math.*, 1993, vol. 13(1-3), pp. 83–102. ISSN 0168-9274. Sixth Conference on the Numerical Treatment of Differential Equations (Halle, 1992), http://dx.doi.org/10.1016/0168-9274(93)90133-C.

[GS09] M.J. Grote and D. Schötzau. Optimal error estimates for the fully discrete interior penalty DG method for the wave equation. *J. Sci. Comput.*, 2009, vol. 40(1-3), pp. 257–272. ISSN 0885-7474. http://dx.doi.org/10.1007/s10915-008-9247-z.

[GS16] M. Günther and A. Sandu. Multirate generalized additive Runge Kutta methods. *Numer. Math.*, 2016, vol. 133(3), pp. 497–524. ISSN 0029-599X. http://dx.doi.org/10.1007/s00211-015-0756-z.

[GSS99] B. García-Archilla, J.M. Sanz-Serna, and R.D. Skeel. Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.*, 1999, vol. 20(3), pp. 930–963. ISSN 1064-8275. http://dx.doi.org/10.1137/S1064827596313851.

[GSS06] M.J. Grote, A. Schneebeli, and D. Schötzau. Discontinuous Galerkin finite element method for the wave equation. *SIAM J. Numer. Anal.*, 2006, vol. 44(6), pp. 2408–2431. ISSN 0036-1429. http://dx.doi.org/10.1137/05063194X.

[GW84] C.W. Gear and D.R. Wells. Multirate linear multistep methods. *BIT*, 1984, vol. 24(4), pp. 484–502. ISSN 0006-3835. http://dx.doi.org/10.1007/BF01934907.

[Hen62] P. Henrici. *Discrete variable methods in ordinary differential equations*. John Wiley & Sons, Inc., New York-London, 1962, pp. xi+407.

[HHW21] L. He, W. Han, and F. Wang. On a family of discontinuous Galerkin fully-discrete schemes for the wave equation. *Comput. Appl. Math.*, 2021, vol. 40(2), pp. 1–24, Paper No. 56. ISSN 2238-3603. http://dx.doi.org/10.1007/s40314-021-01423-8.

[Hig08] N.J. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, pp. xx+425. Theory and computation. ISBN 978-0-89871-646-7. http://dx.doi.org/10.1137/1.9780898717778.

[HL21] M. Hochbruck and J. Leibold. An implicit-explicit time discretization scheme for second-order semilinear wave equations with application to dynamic boundary conditions. *Numer. Math.*, 2021, vol. 147(4), pp. 869–899. ISSN 0029-599X. http://dx.doi.org/10.1007/s00211-021-01184-w.

[HLW03] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the Störmer-Verlet method. *Acta Numer.*, 2003, vol. 12, pp. 399–450. ISSN 0962-4929. http://dx.doi.org/10.1017/S0962492902000144.

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, *Springer Series in Computational Mathematics*, vol. 31. Springer-Verlag, Berlin, 2006, second ed., pp. xviii+644. Structure-preserving algorithms for ordinary differential equations. ISBN 3-540-30663-3; 978-3-540-30663-4. http://dx.doi.org/10.1007/3-540-30666-8.

[HNW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations I: Nonstiff problems*, *Springer Series in Computational Mathematics*, vol. 8. Springer-Verlag, Berlin, 1993, second ed. ISBN 3-540-56670-8. http://dx.doi.org/10.1007/978-3-540-78862-1.

[Hof76] E. Hofer. A partially implicit method for large stiff systems of ODEs with only few equations introducing small time-constants. *SIAM J. Numer. Anal.*, 1976, vol. 13(5), pp. 645–663. ISSN 0036-1429. http://dx.doi.org/10.1137/0713054.

[HS80] 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.*, 1980, vol. 60(10), pp. 479–485. ISSN 0044-2267. http://dx.doi.org/10.1002/zamm.19800601005.

[HS16] M. Hochbruck and A. Sturm. Error analysis of a second-order locally implicit method for linear Maxwell's equations. *SIAM J. Numer. Anal.*, 2016, vol. 54(5), pp. 3167–3191. ISSN 0036-1429. http://dx.doi.org/10.1137/15M1038037.

[HS18] M. Hochbruck and A. Sturm. On leap-frog-Chebyshev schemes. CRC 1173 Preprint 2018/17, Karlsruhe Institute of Technology, 2018. Outdated, see major revision CRC 1173 Preprint 2019/19. http://dx.doi.org/10.5445/IR/1000085527.

[HV03] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, *Springer Series in Computational Mathematics*, vol. 33. Springer-Verlag, Berlin, 2003, pp. x+471. ISBN 3-540-03440-4. http://dx.doi.org/10.1007/978-3-662-09017-6.

[HW08] J.S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, *Texts in Applied Mathematics*, vol. 54. Springer, New York, 2008, pp. xiv+500. Algorithms, analysis, and applications. ISBN 978-0-387-72065-4. http://dx.doi.org/10.1007/978-0-387-72067-8.

[HZ05] R.A. Horn and F. Zhang. Basic Properties of the Schur Complement. In: F. Zhang (Ed.), *The Schur complement and its applications*, *Numerical Methods and Algorithms*, vol. 4. Springer-Verlag, New York, 2005, pp. xvi+295. ISBN 0-387-24271-6. http://dx.doi.org/10.1007/b105056.

[JN81] R. Jeltsch and O. Nevanlinna. Stability of explicit time discretizations for solving initial value problems. *Numer. Math.*, 1981, vol. 37(1), pp. 61–91. ISSN 0029-599X. http://dx.doi.org/10.1007/BF01396187.

[Jol03] P. Joly. Variational methods for time-dependent wave propagation problems. In: *Topics in computational wave propagation*, *Lect. Notes Comput. Sci. Eng.*, vol. 31. Springer, Berlin, 2003, pp. 201–264. http://dx.doi.org/10.1007/978-3-642-55483-4_6.

[JR10] P. Joly and J. Rodríguez. Optimized higher order time discretization of second order hyperbolic problems: construction and numerical study. *J. Comput. Appl. Math.*, 2010, vol. 234(6), pp. 1953–1961. ISSN 0377-0427. http://dx.doi.org/10.1016/j.cam.2009.08.046.

[Kar11] S. Karaa. Finite element θ-schemes for the acoustic wave equation. *Adv. Appl. Math. Mech.*, 2011, vol. 3(2), pp. 181–203. ISSN 2070-0733. http://dx.doi.org/10.4208/aamm.10-m1018.

[Kar12] S. Karaa. Stability and convergence of fully discrete finite element schemes for the acoustic wave equation. *J. Appl. Math. Comput.*, 2012, vol. 40(1-2), pp. 659–682. ISSN 1598-5865. http://dx.doi.org/10.1007/s12190-012-0558-8.

[Kvæ00] A. Kværnø. Stability of multirate Runge-Kutta schemes. *Int. J. Differ. Equ. Appl.*, 2000, vol. 1A(1), pp. 97–105. ISSN 1311-2872. Tenth International Colloquium on Differential Equations (Plovdiv, 1999).

[Lee59] M. Lees. Approximate solutions of parabolic equations. *J. Soc. Indust. Appl. Math.*, 1959, vol. 7, pp. 167–183. ISSN 0368-4245. http://www.jstor.org/stable/2099090.

[LM15] B. Leimkuhler and C. Matthews. *Molecular dynamics*, *Interdisciplinary Applied Mathematics*, vol. 39. Springer, Cham, 2015, pp. xxii+443. With deterministic and stochastic numerical methods. ISBN 978-3-319-16374-1; 978-3-319-16375-8. http://dx.doi.org/10.1007/978-3-319-16375-8.

[LR04] B. Leimkuhler and S. Reich. *Simulating Hamiltonian dynamics*, *Cambridge Monographs on Applied and Computational Mathematics*, vol. 14. Cambridge University Press, Cambridge, 2004, pp. xvi+379. ISBN 0-521-77290-7. http://dx.doi.org/10.1017/CBO9780511614118.

[Lub83] C. Lubich. On the stability of linear multistep methods for Volterra convolution equations. *IMA J. Numer. Anal.*, 1983, vol. 3(4), pp. 439–465. ISSN 0272-4979. http://dx.doi.org/10.1093/imanum/3.4.439.

[Mar90] A.A. Markov. On a question of D. I. Mendeleev. *Zapiski Imp. Akad. Nauk*, 1890, vol. 62, pp. 1–24. English translation, https://hat.net.technion.ac.il/files/2021/02/markov4.pdf.

[MG16] V.A. Markov and J. Grossmann. Über Polynome, die in einem gegebenen Intervalle möglichst wenig von Null abweichen. *Math. Ann.*, 1916, vol. 77(2), pp. 213–258. ISSN 0025-5831. http://dx.doi.org/10.1007/BF01456902.

[MH03] J.C. Mason and D.C. Handscomb. *Chebyshev polynomials*. Chapman & Hall/CRC, Boca Raton, FL, 2003, pp. xiv+341. ISBN 0-8493-0355-9. https://doi.org/10.1201/9781420036114.

[MO17] K.R. Meyer and D.C. Offin. *Introduction to Hamiltonian dynamical systems and the N-body problem*, *Applied Mathematical Sciences*, vol. 90. Springer, Cham, 2017, third ed., pp. xiii + 384. ISBN 978-3-319-53690-3; 978-3-319-53691-0. http://dx.doi.org/10.1007/978-3-319-53691-0.

[NS87] J. Norbury and A.M. Stuart. Volterra integral equations and a new Gronwall inequality. I. The linear case. *Proc. Roy. Soc. Edinburgh Sect. A*, 1987, vol. 106(3-4), pp. 361–373. ISSN 0308-2105. http://dx.doi.org/10.1017/S0308210500018473.

[Pac98] B.G. Pachpatte. *Inequalities for differential and integral equations*, *Mathematics in Science and Engineering*, vol. 197. Academic Press, Inc., San Diego, CA, 1998, pp. x+611. ISBN 0-12-543430-8.

[Paz83] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*, *Applied Mathematical Sciences*, vol. 44. Springer-Verlag, New York, 1983, pp. viii+279. ISBN 0-387-90845-5. http://dx.doi.org/10.1007/978-1-4612-5561-1.

[PW10] J.W. Prüss and M. Wilke. *Gewöhnliche Differentialgleichungen und dynamische Systeme*. Grundstudium Mathematik. [Basic Study of Mathematics]. Birkhäuser/Springer Basel AG, Basel, 2010, pp. xvi+318. ISBN 978-3-0348-0001-3; 978-3-0348-0002-0. http://dx.doi.org/10.1007/978-3-0348-0002-0.

[Qin16] Y. Qin. *Integral and discrete inequalities and their applications. Vol. I*. Birkhäuser/Springer, [Cham], 2016, pp. xvi+989. Linear inequalities. ISBN 978-3-319-33300-7; 978-3-319-33301-4. http://dx.doi.org/10.1007/978-3-319-33304-5_8.

[Ric60] J.R. Rice. Split Runge-Kutta method for simultaneous equations. *J. Res. Nat. Bur. Standards Sect. B*, 1960, vol. 64B, pp. 151–170. ISSN 0022-4340. http://dx.doi.org/10.6028/jres.064B.018.

[Riv90] T.J. Rivlin. *Chebyshev polynomials*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, 1990, second ed., pp. xvi+249. From approximation theory to algebra and number theory. ISBN 0-471-62896-4.

[RLS+21] S. Roberts, J. Loffeld, A. Sarshar, C.S. Woodward, and A. Sandu. Implicit multirate GARK methods. *J. Sci. Comput.*, 2021, vol. 87(1), pp. Paper No. 4, 32. ISSN 0885-7474. http://dx.doi.org/10.1007/s10915-020-01400-z.

[SA89] S. Skelboe and P.U. Andersen. Stability properties of backward Euler multirate formulas. *SIAM J. Sci. Statist. Comput.*, 1989, vol. 10(5), pp. 1000–1009. ISSN 0196-5204. http://dx.doi.org/10.1137/0910059.

[SB87] G.R. Shubin and J.B. Bell. A modified equation approach to constructing fourth-order methods for acoustic wave propagation. *SIAM J. Sci. Statist. Comput.*, 1987, vol. 8(2), pp. 135–151. ISSN 0196-5204. http://dx.doi.org/10.1137/0908026.

[SC20] R.D. Skeel and J.L. Cieśliński. On the famous unpublished preprint "Methods of integration which preserve the contact transformation property of the Hamilton equations" by René De Vogelaere, 2020, arXiv:2003.12268 [math.NA]. http://arxiv.org/abs/2003.12268.

[Sch05] I. Schneider. Chapter 7 - Abraham de Moivre, The doctrine of chances (1718, 1738, 1756). In: Grattan-Guinness et al. [GCC+05], pp. 105 – 120. http://dx.doi.org/10.1016/B978-044450871-3/50088-7.

[Sho97] R.E. Showalter. *Monotone operators in Banach space and nonlinear partial differential equations*, *Mathematical Surveys and Monographs*, vol. 49. American Mathematical Society, Providence, RI, 1997, pp. xiv+278. ISBN 0-8218-0500-2. http://dx.doi.org/10.1090/surv/049.

[SHV07] V. Savcenco, W. Hundsdorfer, and J.G. Verwer. A multirate time stepping strategy for stiff ordinary differential equations. *BIT*, 2007, vol. 47(1), pp. 137–155. ISSN 0006-3835. http://dx.doi.org/10.1007/s10543-006-0095-7.

[Ske93] R.D. Skeel. Variable step size destabilizes the Störmer/leapfrog/Verlet method. *BIT*, 1993, vol. 33(1), pp. 172–175. ISSN 0006-3835. http://dx.doi.org/10.1007/BF01990352.

[SM10] V. Savcenco and R.M.M. Mattheij. Multirate numerical integration for stiff ODEs. In: *Progress in industrial mathematics at ECMI 2008*, *Math. Ind.*, vol. 15. Springer, Heidelberg, 2010, pp. 327–332. http://dx.doi.org/10.1007/978-3-642-12110-4_50.

[SRS19] A. Sarshar, S. Roberts, and A. Sandu. Design of high-order decoupled multirate GARK schemes. *SIAM J. Sci. Comput.*, 2019, vol. 41(2), pp. A816–A847. ISSN 1064-8275. http://dx.doi.org/10.1137/18M1182875.

[Sti05] S.M. Stigler. Chapter 24 - P.S. Laplace, Théorie analytique des probabilités, first edition (1812); Essai philosophique sur les probabilités, first edition (1814). In: Grattan-Guinness et al. [GCC+05], pp. 329 – 340. http://dx.doi.org/10.1016/B978-044450871-3/50105-4.

[Stu17] A. Sturm. *Locally Implicit Time Integration for Linear Maxwell's Equations*. Ph.D. thesis, Karlsruher Institut für Technologie (KIT), 2017. http://dx.doi.org/10.5445/IR/1000069341.

[TBM92] M. Tuckerman, B.J. Berne, and G.J. Martyna. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 1992, vol. 97(3), pp. 1990–2001. http://dx.doi.org/10.1063/1.463137.

[Tes12] G. Teschl. *Ordinary differential equations and dynamical systems*, *Graduate Studies in Mathematics*, vol. 140. American Mathematical Society, Providence, RI, 2012, pp. xii+356. ISBN 978-0-8218-8328-0. http://dx.doi.org/10.1090/gsm/140.

[Ver82] J.G. Verwer. A note on a Runge-Kutta-Chebyshev method. *Z. Angew. Math. Mech.*, 1982, vol. 62(10), pp. 561–563. ISSN 0044-2267. http://dx.doi.org/10.1002/zamm.19820621008.

[Ver11] J.G. Verwer. Component splitting for semi-discrete Maxwell equations. *BIT*, 2011, vol. 51(2), pp. 427–445. ISSN 0006-3835. http://dx.doi.org/10.1007/s10543-010-0296-y.

[VHS90] J.G. Verwer, W.H. Hundsdorfer, and B.P. Sommeijer. Convergence properties of the Runge-Kutta-Chebyshev method. *Numer. Math.*, 1990, vol. 57(2), pp. 157–178. ISSN 0029-599X. http://dx.doi.org/10.1007/BF01386405.