European Society of Computational Methods in Sciences and Engineering (ESCMSE)

Journal of Numerical Analysis, Industrial and Applied Mathematics (JNAIAM)
vol. 4, no. 1-2, 2009, pp. 41-64
ISSN 1790–8140

Spectral Approximation of Time Windows in the Solution of Dissipative Linear Differential Equations[1]

K. Burrage[2]

Oxford University Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom, and Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Z. Jackiewicz[3], **B.D. Welfert**[4]

Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona 85287-1804, USA

Received 14 February, 2009; accepted in revised form 9 March, 2009

Dedicated to John Butcher on the occasion of his 75th birthday

Abstract: We establish a relation between the length T of the integration window of a linear differential equation $x'+Ax = b$ and a spectral parameter s^*. This parameter is determined by comparing the exact solution $x(T)$ at the end of the integration window to the solution of a linear system obtained from the Laplace transform of the differential equation by freezing the system matrix. We propose a method to integrate the relation $s^* = s^*(T)$ into the determination of the interval of rapid convergence of waveform relaxation iterations. The method is illustrated with a few numerical examples.

Keywords: Linear differential systems, time window, spectral approximation, waveform relaxation.

Mathematics Subject Classification: 65L05.

1 Introduction

Consider the differential system

$$x'(t) + Ax(t) = b(t), \quad t \in [0, T], \quad x(0) = 0, \tag{1.1}$$

with solution

$$x(t) = \int_0^t e^{-(t-s)A} b(s) ds. \tag{1.2}$$

[1]Published electronically June 15, 2009
[2]E-mails: kevin.burrage@comlab.ox.ac.uk, k.burrage@imb.uq.edu.au
[3]Corresponding author. E-mail: jackiewi@math.la.asu.edu
[4]E-mail: welfert@asu.edu

The matrix $A \in \mathbb{C}^{n\times n}$ in (1.1) is constant and assumed to be positive definite, with eigenvalues λ_i such that $0 < \Re(\lambda_n) \leq \ldots \leq \Re(\lambda_1)$. The integral form of (1.1) is

$$x(t) + \mathcal{I}Ax(t) = \mathcal{I}b(t), \quad t \in [0, T], \tag{1.3}$$

where $\mathcal{I}$ is the (linear) integral operator defined by $\mathcal{I}u(t) = \int_0^t u(s)ds$.

The equivalent of (1.1) in the spectral domain is

$$(sI + A)X(s) = B(s), \tag{1.4}$$

where $X(s)$ and $B(s)$ denote the Laplace transforms $\mathcal{L}\{x(t)\}$ and $\mathcal{L}\{b(t)\}$ of $x(t) \in \mathbb{C}^n$ and $b(t) \in \mathbb{C}^n$, respectively. Fixing $s = s^*$ in the matrix $sI + A$ then yields $(s^*I + A)X(s) \simeq B(s)$, i.e., $(s^*I + A)x(t) \simeq b(t)$ back in the temporal domain. We thus define $y(t) \in \mathbb{R}^n$ by

$$(s^*I + A)y(t) = b(t), \tag{1.5}$$

so that

$$(s^*\mathcal{I} + \mathcal{I}A)y(t) = \mathcal{I}b(t). \tag{1.6}$$

Our goal is to determine s^* such that the solution $x(t)$ of (1.1) approximates, in some sense, the solution $y(t)$ of (1.5). In the context of waveform relaxation applied to (1.1) we thereby hope that if $M \simeq A$ is a preconditioning matrix for the system $Ax(t) = b(t)$ then $s^*I + M$ is a suitable preconditioning matrix for (1.5) and thus for (1.1).

A comparison of (1.3) and (1.6) shows that

$$(I + \mathcal{I}A)(x(t) - y(t)) + (I - s^*\mathcal{I})y(t) = 0$$

and, consequently, that $y(t)$ may be a reasonable approximation of $x(t)$ if $s^*\mathcal{I} \simeq I$. On small time windows $x(t)$ can be approximated by a constant x so that

$$s^*Tx \simeq s^*\mathcal{I}x(T) \simeq Ix(T) = x(T) \simeq x$$

yields

$$s^* \simeq \frac{1}{T}. \tag{1.7}$$

The approximation (1.7) is also consistent with large time windows estimates, at least when $b(t) = b$ is constant. Indeed, the steady state solution of (1.1) is then $x(\infty) = A^{-1}b$ which is equal to the solution $y(\infty) = y$ of (1.5) obtained when setting $s^* = 0$.

The estimate (1.7) was first suggested by Leimkuhler [9] for estimating windows of convergence in waveform relaxation methods applied to (1.1). He based his analysis on the size of spectral radius of the matrix $sI + A$ for $\Re(s) > s^*$. He noted that (1.7) is a simplification, a fact later confirmed, especially for larger time windows, by extensive numerical experiments conducted by Burrage et al. [1], [2]. Jackiewicz et al. [8] proposed instead an estimate of the form

$$s^* = \frac{C}{T} \tag{1.8}$$

for some constant C. They related C to the ϵ-contour of the pseudospectrum [12] of A, namely $C = -\ln\epsilon$, but determine the appropriate value of ϵ only numerically by comparing the pseudospectra of a discrete version of the integral operator defined by (1.2) and of the Laplace transform $(sI + A)^{-1}$ of its kernel e^{-tA}.

The relation (1.8) between the size of the time window and the spectral parameter s^* lays at the heart of a recent strategy developed by Burrage et al. [3] for accelerating the convergence of

waveform relaxation iterations. It is therefore important to make this relation more precise, and in particular to find out whether it can be extended to larger time windows.

In the following $T > 0$ is assumed to be fixed. Our strategy for determining $s^* = s^*(T)$ is based on the minimization of the norm $\|x(T) - y(T)\|_2$ of the difference between the solution $x(T)$ of (1.1) at the end of the interval of integration and the solution $y(T)$ of (1.5) (which depends on s^*); i.e., we seek s^* such that

$$\|x(T) - y(T)\|_2 \to \min . \tag{1.9}$$

We first start in Section 2 by considering specific right-hand sides $b(t)$ and provide a detailed analysis under the assumption that A is hermitian.

Section 3 deals with general right-hand sides. We show that the solution of the minimization problem (1.9) satisfies a nonlinear equation which is well-posed and guaranteed to have a solution for small enough time windows and we suggest a method to integrate the computation of $s^*(t)$ and the resulting window of fast convergence of waveform relaxation iterations for each $0 < t \leq T$ into the ODE solution process. We illustrate the results numerically in Section 4 using an example arising from a discretization of a one-dimensional boundary value problem for the convection-diffusion equation via the method of lines.

The relationship between s^* and T derived in this paper puts the determination of the size of the window of rapid convergence of waveform relaxation iterations on a solid theoretical and practical ground and it is in our opinion the most serious attempt up to date to address this important problem of great practical importance.

2 Minimization with a particular right-hand side

In this section we consider specific right-hand side functions of the form $b(t) = f(t)b$ with $b \in \mathbb{C}^n$, $b \neq 0$, and $f(t)$ a scalar function. We focus on two choices for $f(t)$: monomials t^k with $k \geq 0$ integer and (combinations of) exponentials e^{at}.

2.1 Monomial right-hand side

The choice $b(t) = t^k b$ with $k \geq 0$ integer and $b \in \mathbb{C}^n$, $b \neq 0$, is driven by the observation that $b(t) \simeq t^k \frac{1}{k!} b^{(k)}(0)$ for some $k \geq 0$ is a reasonable approximation of $b(t)$ on small windows, provided $b(t)$ is analytic around $t = 0$.

We shall denote by $R_{k,0}(z)$ and $R_{k,1}(z)$ the $(k,0)-$ and $(k,1)-$Padé approximations of e^z at $z = 0$, respectively [6, p. 48]. It is easy to verify (e.g. by induction) that the solution (1.2) of (1.1) at $t = T$ can then be expressed as

$$x(T) = \int_0^T e^{-(T-s)A} s^k b \, ds = \frac{T}{k+1} \, \varphi_k(-TA) \, b(T) \tag{2.1}$$

with φ_k defined by

$$\varphi_k(z) = (k+1)! \, z^{-(k+1)} \, (e^z - R_{k,0}(z)) \tag{2.2}$$

for $z \neq 0$ and $\varphi_k(0) = 1$ by continuity. The functions $\varphi_k(z)$, $k = 0, \ldots, 4$, are shown in Fig. 1 and shall play an important role in the remainder of this section.

Our first theorem shows that if $s^* = \frac{k+1}{T}$ the quantity $y(T)$ is exactly equal to the approximation of $x(T)$ obtained by replacing e^{-TA} in the expression of $\varphi_k(-TA)$ in (2.1) by a $(k,1)$-Padé approximation. This can for example be the result of solving (1.1) using an appropriate Runge-Kutta or linear multistep method.

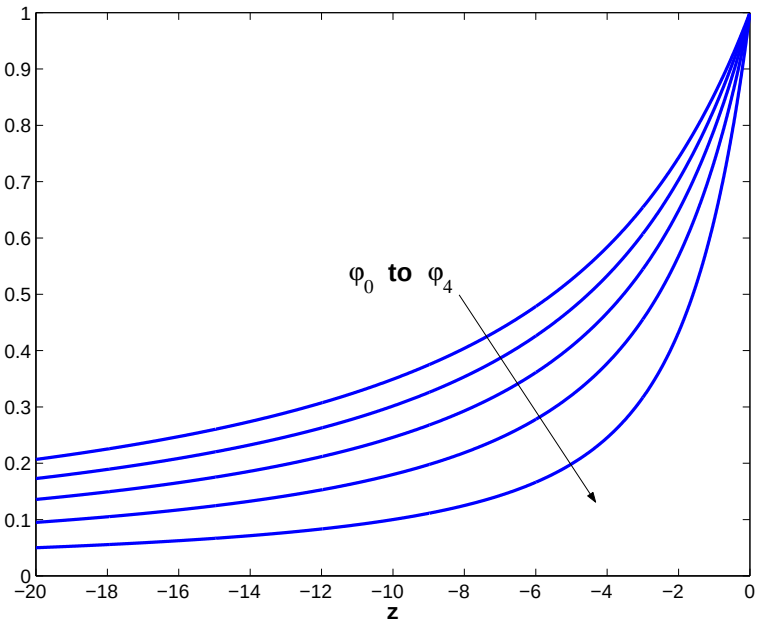

Figure 1: Functions $\varphi_k(z)$, $0 \le k \le 4$, for $-20 \le z \le 0$.

Theorem 2.1 *Assume that* $b(t) = t^k b$, $b \neq 0$, *and let* $s^* = \dfrac{k+1}{T}$. *Then*

$$y(T) = \frac{T}{k+1}\, \psi_k(-TA)\, b(T) \tag{2.3}$$

with

$$\psi_k(z) = (k+1)!\, z^{-(k+1)} \left(R_{k,1}(z) - R_{k,0}(z)\right). \tag{2.4}$$

Proof: From (2.4), Lemma 6.1 (see Appendix), and (1.5) the right-hand side of (2.3) reduces to

$$\begin{aligned} RHS &= \frac{T}{k+1}\,(k+1)!\,(-TA)^{-(k+1)}\frac{(-TA)^{k+1}}{(k+1)!}\left(I + \frac{TA}{k+1}\right)^{-1} b(T) \\ &= \left(\frac{k+1}{T} I + A\right)^{-1} b(T) \;=\; y(T) \end{aligned}$$

provided $s^* = \frac{k+1}{T}$. □

Theorem 2.1 shows that the choice $s^* = \frac{k+1}{T}$ is numerically acceptable when $b(t) = t^k b$. The following result then compares $y(T)$ to the exact solution (2.1) of (1.1) rather than an approximation. The quantity $\mu_2(A)$ represents the logarithmic norm of A with respect to the norm $\|\cdot\|_2$ [4, pp 18-19]. The positive definiteness of A and of its hermitian part $A_H = \frac{A+A^H}{2}$ implies in particular that

$$-\Re(\lambda_n) \le \mu_2(-A) = \max \lambda(-A_H) = -\min \lambda(A_H) < 0 \tag{2.5}$$

(see [4, p. 19]). Note that the leftmost inequality in (2.5) becomes an equality when A is normal.

Theorem 2.2 *Assume that* $b(t) = t^k b$, $b \neq 0$, *and let* $s^* = \dfrac{k+1}{T}$. *Then*

$$\frac{\|x(T) - y(T)\|_2}{\|y(T)\|_2} \leq \frac{\|A\|_2}{|\mu_2(-A)|} \; |\phi_k(T\mu_2(-A))| \tag{2.6}$$

with $\phi_k(z) = \frac{z}{k+1}\varphi_k'(z)$ *for* $z < 0$.

Proof: From (2.1), (1.5), and using Lemma 6.2(b) (see Appendix) we obtain

$$\begin{aligned} x(T) - y(T) &= \frac{T}{k+1}\, \varphi_k(-TA)\, b(T) - y(T) \\ &= \left(I + \frac{TA}{k+1}\right) \varphi_k(-TA)\, y(T) - y(T) \\ &= \frac{TA}{k+1}\, \varphi_k'(-TA)\, y(T). \end{aligned}$$

From Lemma 6.2(c) we have $\varphi_k'(z) = (k+1)! \int_\Omega t_k e^{t_k z}\; dt_k \ldots dt_0 > 0$, where Ω is the region $0 \leq t_k \leq \ldots \leq t_0 \leq 1$ of $\mathbb{R}^{k+1}$. Therefore

$$\begin{aligned} \|\varphi_k'(-TA)\|_2 &= (k+1)! \left\| \int_\Omega t_k e^{-t_k TA}\; dt_k \ldots dt_0 \right\|_2 \\ &\leq (k+1)! \int_\Omega t_k \|e^{-t_k TA}\|_2\; dt_k \ldots dt_0 \\ &\leq (k+1)! \int_\Omega t_k e^{t_k T \mu_2(-A)}\; dt_k \ldots dt_0 \\ &= \varphi_k'(T\mu_2(-A)). \end{aligned}$$

The last inequality follows from the fact that the logarithmic norm $\mu_2(A)$ provides the optimal exponential bound $\exp(\mu_2(A)t)$ for $\|\exp(At)\|_2$, see [4, p. 18]). The bound (2.6) then follows from

$$\|x(T) - y(T)\|_2 \leq \frac{T\|A\|_2}{k+1}\; \varphi_k'(T\mu_2(-A))\; \|y(T)\|_2$$

which completes the proof. □

The function $\phi_k(z)$ introduced in Theorem 2.2 is negative for $z < 0$. We can also verify using Lemma 6.2(a) and (b) that

$$\begin{aligned} \phi_k(z) &= \frac{z}{k+1}\varphi_k'(z) \;=\; \frac{z}{k+1}\left(\left(1 - \frac{k+1}{z}\right)\varphi_k(z) + \frac{k+1}{z}\right) \\ &= \frac{z}{k+1}\varphi_k(z) - \varphi_k(z) + 1 \;=\; \varphi_{k-1}(z) - \varphi_k(z) \end{aligned}$$

for $k \geq 0$. Since

$$\varphi_k(z) = 1 + \frac{z}{k+2} + \mathcal{O}(z^2) \quad \text{as} \quad z \to 0$$

and

$$\varphi_k(z) = -\frac{k+1}{z} + \mathcal{O}\left(\frac{1}{z^2}\right) \quad \text{as} \quad z \to -\infty$$

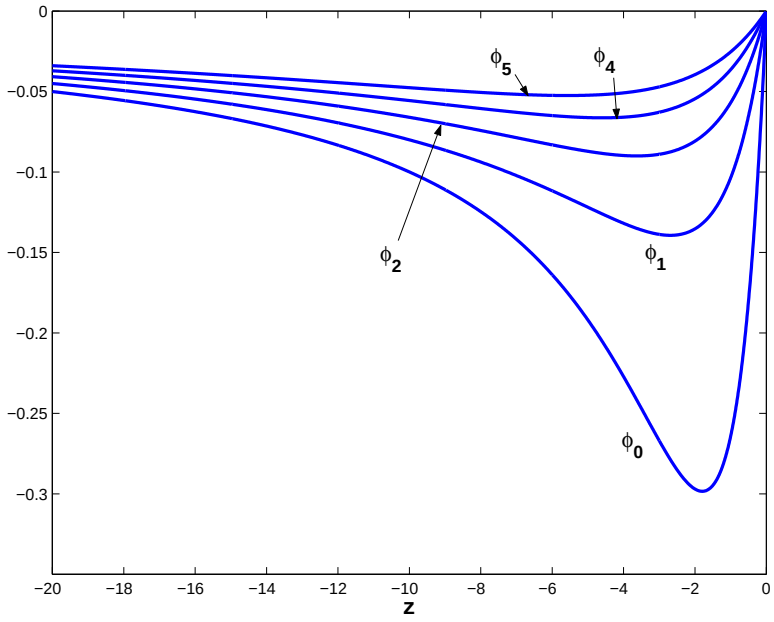

Figure 2: Functions $\phi_k(z)$ from Theorem 2.1, $0 \le k \le 4$, for $z \le 0$.

it follows that

$$\phi_k(z) \simeq \frac{z}{(k+1)(k+2)} \quad \text{as} \quad z \to 0$$

and

$$\phi_k(z) \simeq \frac{1}{z} \quad \text{as} \quad z \to -\infty.$$

Hence, the solution $y(T)$ of (1.5) is a first order approximation of $x(T)$ on short and long time windows, i.e.,

$$\frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2} = \begin{cases} \mathcal{O}(T) & \text{as } T \to 0, \\ \mathcal{O}(T^{-1}) & \text{as } T \to \infty. \end{cases}$$

The hermitian case

A finer analysis can be carried out when A is hermitian and positive definite. We denote by $A = U\Lambda U^H$ the Schur decomposition of A with $\Lambda = \mathrm{diag}(\lambda_i)_{1\le i\le n}$, $0 < \lambda_n \le \ldots \le \lambda_1$, and U a unitary matrix. From (2.1) and (1.5) we obtain

$$\begin{aligned} \|x(T)-y(T)\|_2^2 &= \left\| \frac{T}{k+1}\, \varphi_k(-TA)\, b(T) - (s^*I + A)^{-1} b(T) \right\|_2^2 \\ &= \left\| \frac{T}{k+1}\, \varphi_k(-T\Lambda)\, c - (s^*I + \Lambda)^{-1} c \right\|_2^2 \\ &= \sum_{i=1}^{n} \left| \frac{T}{k+1}\, \varphi_k(-T\lambda_i) - \frac{1}{s^*+\lambda_i} \right|^2 |c_i|^2 \end{aligned}$$

with $c = U^H b(T)$.

Theorem 2.3 *Assume that A is hermitian, positive definite, and that $T > 0$. For each $k \geq 0$ there exists a value μ_k with $0 < \lambda_n \leq \mu_k \leq \lambda_1$ and such that $\|x(T) - y(T)\|_2$ corresponding to $b(t) = t^k b$, $b \neq 0$, considered as a function of $s \geq 0$ is minimal for*

$$s_k^* = \frac{\varphi_{k-1}(-T\mu_k)}{\varphi_k(-T\mu_k)} \frac{k+1}{T}, \tag{2.7}$$

with the convention that $\varphi_{-1}(z) \equiv e^z$. Moreover, the following interlacing property holds:

$$0 < \frac{\lambda_n}{e^{T\lambda_n} - 1} \leq s_0^* \leq \frac{1}{T} \leq s_1^* \leq \frac{2}{T} \leq \ldots \leq \frac{k}{T} \leq s_k^* \leq \frac{k+1}{T}. \tag{2.8}$$

Proof: For fixed $k \geq 0$, $T > 0$ and $\mu > 0$ define the functions $f(\mu, s) = \frac{T}{k+1}\varphi_k(-T\mu) - \frac{1}{s+\mu}$ and $F(s) = \sum_{i=1}^n |f(\lambda_i, s)|^2 |c_i|^2$. The function F is defined and continuously differentiable for all $s \geq 0$. From Lemma 6.2(a) we have

$$f(\mu, 0) = -\frac{-T\mu\varphi_k(-T\mu) + k + 1}{(k+1)\mu} = -\frac{\varphi_{k-1}(-T\mu)}{\mu} < 0$$

and

$$f(\mu, \infty) = \frac{T}{k+1}\varphi_k(-T\mu) > 0.$$

Consequently, the function $F'(s) = 2\sum_{i=1}^n \frac{f(\lambda_i, s)}{(s+\lambda_i)^2}|c_i|^2$ admits at least one zero $s = s_k^* > 0$. Using Lemma 6.2(a) again, the relation $F'(s_k^*) = 0$ can be written as

$$\sum_{i=1}^n \frac{T(s_k^* + \lambda_i)\varphi_k(-T\lambda_i) - k - 1}{(k+1)(s_k^* + \lambda_i)^3}|c_i|^2 = \sum_{i=1}^n \alpha_i \left(s_k^* - \frac{\varphi_{k-1}(-T\lambda_i)}{\varphi_k(-T\lambda_i)} \frac{k+1}{T} \right) = 0$$

with $\alpha_i = \frac{T\varphi_k(-T\lambda_i)}{(k+1)(s_k^* + \lambda_i)^3}|c_i|^2 \geq 0$. Note that $\alpha_i > 0$ if $c_i \neq 0$ so that $b \neq 0$ implies $c \neq 0$ and $\sum_{i=1}^n \alpha_i > 0$. By Lemma 6.2(d) the function $\frac{\varphi_{k-1}(z)}{\varphi_k(z)}$ is non-decreasing for $z < 0$. Therefore s_k^* satisfies

$$\begin{aligned} \frac{\varphi_{k-1}(-T\lambda_1)}{\varphi_k(-T\lambda_1)} \frac{k+1}{T} &\leq s_k^* = \sum_{i=1}^n \left(\frac{\alpha_i}{\sum_{j=1}^n \alpha_j} \right) \frac{\varphi_{k-1}(-T\lambda_i)}{\varphi_k(-T\lambda_i)} \frac{k+1}{T} \\ &\leq \frac{\varphi_{k-1}(-T\lambda_n)}{\varphi_k(-T\lambda_n)} \frac{k+1}{T} \end{aligned} \tag{2.9}$$

so that (2.7) holds for some $\lambda_n \leq \mu_k \leq \lambda_1$. The fact that $F(s)$ reaches its minimum at s_k^* follows from

$$F'(0) = 2\sum_{i=1}^n \frac{T\lambda_i\varphi_k(-T\lambda_i) - k - 1}{(k+1)\lambda_i^3}|c_i|^2 = -2\sum_{i=1}^n \frac{\varphi_{k-1}(-T\lambda_i)}{(k+1)\lambda_i^3}|c_i|^2 < 0$$

and $F'(\infty) = 0^+$. Finally, the interlacing property (2.8) is a consequence of

$$\frac{\varphi_{k-1}(-\infty)}{\varphi_k(-\infty)} = \frac{k}{k+1} \quad \text{and} \quad \frac{\varphi_{k-1}(0)}{\varphi_k(0)} = 1.$$

For $k = 0$ we obtain

$$s_0^* = \frac{\mu_0}{e^{T\mu_0} - 1} = \frac{\varphi_{-1}(-T\mu_0)}{\varphi_0(-T\mu_0)} \frac{1}{T} \geq \frac{\varphi_{-1}(-T\lambda_n)}{\varphi_0(-T\lambda_n)} \frac{1}{T} = \frac{\lambda_n}{e^{T\lambda_n} - 1}.$$

This completes the proof. □

If $c_i = c_j$ then it is easy to verify that the weights α_i in the proof of Theorem 2.1 satisfy $0 \le \alpha_1 \le \ldots \le \alpha_n$ so that, in such a case,

$$\mu_k \simeq \lambda_n.$$

For *small time windows* T the value s_k^* given by (2.7) becomes

$$s_k^* \simeq \frac{k+1}{T} \tag{2.10}$$

for $k \ge 0$. On the other hand for *large time windows* T (e.g. $T\lambda_n \gg 1$) we obtain

$$s_k^* \simeq \frac{k}{T} \tag{2.11}$$

for $k > 0$ and

$$s_0^* \simeq \lambda_n e^{-T\lambda_n}. \tag{2.12}$$

Remark 1: If $c_{p+1} = \ldots = c_n = 0$ for some $0 < p < n$ (i.e., $c = U^H b(T)$ is ℓ^2-orthogonal to the first $n-p$ eigenvectors of A) then $\alpha_{p+1} = \ldots = \alpha_n = 0$ and $s_0^* \simeq \lambda_p\, e^{-T\lambda_p}$ instead for large T. Such a situation may occur in particular when b is randomly chosen. Indeed, such vector is typically "oscillatory" and likely to be orthogonal to the "smooth" eigenvectors of the matrix A associated to the smallest eigenvalues.

Note however that round-off errors in the numerical determination of the vector c may prevent c_n from vanishing exactly and still may yield, for larger values of T and in the case $k = 0$, $\alpha_n \gg \alpha_{n-1}$ if $c_n \approx c_{n-1}$.

Remark 2: Other strategies for finding an optimal s^* may be used. For example it is possible, for a given window $[0,T]$, to minimize the "average error" $\left\| \int_0^T (x(t) - y(t))dt \right\|$. If x_k and y_k denote the solutions of (1.1) and (1.5) when $b(t) = t^k b$, respectively, note however that

$$\begin{aligned}
\int_0^T (x_k(t) - y_k(t))dt &= \int_0^T \left(\int_0^t e^{-(t-s)A} s^k b\, ds - (s^* I + A)^{-1} t^k b \right) dt \\
&= \int_0^T \left(\frac{1}{k+1} \left(e^{-(t-s)A} s^{k+1} b \Big|_0^t - A \int_0^t e^{-(t-s)A} s^{k+1} b\, ds \right) - (s^* I + A)^{-1} t^k b \right) dt \\
&= \frac{1}{k+1} \left(\int_0^T \left(t^{k+1} b - A x_{k+1}(t) \right) dt - (s^* I + A)^{-1} T^{k+1} b \right) \\
&= \frac{1}{k+1} \left(\int_0^T x'_{k+1}(t)\, dt - y_{k+1}(T) \right) = \frac{1}{k+1} \left(x_{k+1}(T) - y_{k+1}(T) \right),
\end{aligned}$$

i.e., the optimization will lead to $s^* = s_{k+1}^*$ rather than $s^* = s_k^*$.

2.2 Exponential right-hand side

The choice of exponential right-hand sides $b(t) = e^{at} b$ with $a \in \mathbb{C}$ is essentially guided by the fact that many physical processes are driven by exponentially growing forcing terms.

We shall restrict ourselves to values of a such that the matrix $A+aI$ remains positive definite, i.e., in particular $\Re(a+\lambda_n) > 0$. From (1.2) we obtain

$$\begin{aligned} x(T) &= \int_0^T e^{-(T-s)A+asI} b\, ds = \int_0^T e^{-(T-s)(A+aI)} ds\, b(T) \\ &= T\varphi_0(-T(A+aI))\, b(T) \end{aligned} \tag{2.13}$$

which is exactly the solution (2.1) of (1.1) with $k=0$ and A replaced by $A+aI$. The expression (2.13) is now to be compared to

$$y(T) = (s^*I + A)^{-1} b(T) = ((s^*-a)I + A + aI)^{-1}\, b(T).$$

From Section 2.1 the optimal choice of s^* satisfies, in the hermitian case and with $a \in \mathbb{R}$ (so that $A+aI$ is also hermitian),

$$s^* - a = \frac{\varphi_{-1}(-T\mu)}{\varphi_0(-T\mu)} \frac{1}{T} = \frac{\mu}{e^{T\mu}-1} \tag{2.14}$$

for some $a+\lambda_n \le \mu \le a+\lambda_1$. On short time windows $\frac{\mu}{e^{T\mu}-1} \simeq \frac{1}{T}$ so that $s^* \simeq a + \frac{1}{T}$ while $s^* \to a$ exponentially fast as T increases, provided $a > 0$.

We now consider a right-hand side $b(t) = \sin(\omega t) b = \frac{1}{2i}\left(e^{i\omega t} - e^{-i\omega t}\right) b$ where $b \neq 0$ is a constant vector (e.g. a Fourier mode for general functions $b(t)$). The solution of (1.1) at $t=T$ is now

$$\begin{aligned} x(T) &= \frac{T}{2i}\left(\varphi_0(-T(A+i\omega I))e^{i\omega T} - \varphi_0(-T(A-i\omega I))e^{-i\omega T}\right) b \\ &= \frac{1}{2i}\left((A+i\omega I)^{-1}(e^{i\omega T} I - e^{-TA}) - (A-i\omega I)^{-1}(e^{-i\omega T} I - e^{-TA})\right) b \\ &= \frac{1}{2i}(A^2+\omega^2 I)^{-1}\left((A-i\omega I)(e^{i\omega T} I - e^{-TA}) - (A+i\omega I)(e^{-i\omega T} I - e^{-TA})\right) b \\ &= (A^2+\omega^2 I)^{-1}\left(\sin(\omega T) A - \omega\cos(\omega T) I + \omega e^{-TA}\right) b \end{aligned}$$

while the solution $y(T)$ of (1.5) is

$$y(T) = \sin(\omega T)\ (s^*I + A)^{-1}\, b.$$

Note that the optimization process to determine the optimal value of s^* breaks down when $\omega T = \pi$ since $y(T)$ then becomes independent of s^*. For values $T \ll \frac{\pi}{\omega}$ the approximation $b(t) \simeq \omega t$ holds and we expect the optimal choice of s^* to follow the recommendations from Section 2.1 with $k=1$.

3 Minimization with general right-hand side

We now consider a general right-hand side $b(t)$ and derive a nonlinear equation for the optimal value s^* obtained in the case $\|\cdot\| = \|\cdot\|_2$.

Let $A_{s^*} = s^*I + A$. We write

$$x(T) - y(T) = A_{s^*}^{-1}\left(s^* x(T) + Ax(T) - b(T)\right) = A_{s^*}^{-1}\left(s^* x(T) - x'(T)\right). \tag{3.1}$$

Since $x(T)$ does not depend on s^* we have also

$$\frac{d}{ds^*}(x(T) - y(T)) = -\frac{d}{ds^*} A_{s^*}^{-1} b(T) = A_{s^*}^{-2} b(T).$$

Therefore

$$\begin{aligned} \frac{d}{ds^*}\|x(T)-y(T)\|_2^2 &= 2\Re\left((x(T)-y(T))^H\frac{d}{ds^*}(x(T)-y(T))\right) \\ &= 2\Re\left(\left(s^*x(T)-x'(T)\right)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right) \end{aligned} \tag{3.2}$$

vanishes for

$$s^* = \frac{\Re\left(x'(T)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right)}{\Re\left(x(T)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right)}. \tag{3.3}$$

The expression (3.3) defines s^* as the solution of a nonlinear equation. The following result shows that for sufficiently small windows this equation admits at least one solution $s^* > 0$.

Theorem 3.1 *Assume that $b(t)$ is continuous on an interval $[0,T^+]$ for some $T^+ > 0$. Then there exists an interval $[0,T^-]$ with $0 < T^- \le T^+$ such that the equation (3.3) is well-posed and admits a solution $s^* > 0$ for all $0 < T \le T^-$.*

Proof: From (1.1) we have $x'(0) = b(0)$. We first assume that $b(0) \neq 0$ and show that both numerator and denominator on the right-hand side of (3.3) are positive for any $s^* > 0$ provided $T > 0$ is small enough. By the continuity of x' on $[0,T^+]$ and the positive definiteness of A_{s^*} and $A_{s^*}^{-1}$ for any $s^* > 0$ there exists $0 < T_1 \le T^+$ such that

$$\Re\left(x'(t)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right) = \Re\left(\left(\left(A_{s^*}^{-1}x'(t)\right)^H A_{s^*}^{-1}\left(A_{s^*}^{-1}b(T)\right)\right) > 0$$

for all t such that $0 \le t \le T \le T_1$. In particular

$$\Re\left(x'(T)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right) > 0$$

and

$$\Re\left(x(T)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right) = \int_0^T \Re\left(x'(t)^H A_{s^*}^{-H}A_{s^*}^{-2}b(T)\right)dt > 0$$

for all $T \in (0,T_1]$ and $s^* > 0$.

We next show that the minimum of $\|x(T)-y(T)\|_2$ is solution of (3.3). Similarly as (3.2) we obtain

$$\begin{aligned} \frac{d}{ds^*}\|x(0)-y(0)\|_2^2\Big|_{s^*=0} &= -2\Re\left(x'(0)^H A_0^{-H}A_0^{-2}b(0)\right) \\ &= -2\Re\left(\left(A^{-1}b(0)\right)^H A^{-1}\left(A^{-1}b(0)\right)\right) < 0 \end{aligned}$$

because A and A^{-1} are positive definite. By continuity there exists $T_2 > 0$ such that

$$\frac{d}{ds^*}\|x(T)-y(T)\|_2^2 < 0$$

for all $T \in [0,T_2]$.

Since $x(0) = 0$, $x'(0) \neq 0$ and x' is continuous on $[0,T^+]$ there also exists $0 < T_3 \le T^+$ such that $\frac{d}{dT}\|x(T)\|_2^2 > 0$ for all $T \in (0,T_2]$. Therefore

$$\Re\left(x(T)^H b(T)\right) = \frac{1}{2}\frac{d}{dT}\|x(T)\|_2^2 + \Re\left(x(T)^H Ax(T)\right) > 0$$

for all $T \in (0,T_3]$. Since $A_{s^*} \simeq s^*I$ as $s^* \to \infty$ we obtain

$$\frac{d}{ds^*}\|x(T)-y(T)\|_2^2\Big|_{s^*\to\infty} \simeq 2\,\frac{\Re\left(x(T)^H b(T)\right)}{(s^*)^2} > 0$$

for all $T \in (0, T_3]$. This shows that for any T such that $0 < T \leq T^- = \min(T_1, T_2, T_3)$ the quantity $\|x(T) - y(T)\|_2$ reaches its minimum at a critical point $0 < s^* < \infty$ which satisfies (3.3).

The result can also be shown to hold in the case $b(0) = 0$ using a continuity argument based on $b(t) \neq 0$ for any $t > 0$ sufficiently small. □

Observe that (3.3) can be interpreted as a weak form of the condition $y(T) - x(T) = 0$. Indeed, we obtain from (1.5), (3.1) and (3.3)

$$\Re\left((y(T) - x(T))^H A_{s^*}^{-1} y(T)\right) = \Re\left(\left(s^* x(T) - x'(T)\right)^H A_{s^*}^{-H} A_{s^*}^{-2} b(T)\right) = 0. \tag{3.4}$$

On long time windows the right-hand side of (3.3) may vanish or become negative, in particular when $b(T)$ itself vanishes as in the case of sinusoidal right-hand side (see Section 2.2). In this case $y(T)$ vanishes as well and (3.4) can no longer be used to determine s^*.

3.1 Short time window estimate

For right-hand sides $b(t)$ of the form

$$b(t) = f(t)b$$

where $f(t)$ is a continuous scalar function and $b \in \mathbb{R}^n$, $b \neq 0$, we obtain from (1.2)

$$x(T) = \int_0^T \left(I + \mathcal{O}(T - t)\right) f(t) b \, dt = \left(\int_0^T f(t) dt\right) \left(1 + \mathcal{O}(T)\right) b$$

and

$$x'(T) = f(T)b - Ax(T) = f(T)\left(1 + o(1)\right) b$$

for small time windows. Then (3.3) reduces to

$$s^* \simeq \frac{f(T)}{\int_0^T f(t) dt} \tag{3.5}$$

for small T independently of A (although the time interval on which (3.5) remains a good approximation does depend in general on A). In particular for $f(t) = t^\alpha$ with $\alpha > 0$ we obtain the estimate $s^* \simeq \frac{\alpha+1}{T}$, which agrees with the results of Section 2 for integer values of α. For $f(t) = e^{at}$ we also obtain $s^* \simeq \frac{a e^{aT}}{e^{aT} - 1}$ which is identical to (2.14) with $\mu = a$. The choice $f(t) = \sin(\omega t)$ yields $s^* \simeq \omega \cot\left(\frac{\omega T}{2}\right)$.

3.2 Numerical solution of (3.3)

From a practical point of view s^* can be computed from (3.3) for increasing values of T once $x(T)$ and $x'(T)$ have been determined (for example using an adaptive solver), until the right-hand side of (3.3) fails to remain positive. The nonlinear equation (3.3) is of the form $s^* = F(s^*)$. In all our numerical tests a Picard iteration $s^*_{p,m} = F(s^*_{p,m-1})$ for $m > 0$ starting with the solution of (3.3) obtained at the previous time step T (and with a predicted $s^*_{1,0} = \frac{1}{h_0}$ for the first step) was successfully used to solve (3.3) at time $T = h_0 + \ldots + h_{p-1}$, see Algorithm 1.

Algorithm 1: Adaptive determination of $s^* = s^*(T)$ for $T = \sum_{p=0}^{q-1} h_p$.

$x_0 = 0$, $h_0 > 0$ given, $t_0 = 0$, $s^*_{1,\text{predicted}} = \frac{1}{h_0}$
for $p = 1, 2, \ldots, q$
    compute h_{p-1} and $x_p \simeq x(t_{p-1} + h_{p-1})$ using an (adaptive) ODE solver

$t_p = t_{p-1} + h_{p-1}$
$x'_p = b(t_p) - Ax_p$
$s^*_{p,0} = s^*_{p,\text{predicted}}$
for $m = 1, 2, \ldots, M$
$d = A^{-H}_{s^*_{p,m-1}} A^{-2}_{s^*_{p,m-1}} b(t_p)$
$s^*_{p,m} = \frac{\Re\left({x'_p}^H d\right)}{\Re\left(x_p^H d\right)}$
if $s^*_{p,m} \le 0$ stop
if $|s^*_{p,m} - s^*_{p,m-1}| \le TOL$ break
end(m)
$s^*_p = s^*_{p,M}$
$s^*_{p+1,\text{predicted}} = s^*_p$
end(p)

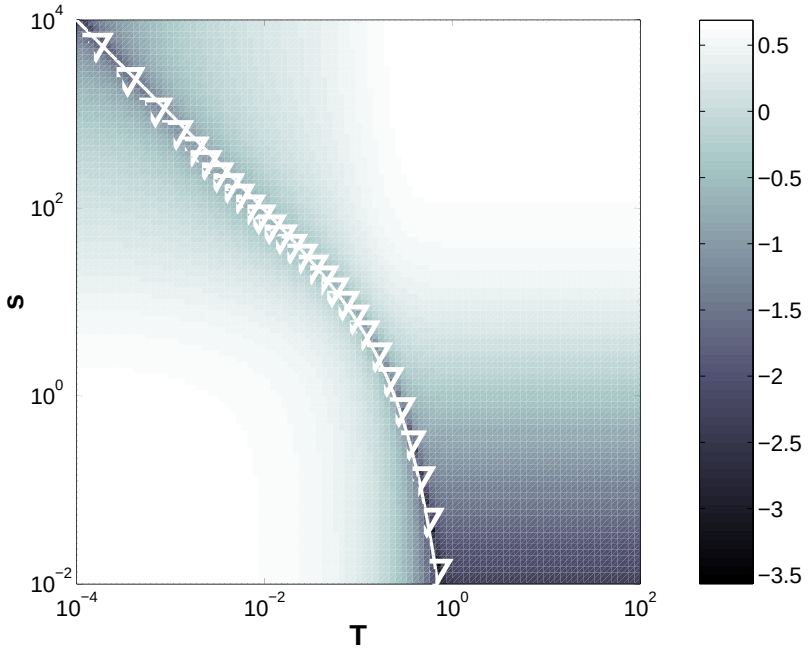

Figure 3: Levels of $\log_{10} \frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2}$ for the second-order central differentiation matrix $A = (n+1)^2$ tridiag$(-1, 2, -1)$ and the right-hand side $b(t) = (n+1)^2\ [1, 0 \ldots, 0, 1]^H$ of dimension $n = 24$ ($\lambda_n \simeq \pi^2$), function $s = \frac{\varphi_{-1}(-T\lambda_n)}{\varphi_0(-T\lambda_n)} \frac{1}{T}$ (while line) and values obtained from (3.3) (white inverted triangles).

One or two Picard iteration(s) ($m \le 2$) are generally sufficient to get satisfactory results because of the near independence of the right-hand side of (3.3) on s^* for small s^* ($A_{s^*} \simeq A$) and large s^* ($A_{s^*} \simeq s^* I$ so that the limit of the right-hand side for $s^* \to \infty$ is independent of s^*).

Remark 3: The simplified expression $d = b(t_q)$ in Algorithm 1 was also tested. It did not significantly affect the numerical results while making the process more efficient.

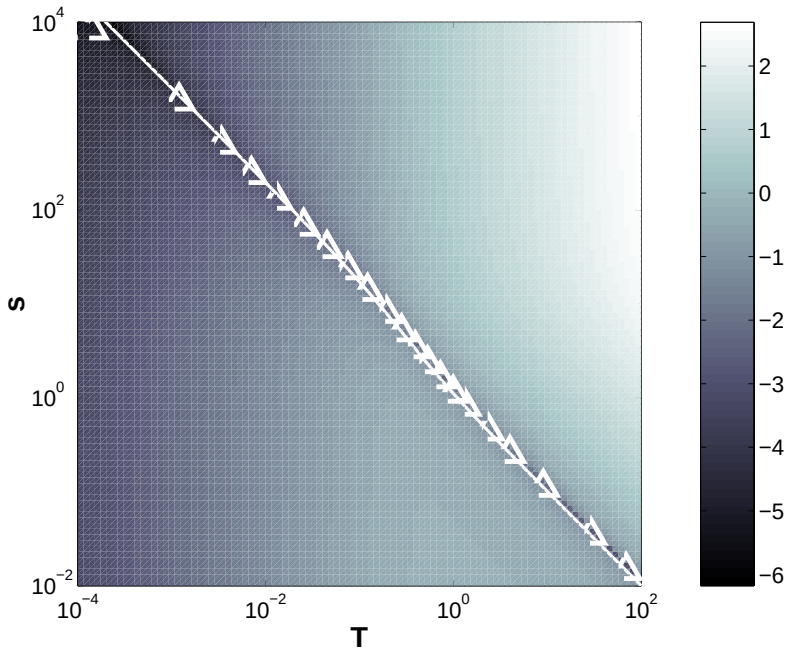

Figure 4: Levels of $\log_{10} \frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2}$ for the second-order central differentiation matrix $A = (n+1)^2$ tridiag$(-1, 2, -1)$ and the right-hand side $b(t) = t(n+1)^2\ [1, 0 \ldots, 0, 1]^H$ of dimension $n = 24$ ($\lambda_n \simeq \pi^2$), function $s = \frac{\varphi_0(-T\lambda_n)}{\varphi_1(-T\lambda_n)} \frac{1}{T}$ (while line) and values obtained from (3.3) (white triangles).

3.3 Effect on (3.3) of using a numerical integration scheme

Suppose the problem (1.1) is solved by applying q steps of a k-step multistep method

$$\sum_{j=0}^{k} \alpha_j x_{p-j} = h \sum_{j=0}^{k} \beta_j \left(b_{p-j} - A x_{p-j}\right) \tag{3.6}$$

where $x_j \simeq x(jh)$ and $b_j = b(jh)$ for $0 \le j \le q$ such that $qh = T$. The relation (3.6) can be written as $(\alpha \otimes I)\, X + (h\beta \otimes A)\, X = h\beta \otimes B$ or

$$\left((h\beta)^{-1}\alpha \otimes I\right) X + (I \otimes A)\, X = B, \tag{3.7}$$

with

$$\alpha = \left(\begin{array}{ccc|ccccc} S_\alpha & & & & & & & \\ \hline \alpha_k & \ldots & & \alpha_0 & & & & \\ & \ddots & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \alpha_k & \ldots & \alpha_0 \end{array}\right) = \left(\begin{array}{c|c} S_\alpha & 0 \\ \hline U_\alpha & L_\alpha \end{array}\right) \in \mathbb{R}^{(q+1)\times(q+1)}, \tag{3.8}$$

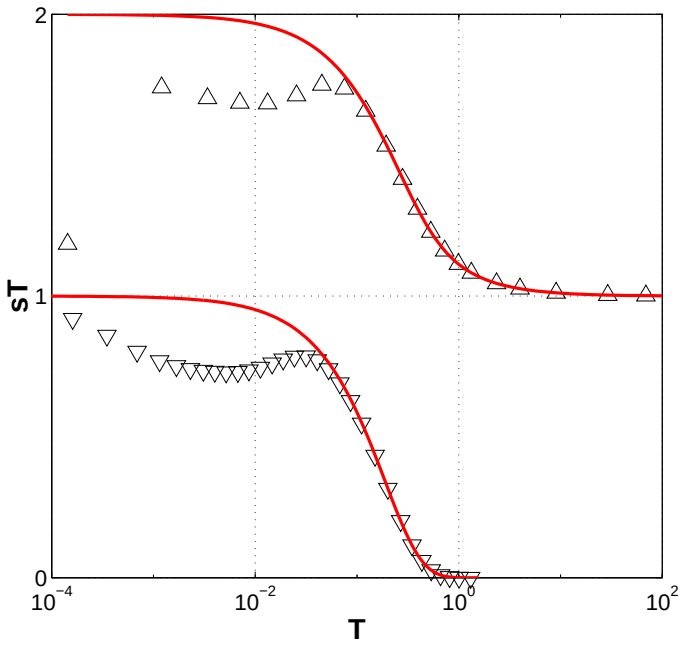

Figure 5: Product $C = s^*T$ versus T for the examples of Fig. 3 (right-hand side $b(t) = b$) ($\bigtriangledown$) and Fig. 4 (right-hand side $b(t) = tb$) ($\bigtriangleup$) together with the corresponding theoretical estimates (2.7) with $\mu_k = \lambda_n \simeq \pi^2$ for $k = 0, 1$.

$$\beta = \left(\begin{array}{c|ccc} S_\beta & & & \\ \hline \beta_k & \dots & \beta_0 & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \\ & & & \beta_k & \dots & \beta_0 \end{array}\right) = \left(\begin{array}{c|c} S_\beta & 0 \\ \hline U_\beta & L_\beta \end{array}\right) \in \mathbb{R}^{(q+1)\times(q+1)}, \tag{3.9}$$

and

$$X = \left(\begin{array}{c} 0 \\ x_1 \\ \vdots \\ \hline x_{k-1} \\ x_k \\ \hline \vdots \\ x_q \end{array}\right), \quad B = \left(\begin{array}{c} b_0 \\ b_1 \\ \vdots \\ \hline b_{k-1} \\ b_k \\ \hline \vdots \\ b_q \end{array}\right). \tag{3.10}$$

The symbol '$\otimes$' denotes the Kronecker product. The matrices S_α and S_β correspond to the starting procedure. In (3.7) we have assumed that S_β is nonsingular, $\beta_0 \neq 0$ and that the same time-step h is used in the starting procedure. Multiplying (3.7) by $e_{q+1}^H \otimes I$ with $e_{q+1}^H = (0, \dots, 0, 1)$ of dimension $q + 1$, we obtain

$$\left(e_{q+1}^H (h\beta)^{-1}\alpha \otimes I\right) X + Ax_q = b_q = b(T),$$

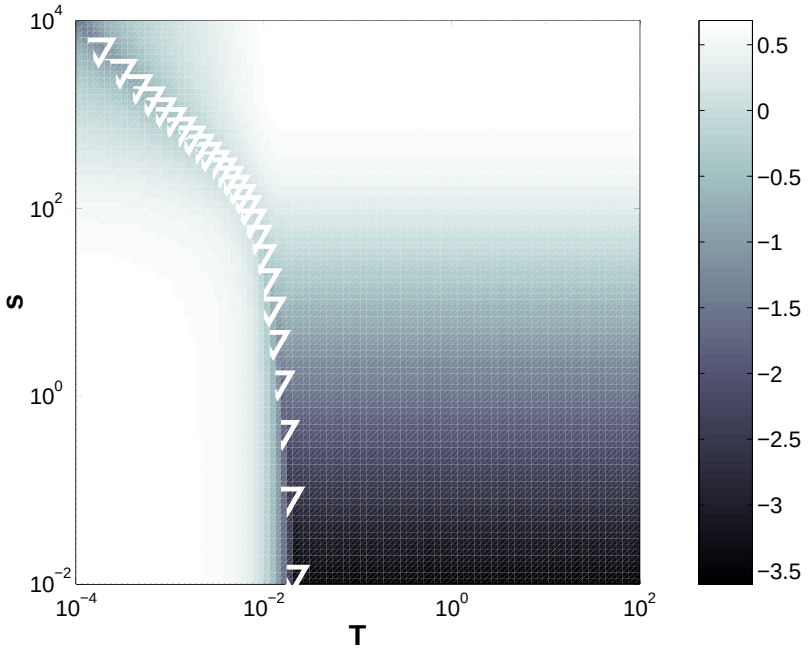

Figure 6: Levels of $\log_{10} \frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2}$ for the second-order central differentiation matrix $A = (n+1)^2$ tridiag$(-1, 6, -5)$ and the right-hand side $b(t) = (n+1)^2\ [5, 0 \ldots, 0, 1]^H$ of dimension $n = 24$ and values obtained from (3.3) (white inverted triangles).

i.e, the quantity $\left(e_{q+1}^H (h\beta)^{-1}\alpha \otimes I\right) X$ is the discrete equivalent of $x'(T)$. This leads to

$$s^* = \frac{\Re\left(X^H \left(e_{q+1}^H (h\beta)^{-1}\alpha \otimes I\right)^H A_{s^*}^{-H} A_{s^*}^{-2} b(T)\right)}{\Re\left(x_q^H A_{s^*}^{-H} A_{s^*}^{-2} b(T)\right)}. \tag{3.11}$$

A simpler estimate for the expression (3.11) can be derived if we assume that the (first q steps of the) computed solution X is of the form

$$X = \begin{pmatrix} 0 \\ zy \\ \vdots \\ z^q y \end{pmatrix} = u \otimes y, \qquad u = \begin{pmatrix} 0 \\ z \\ \vdots \\ z^q \end{pmatrix}, \tag{3.12}$$

for some $z > 1$ and some nonzero vector $y \in \mathbb{R}^n$.

Theorem 3.2 *Assume that the IVP (1.1) is solved using a k-step method (3.6) of order r and that the numerical solution can be approximated by a vector of the form (3.12). Then the optimal s^* given by (3.11) obtained after q steps reduces to*

$$s^* = \frac{\ln(z)}{h} + \frac{1}{h}\mathcal{O}\left(\frac{1}{z^{q-k+1}} + (z-1)^{r+1}\right). \tag{3.13}$$

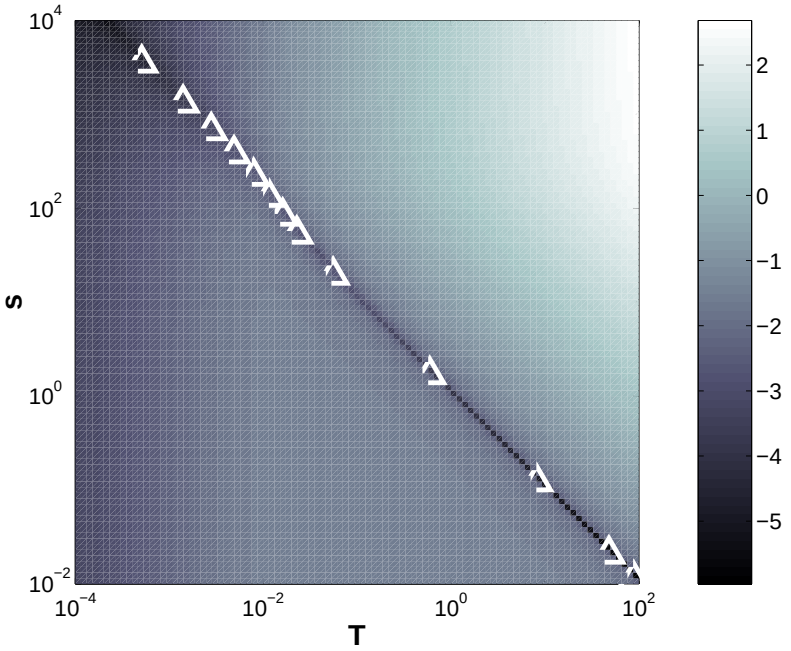

Figure 7: Levels of $\log_{10} \frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2}$ for the second-order central differentiation matrix $A = (n+1)^2$ tridiag$(-1, 6, -5)$ and the right-hand side $b(t) = t(n+1)^2\ [5, 0 \ldots, 0, 1]^H$ of dimension $n = 24$ and values obtained from (3.3) (white triangles).

Proof: From (3.8), (3.9), (3.12) and the properties of symbols σ of lower triangular Toeplitz matrices (see Lemma 6.3) we obtain

$$\begin{aligned}
\left(e_{q+1}^H (h\beta)^{-1}\alpha \otimes I\right) X &= e_{q+1}^H \left(\begin{array}{c|c} * & 0 \\ \hline * & (hL_\beta)^{-1} L_\alpha \end{array}\right) u \otimes y \\
&= \left[e_{q-k+1}^H (hL_\beta)^{-1} L_\alpha \begin{pmatrix} z^k \\ \vdots \\ z^q \end{pmatrix} + \frac{1}{h}\mathcal{O}\left(z^{k-1}\right)\right] \otimes y \\
&= \left[z^q \sigma_{(hL_\beta)^{-1}L_\alpha}\left(z^{-1}\right) + \frac{1}{h}\mathcal{O}\left(z^{k-1}\right)\right] y \\
&= \left[z^q \left(\frac{\sigma_{L_\alpha}\left(z^{-1}\right)}{h\sigma_{L_\beta}\left(z^{-1}\right)} + \frac{1}{h}\mathcal{O}\left(\left(z^{-1}\right)^{q-k+1}\right)\right) + \frac{1}{h}\mathcal{O}\left(z^{k-1}\right)\right] y \\
&= \left[z^q \frac{\underline{\alpha}(z)}{h\underline{\beta}(z)} + \frac{1}{h}\mathcal{O}\left(z^{k-1}\right)\right] y
\end{aligned}$$

where $\underline{\alpha}(z) = \sum_{j=0}^k \alpha_j z^{k-j}$ and $\underline{\beta}(z) = \sum_{j=0}^k \beta_j z^{k-j}$ are the characteristic polynomials of the multistep method (3.6). It then follows from $x_q = z^q y$ and (3.11) that

$$s^* = \frac{\underline{\alpha}(z)}{h\underline{\beta}(z)} + \frac{1}{h}\mathcal{O}\left(\frac{1}{z^{q-k+1}}\right). \tag{3.14}$$

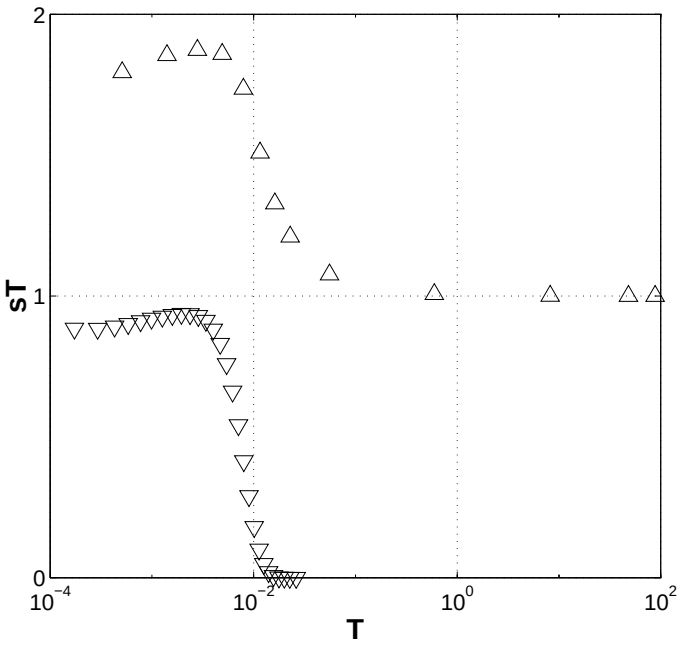

Figure 8: Product $C = s^*T$ versus T for the nonsymmetric examples of Fig. 6 (right-hand side $b(t) = b$) ($\bigtriangledown$) and Fig. 7 (right-hand side $b(t) = tb$) ($\bigtriangleup$).

The estimate

$$\frac{\underline{\alpha}(z)}{\underline{\beta}(z)} = \ln(z) + \mathcal{O}\left((z-1)^{r+1}\right) \tag{3.15}$$

follows from [7, p. 227] (see also [5, Theorem 2.4, p. 370]), $\underline{\beta}(z) = \underline{\beta}(1) + \mathcal{O}(z-1) = \mathcal{O}(1)$ and $\ln(z) = \mathcal{O}(z-1)$ for $z \to 1$. Combining (3.14) and (3.15) finally yields (3.13). □

Note that the leading term in the estimate (3.13) is independent of the particular numerical method used.

The substitution of $z = \epsilon^{-\frac{1}{q}}$ with $0 < \epsilon < 1$ into (3.14) leads to

$$s^* \simeq -\frac{\ln \epsilon}{T} + \mathcal{O}\left(\epsilon^{1-\frac{k-1}{q}} + \frac{(\ln \epsilon)^{r+1}}{q^r T}\right), \tag{3.16}$$

an expression already found in [8] but without proper mention of how ϵ relates to the IVP (1.1). From (3.12) we have

$$\frac{\|x_{q-1}\|}{\|x_q\|} = \frac{z^{q-1}\|y\|}{z^q\|y\|} = \frac{1}{z} = \epsilon^{\frac{1}{q}}$$

so that

$$\epsilon = \left(\frac{\|x_{q-1}\|}{\|x_q\|}\right)^q. \tag{3.17}$$

Thus ϵ is a measure of the growth in the solution from step $q-1$ to step q. Note that other expressions of ϵ can be obtained from (3.12), such as

$$\epsilon = \left(\frac{\|x_1\|}{\|x_q\|}\right)^{\frac{q}{q-1}}. \tag{3.18}$$

In practice, however, the numerical solution X cannot be written exactly in the form (3.12) so that the expressions (3.17) and (3.18) are not equivalent. The formula (3.17) is preferable since it minimizes the influence of the starting procedure and tends to better reflect changes in the solution as q increases.

In the case $b(t) = t^k b$ for some $b \neq 0$ we have $x_q \simeq (qh)^{k+1} y$ for some y (dependent on b and A) so that (3.17) yields

$$\epsilon \simeq \left(\frac{(q-1)^{k+1} h^{k+1} \|y\|}{q^{k+1} h^{k+1} \|y\|} \right)^q = \left(1 - \frac{1}{q} \right)^{(k+1)q} \simeq e^{-(k+1)} \tag{3.19}$$

for larger values of q and $-\ln \epsilon \simeq k+1$. This is consistent with the results of Section 2.1. Note however that the estimate (3.19) is obtained for larger values of q, i.e., large T, while (3.17) is based on the approximation (3.12) of the solution X, which does not generally hold on large windows.

4 Numerical example

The one-dimensional convection-diffusion equation

$$\begin{cases} x_t + a x_u - x_{uu} = 0, \quad 0 < u < 1, \ t \geq 0, \\ x(0,u) = 0, \quad 0 < u < 1 \\ x(t,0) = x(t,1) = f(t) \end{cases} \tag{4.1}$$

($a \geq 0$) is discretized using the method of lines with spatial step $h = \frac{1}{n+1} = \frac{1}{25}$.

A backward difference scheme is used for the first order space derivative (convection) while the usual central difference approximation is used for the second order space derivative (diffusion). This leads to a differential system of the form (1.1) with

$$A = \frac{1}{h^2} \begin{pmatrix} d & -1 & & \\ e & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & e & d \end{pmatrix}, \quad b(t) = \frac{f(t)}{h^2} \begin{pmatrix} -e \\ \vdots \\ \\ \vdots \\ 1 \end{pmatrix}, \tag{4.2}$$

$d = 2 + ah$ and $e = -1 - ah$, so that A is positive definite.

We first determine $\|x(T) - y(T)\|_2$ when $a = 0$. The matrix A is then symmetric. Since $c_n = \left(U^H b(T)\right)_n \neq 0$ we use $\mu_k = \lambda_n \simeq \pi^2$. The solution $x(T)$ and $x'(T)$ is obtained using the MATLAB adaptive stiff ode solver `ode15s`.

Fig. 3 shows the levels of $\log_{10} \frac{\|x(T)-y(T)\|_2}{\|y(T)\|_2}$ in the (T,s) plane over the range $[10^{-4}, 10^2] \times [10^{-2}, 10^4]$ for $f(t) = 1$, together with the curve $s = \frac{\varphi_{-1}(-\lambda_n T)}{\varphi_0(-\lambda_n T)} \frac{1}{T}$ obtained from (2.7) with $k = 0$ and $\mu_0 = \lambda_n$, and the discrete estimates computed from (3.3) using Algorithm 1 with only one Picard iteration (every fourth data point is shown). There is an excellent agreement between the discrete points, the theoretical curve and the actual position of the minima. Fig. 4 shows similar results for the boundary condition $f(t) = t$ ($k = 1$). As expected from the analysis of Section 2.1 the optimal value s^* now behaves as $\frac{1}{T}$ for larger values of T, compare (2.11). There is again excellent agreement between theoretical estimates and actual values. Fig. 5 displays the product $C = s^* T$ versus T for both cases $b(t) = b$ (bottom curve/points) and $b(t) = tb$ (top curve/points). Both sets of discrete estimates start with $s^* T = 1$ according to the initialization used in Algorithm 1. Note that for the case $b(t) = tb$ this initialization does not correspond to the correct limit $s^* T = 2$ given by the theoretical estimate (2.10). As the solution is being computed its behavior

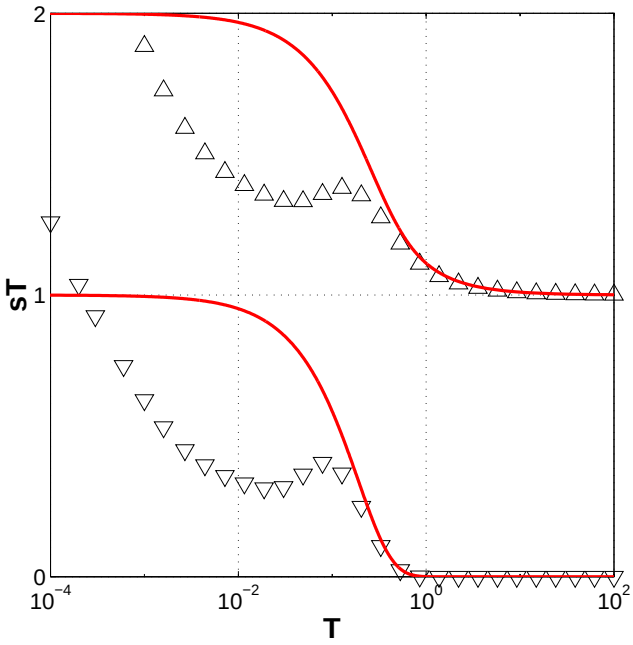

Figure 9: Product $C = s^*T = -\ln\epsilon$ versus T with ϵ determined from (3.17) for the symmetric examples of Fig. 3 (right-hand side $b(t) = b$) ($\bigtriangledown$) and Fig. 4 (right-hand side $b(t) = tb$) ($\bigtriangleup$) together with the corresponding theoretical estimates (2.7) (continuous lines). The numerical solution x_q is obtained using the explicit Euler method with time step $h = 10^{-4}$ ($h\rho(A) \simeq 0.25 < 2$; not all data points are shown). Compare with Fig. 5.

is increasingly taken into account by the numerical procedure 1 and the discrete points quickly move to a position closer to the theoretical curve defined by (2.7). Note also that in both cases the computed estimates remain between the theoretical bounds given by (2.8).

Numerical results for $b(t) = t^k b$ with $k > 1$ are similar to the case $k = 1$ modulo a shift up by $(k-1)$ unit(s) for the curves in Fig. 5. Although the estimates in Section 2.1 were derived for integer values of k, numerical experiments for non-integer values of k yield similar results.

We next consider the problem (4.1) with $a = 100$. The resulting matrix A in (4.2) is non-symmetric (and non-normal because of boundary effects). Figs. 6 and 7 show the value

$$\log_{10} \frac{\|x(T) - y(T)\|_2}{\|y(T)\|_2}$$

and should be compared to Figs. 3 and 4, respectively. The curve of minima is essentially shifted left by $\log_{10} 100 = 2$ units as soon as $T \simeq \frac{1}{100}$ in the case $b(t) = b$ but no significant change is observed in the case $b(t) = tb$. Since A is not hermitian the formula (2.7) is no longer valid, but it is tempting to generalize the results of Section 2.1 to the nonsymmetric example using instead the smallest singular value $\mu_0 = \sigma_n \simeq 167$ of A for the case $k = 0$ in (2.7). This modification yields however a curve which matches the computed minima for lower values of T only ($\sigma_n T < 1$).

Fig. 8 displays the corresponding estimates of the product s^*T obtained from Algorithm 1 and can be compared to Fig. 5. Note that the bounds given by (2.8) seem to remain valid in this nonsymmetric case.

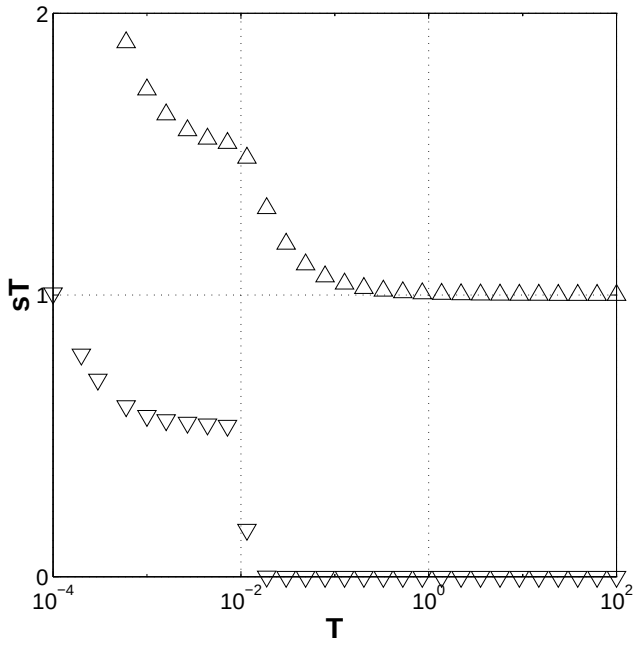

Figure 10: Same as Fig. 9 for the nonsymmetric case $a = 100$ with no theoretical estimate ($h\rho(A) \simeq 0.65 < 2$; not all data points are shown). Compare with Fig. 8.

Similar results where s^* is now obtained from (3.16) with ϵ specified by (3.17) are plotted in Figs. 9 and 10) for the cases $b(t) = b$ and $b(t) = tb$, respectively. The numerical solution x_q is determined by using the explicit Euler method with $h = 10^{-4}$ (so that it remains stable). These results compare quite well with Figs. 5 and 8. We have also checked that they were not significantly affected by the use of a different integrator such as the trapezoidal method applied to (1.1) or the trapezoidal rule applied to (1.2).

5 Conclusion

We have presented a new approach to determine the relation between the time window $[0, T]$ of integration of a linear initial value problem and a quantity s^* obtained by "freezing" the spectral parameter s in the Laplace transform of the kernel operator of the solution. This relation is of the form $s^*T = C$ where C is optimized by comparing the solutions of the initial value problem and the solution of the "frozen" spectral equation. This relation generalizes previous results by Leimkuhler [9] and Jackiewicz et. al. [8]. Theoretical and numerical results show that in shorter time windows C can be approximated by a constant which depends on the data of the problem itself.

We also proposed an algorithm which adaptively adjusts the optimal choice of s^* as the time window is increased. Applications to the preconditioning of differential linear systems and to the determination of optimal time windows in waveform relaxation methods are currently under investigation. Future work will also address the extension of the theory developed in this paper to the case the matrix A appearing in (1.1) is time dependent and to nonlinear problems. In particular, the analysis of Section 3 can be extended to time dependent case, although possibly

with a deterioration of the convergence rate of the Picard iteration in Algorithm 1 and a reduction of the time window on which convergence takes place. On the other hand, the more specific results of Section 2 seem more difficult to generalize unless additional assumptions such as commuting properties between $A(t)$ and its antiderivatives are made.

References

[1] K. Burrage, Z. Jackiewicz, S. P. Nørsett and R. Renaut, Preconditioning waveform relaxation iterations for differential systems, BIT 36(1996), 54–76.

[2] K. Burrage, Z. Jackiewicz and R. Renaut, Waveform relaxation techniques for pseudospectral methods, Numer. Methods Partial Differential Equations 12(1996), 245–263.

[3] K. Burrage, G. Hertono, Z. Jackiewicz and B. D. Welfert, Acceleration of convergence of static and dynamic iterations, BIT 41(2001), 645–655.

[4] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, North-Holland, Amsterdam (1984).

[5] E. Hairer, S. P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer Series in Computational Math., 2nd rev. ed., Springer (1993).

[6] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential–Algebraic Problems*, Springer Series in Computational Math., 2nd rev ed., Springer (1996).

[7] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York (1962).

[8] Z. Jackiewicz, B. Owren and B. Welfert, Pseudospectra of waveform relaxation operators, Comp. & Math. with Appls 36(1998), 67–85.

[9] B. Leimkuhler, Estimating waveform relaxation convergence, SIAM J. Sci. Comput. 14(1993), 872–889.

[10] E. Lelarasmee, *The waveform relaxation methods for the time domain analysis of large scale nonlinear dynamical systems*, Ph.D. Thesis, University of California, Berkeley, California 1982.

[11] L. Reichel and L.N. Trefethen, Eigenvalues and pseudo-eigenvalues of Toeplitz matrices, Linear Algebra Appl. 162(1992), 153–185.

[12] L.N. Trefethen, Pseudospectra of linear operators, SIAM Rev. 39(1997), 383–406.

6 Appendix

This appendix contains three technical lemmas about

1. The relationship between the $(k,1)$- and $(k,0)$-Padé approximations $R_{k,0}(z)$ and $R_{k,1}(z)$, respectively, to e^z at $z=0$.

2. Properties of the functions $\varphi(z)$ introduced in (2.2).

3. Symbols of lower triangular Toeplitz matrices.

Lemma 6.1 *For $k \geq 0$ we have*

$$R_{k,1}(z) = R_{k,0}(z) + \frac{z^{k+1}}{(k+1)!} \left(1 - \frac{z}{k+1}\right)^{-1}. \tag{6.1}$$

Proof: The explicit form of $R_{k,1}(z)$ is not needed to prove the lemma but can be found, for example, in [6, p. 48]. The numerator of

$$R_{k,0}(z) + \frac{z^{k+1}}{(k+1)!} \frac{1}{1 - \frac{z}{k+1}} = \frac{(1 - \frac{z}{k+1}) \left(\sum_{j=0}^{k} \frac{z^j}{j!}\right) + \frac{z^{k+1}}{(k+1)!}}{1 - \frac{z}{k+1}}$$

is a polynomial of degree at most $k+1$ whose z^{k+1} coefficient is $-\frac{1}{k+1}\frac{1}{k!} + \frac{1}{(k+1)!} = 0$, i.e., is of degree at most k. On the other hand, we have, for $z \to 0$,

$$R_{k,0}(z) + \frac{z^{k+1}}{(k+1)!} \frac{1}{1 - \frac{z}{k+1}} = \sum_{j=0}^{k+1} \frac{z^j}{j!} + \mathcal{O}(z^{k+2}).$$

Therefore the right-hand side of (6.1) is a rational approximation to e^z of order $k+1$ at $z = 0$ such that the degree of the numerator and denominator are k and 1, respectively, thus *is* the $(k, 1)$-Padé approximation by uniqueness of such approximation (see [6, Theorem 3.11 p. 48]). □

Lemma 6.2 *The functions φ_k defined in (2.2) satisfy (we define $\varphi_{-1}(z) = e^z$):*

(a) $\varphi_k(z) = \frac{k+1}{z} (\varphi_{k-1}(z) - 1)$ *for $z \neq 0$ and $k \geq 0$;*

(b) $\varphi_k'(z) = \left(1 - \frac{k+1}{z}\right) \varphi_k(z) + \frac{k+1}{z}$ *for $z \neq 0$ and $k \geq 0$;*

(c) $\varphi_k(z) = (k+1)! \int_{\Omega_k} e^{t_k z}\, dt_k \ldots dt_0 > 0$ *for $z \in \mathbb{C}$ and $k > 0$, where*

$$\Omega_k = \{(t_k, t_{k-1}, \ldots, t_0) : \ 0 \leq t_k \leq \cdots \leq t_0 \leq 1\} \subset \mathbb{R}^{k+1};$$

(d) $\left(\frac{\varphi_{k-1}(z)}{\varphi_k(z)}\right)' \geq 0$ *for $z \leq 0$ and $k \geq 0$.*

Proof: A direct calculation yields

$$\varphi_k(z) = (k+1)!\, z^{-(k+1)} \left(e^z - R_{k-1,0}(z) - \frac{z^k}{k!}\right) = \frac{k+1}{z} (\varphi_{k-1}(z) - 1)$$

for $z \neq 0$ and $k > 0$. It is easy to check that (a) also holds for $k = 0$ with the given definition of φ_{-1}. The relation (b) can be shown for example by using (a):

$$\begin{aligned} \varphi_k'(z) &= (k+1)! \left(z^{-(k+1)} (e^z - R_{k-1,0}(z)) - (k+1) z^{-(k+2)} (e^z - R_{k,0}(z))\right) \\ &= \frac{k+1}{z} (\varphi_{k-1}(z) - \varphi_k(z)) \ = \ \frac{k+1}{z} \left(\frac{z}{k+1} \varphi_k(z) + 1 - \varphi_k(z)\right) \\ &= \left(1 - \frac{k+1}{z}\right) \varphi_k(z) + \frac{k+1}{z}. \end{aligned}$$

The integral form (c) of $\varphi_k(z)$ can easily be checked for $k = 0$ and is proved for $k > 0$ by using the induction and (a). Since $\int_{\Omega_{k-1}} dt_{k-1} \dots dt_0 = \frac{1}{k!}$ we obtain for $z \neq 0$

$$\begin{aligned}
\varphi_k(z) &= \frac{k+1}{z} \left(k! \int_{\Omega_{k-1}} e^{t_{k-1} z} dt_{k-1} \dots dt_0 \ - 1 \right) \\
&= (k+1)! \int_{\Omega_{k-1}} \frac{e^{t_{k-1} z} - 1}{z} \, dt_{k-1} \dots dt_0 \\
&= (k+1)! \int_{\Omega_{k-1}} \left(\int_0^{t_{k-1}} e^{t_k z} dt_k \right) dt_{k-1} \dots dt_0 \\
&= (k+1)! \int_{\Omega_k} e^{t_k z} \, dt_k \dots dt_0.
\end{aligned}$$

Since $\varphi_k(0) = 1 = (k+1)! \int_{\Omega_k} dt_k \dots dt_0$ the result holds also for $z = 0$.

The formula (c) implies that $\varphi_k(z) > 0$ for $z \in \mathbb{R}$. Moreover, the p-th derivative of φ_k becomes

$$\varphi_k^{(p)}(z) = (k+1)! \int_{\Omega_k} t_k^p e^{t_k z} \, dt_k \dots dt_0 > 0$$

for $z \in \mathbb{R}$. For the functions f, $g : \Omega_k \to \mathbb{R}$ define the scalar inner product by

$$\langle f, g \rangle = \int_{\Omega_k} f(t_k, \dots, t_0) g(t_k, \dots, t_0) dt_k \dots dt_0.$$

Applying the Cauchy-Schwarz inequality to the functions

$$f(t_k, \dots, t_0) = e^{\frac{t_k z}{2}}, \quad g(t_k, \dots, t_0) = t_k e^{\frac{t_k z}{2}}$$

we obtain

$$\begin{aligned}
\frac{(\varphi_k'(z))^2}{((k+1)!)^2} &= \langle f, g \rangle^2 \ = \ \left(\int_{\Omega_k} t_k e^{t_k z} dt_k \dots dt_0 \right)^2 \\
&\leq \int_{\Omega_k} e^{t_k z} dt_k \dots dt_0 \int_{\Omega_k} t_k^2 e^{t_k z} dt_k \dots dt_0 \ = \ \frac{\varphi_k(z) \varphi_k''(z)}{((k+1)!)^2}.
\end{aligned}$$

This leads to

$$\left(\varphi_k'(z)\right)^2 \leq \varphi_k(z) \varphi_k''(z)$$

for $z \in \mathbb{R}$. Hence,

$$\left(\frac{\varphi_k'(z)}{\varphi_k(z)} \right)' = \frac{\varphi''(z)\varphi_k(z) - (\varphi_k'(z))^2}{\varphi_k^2(z)} \geq 0$$

and, as a result, the function $\frac{\varphi_k'(z)}{\varphi_k(z)}$ is non-decreasing on $\mathbb{R}$ for all $k \geq 0$. On the other hand (a) and (b) yield

$$\varphi_{k-1}'(z) = \varphi_{k-1}(z) - \frac{k}{z} \left(\varphi_{k-1}(z) - 1 \right) = \varphi_{k-1}(z) - \frac{k}{k+1} \, \varphi_k(z)$$

for $z \neq 0$ and $k > 0$. Consequently, the function $\frac{\varphi_k(z)}{\varphi_{k-1}(z)} = \frac{k+1}{k} \left(1 - \frac{\varphi_{k-1}'(z)}{\varphi_{k-1}(z)} \right)$ is non-increasing, i.e., $\left(\frac{\varphi_k(z)}{\varphi_{k-1}(z)} \right)' \leq 0$ and $\left(\frac{\varphi_{k-1}(z)}{\varphi_k(z)} \right)' \geq 0$ for $z \leq 0$. The property (d) can easily be verified in the

case $k = 0$. □

The symbol of a lower triangular Toeplitz matrix

$$R = \begin{pmatrix} r_0 & & \\ \ddots & \ddots & \\ r_{n-1} & \ddots & r_0 \end{pmatrix} = \text{toeplitz}(r_0, \ldots, r_{n-1}) \in \mathbb{R}^{n\times n}$$

is defined as the (polynomial) function

$$\sigma_R(z) = \sum_{j=0}^{n-1} r_j z^j = (0, \ldots, 1) R \begin{pmatrix} z^{n-1} \\ \vdots \\ 1 \end{pmatrix}$$

(see [11] for example).

Lemma 6.3 *Let T_1 and $T_2 = \text{toeplitz}(\beta_0, \ldots, \beta_{n-1})$ be two $n \times n$ lower triangular Toeplitz matrices. Then*

$$\sigma_{T_2T_1}(z) = \sigma_{T_2}(z)\sigma_{T_1}(z) + \mathcal{O}(z^n). \tag{6.2}$$

for $|z| < 1$. Moreover, if $\beta_0 \neq 0$ and if $\sum_{j=0}^{n-1} |\beta_j|$ can be bounded independently of n, then

$$\sigma_{T_2^{-1}T_1}(z) = \frac{\sigma_{T_1}(z)}{\sigma_{T_2}(z)} + \mathcal{O}(z^n). \tag{6.3}$$

Proof: The product T_2T_1 is again a lower triangular Toeplitz matrix. The relation (6.2) is a discrete convolution formula which is easy to verify by multiplying out the symbols of the two matrices and comparing it with the symbol of the product of the two matrices. The second relation (6.3) can be derived from (6.2) as

$$\sigma_{T_1}(z) = \sigma_{T_2\left(T_2^{-1}T_1\right)}(z) = \sigma_{T_2}(z)\sigma_{T_2^{-1}T_1}(z) + \mathcal{O}(z^n)$$

and the fact that $|\sigma_{T_2}(z)| \leq \sum_{j=0}^{n-1} |\beta_j| = \mathcal{O}(1)$ for $|z| < 1$ independently of n. □

Note that the condition in Lemma 6.3 that $\sum_{j=0}^{n-1} |\beta_j|$ be bounded independently of the dimension n of T_2 is in particular satisfied if T_2 has a fixed bandwidth.