Numerical stability in the presence of variable coefficients

Ernst Hairer
Section de mathématiques
Université de Genève
Email: Ernst.Hairer@unige.ch

Arieh Iserles
Department of Applied Mathematics and Theoretical Physics
Centre for Mathematical Sciences
University of Cambridge
Email: A.Iserles@damtp.cam.ac.uk

April 3, 2015

Abstract

The main concern of this paper is with the stable discretisation of linear partial differential equations of evolution with time-varying coefficients. We commence by demonstrating that an approximation of the first derivative by a skew-symmetric matrix is fundamental in ensuring stability for many differential equations of evolution. This motivates our detailed study of skew-symmetric differentiation matrices for univariate finite-difference methods.

We prove that, in order to sustain a skew-symmetric differentiation matrix of order $p \geq 2$, a grid must satisfy $2p-3$ polynomial conditions. Moreover, once it satisfies these conditions, it supports a banded skew-symmetric differentiation matrix of this order and of the bandwidth $2p-1$ which can be derived in a constructive manner.

Some applications require not just skew-symmetry but also that the growth in the elements of the differentiation matrix is at most linear in the number of unknowns. This is always true for our tridiagonal matrices of order 2 but need not be true otherwise, a subject which we explore further.

Another subject which we examine is the existence and practical construction of grids that support skew-symmetric differentiation matrices of a given order. We resolve this issue completely for order-two methods.

We conclude the paper with a list of open problems and their discussion.

[0] ***Communicated by Peter Olver***
[0] AMS (MOC) Subject classification: 65M12, 65D25
[0] **Keywords:** Partial differential equations, finite difference methods, numerical stability, skew-symmetric differentiation matrices, order conditions

1 On the importance of being skew-symmetric

Classical stability theory for discretised time-dependent partial differential equations (PDEs), since its origins in the work of John von Neumann and Richard Courant, is fundamental to modern numerical analysis. The *Lax Equivalence Theorem* (Richtmyer & Morton 1967) means that stability, far from being an optional extra for a numerical method, is a necessary condition for the discretisation to converge to the exact solution (subject to very generous side conditions) once the number of degrees of freedom becomes infinite. Yet, it is clear that numerical stability theory is replete with lacunæ and open problems. The main purpose of this paper is to address a major gap in our understanding of this issue, namely numerical stability in the presence of variable coefficients.

Standard treatment of numerical stability commences from the PDE

$$\frac{\partial u}{\partial t} = \mathcal{L}u + f, \qquad \boldsymbol{x} \in \Omega, \quad t \geq 0, \tag{1.1}$$

where $u = u(\boldsymbol{x}, t)$, given with initial conditions for $u(\boldsymbol{x}, 0)$, $\boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$, and suitable boundary conditions on $\partial\Omega$. Here $\mathcal{L}$ is a linear, time-independent differential operator, while both the forcing term f and the initial and boundary conditions are suitably smooth. We denote by K the degree of the highest spatial derivative present in $\mathcal{L}$.

The work of this paper applies to all discretisation methods which are concerned with nodal values – in other words, whose unknowns are approximate solution values at a given number of points in $\mathbb{R}^d$. Hence, at least in principle they apply to finite differences, finite elements, finite volumes and spectral collocation, but not to spectral methods. For the sake of simplicity, we henceforth restrict the narrative to finite difference methods. We also for the time being assume zero Dirichlet boundary conditions but this restriction can be easily lifted.

Discretising the PDE (1.1), the outcome is the recurrence

$$\boldsymbol{u}_N^{n+1} = \mathcal{A}_N \boldsymbol{u}_N^n + \boldsymbol{f}_N^n, \qquad t \geq 0, \tag{1.2}$$

where $\boldsymbol{u}_N^n = (u_{N,1}^n, \ldots, u_{N,N}^n)$: for example, in the case of finite differences, $u_{N,m}^n \approx u(\boldsymbol{x}_m, n\Delta t)$, where $\boldsymbol{x}_m$ is a grid point (assuming for simplicity Dirichlet boundary conditions, there are N such points in the interior of Ω) and $\Delta t > 0$ is the time step. Two conditions join to assure us that, as $N \to \infty$, $\boldsymbol{u}_N^n$ converges pointwise to the exact solution of (1.1): *consistency* (i.e., that *locally* (1.2) matches the original PDE up to $\mathcal{O}\big(N^{-K-1}\big)$ for some $K \geq 1$, something that can be usually verified easily by Taylor expansion) and *stability:* uniform well-posedness of the operators $\{\mathcal{A}_N\}$ as $N \to \infty$ in the time interval $[0, T]$ (Richtmyer & Morton 1967). In other words, letting $\| \cdot \|_N$ be (without loss of generality) the ℓ_2 norm on $\mathbb{R}^N$, we require that, once $N^K \Delta t$ is uniformly bounded as $N \to \infty$, we have

$$\limsup_{N \to \infty} \|\mathcal{A}_N^n\|_N < \infty, \qquad n\Delta t < T. \tag{1.3}$$

We note for further reference that a very helpful sufficient condition for (1.3) is that

$$\|\mathcal{A}_N\|_N \leq 1 + c\Delta t, \qquad N \gg 1, \tag{1.4}$$

where $c \geq 0$ is independent of N.

Once the coefficients of $\mathcal{L}$ are constant, two powerful techniques – eigenvalue analysis and Fourier analysis – can be used to investigate a wide range of numerical methods (Iserles 2008). Although immensely useful, the two techniques are far from comprehensive, and this has led to a wide range of further tools, e.g. the Kreiss matrix theorem (Kreiss 1962), the GKS theory (Gustafsson, Kreiss & Sundström 1972, Trefethen 1983) and pseudo-eigenvalues (Reddy & Trefethen 1992). Finally, if everything else fails, one can always attempt the *energy method* (Richtmyer & Morton 1967) – essentially, proving (1.4) from first principles.

Matters become considerably more complicated once the coefficients of $\mathcal{L}$ are allowed to depend on $\boldsymbol{x}$. Fourier analysis is no longer suitable, eigenvalue analysis has a fairly limited scope and the main practical tool is the energy method – and the latter is suitable only for fairly simple methods.

Consider the following differential equations:

$$\begin{aligned}
&\text{The diffusion equation:} && \frac{\partial u}{\partial t} = \boldsymbol{\nabla}^\top a(\boldsymbol{x}) \boldsymbol{\nabla} u, \qquad \min_{\boldsymbol{x} \in \Omega} a(\boldsymbol{x}) > 0, \\
&\text{The Liouville equation:} && \frac{\partial u}{\partial t} + \boldsymbol{V}(x) \cdot \boldsymbol{\nabla} u = 0, \\
&\text{Convection–diffusion:} && \frac{\partial u}{\partial t} + \boldsymbol{V}(x) \cdot \boldsymbol{\nabla} u = \varepsilon \Delta u, \qquad 0 < \varepsilon \ll 1, \\
&\text{The Fokker–Planck equation:} && \frac{\partial u}{\partial t} + \sum_{i=1}^{d} \frac{\partial \mu_i(\boldsymbol{x}) u}{\partial x_i} = \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 \mathcal{R}_{i,j}(\boldsymbol{x}) u}{\partial x_i \partial x_j}
\end{aligned}$$

where the matrix $\mathcal{R}$ is positive semidefinite.

All these equations share several features. Firstly, they all feature in a long list of practical applications and their numerical solution is more than a matter of purely intellectual interest. Secondly, they all involve variable coefficients. Thirdly, the present state of knowledge restricts their stability analysis to simple, low-order methods. In this paper we demonstrate that they also share another feature: they all can be discretised stably *once first space derivatives are approximated by skew-symmetric matrices!*

The aforementioned statement is trivial in the case of the diffusion equation. Thus, given a grid $\{\boldsymbol{x}_m\}_{m=0}^{N+1}$, let $u_m(t) \approx u(\boldsymbol{x}_m, t)$. Ordering grid points in any manner, we denote the vector of approximate nodal values by $\boldsymbol{u}$. Supposing that we approximate $\partial \boldsymbol{w}/\partial x \approx \mathcal{D}\boldsymbol{w}$: in one space dimension this leads to the *semi-discretisation*

$$\boldsymbol{u}' = \mathcal{D}\mathcal{A}\mathcal{D}\boldsymbol{u}, \tag{1.5}$$

where $\boldsymbol{A}$ is a diagonal matrix with the (positive!) diffusion coefficients along its diagonal. Since $\mathcal{A}$ is positive definite, once $\mathcal{D}$ is skew-symmetric, it is trivial that $\mathcal{D}\mathcal{A}\mathcal{D}$ is negative definite and the solution of (1.5) with any A-stable ODE method (Iserles 2008) produces a stable discretisation (1.2). Moreover, once the differentiation matrix approximates the exact derivative to order $p \geq 1$ and the semi-discretised equations (1.5) are solved by an order-$\lfloor p/2 \rfloor$ ODE method, the outcome is a method of order p. (The reason for the mismatch in the order of the differentiation matrix and the ODE method is that $N^2(\Delta t) = \mathcal{O}(1)$.) This argument can be easily extended to any

number of space dimensions because a sum of negative-definite matrices is negative definite.

We shift our gaze to equations involving the *material derivative* $\partial/\partial t + \boldsymbol{V}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}$, e.g. the Liouville equation and the (linear) convection-diffusion equation.

Definition *Let* $\{\mathcal{A}_N\}$ *be an infinite family of matrices such that* $\mathcal{A}_N \in \mathrm{M}_N[\mathbb{C}]$. *We say that it is* stable *if there exists* $c > 0$ *such that* $\limsup_{N\to\infty} \|\mathrm{e}^{t\mathcal{A}_N}\|_N \le 1 + ct$ *for sufficiently small* $t \ge 0$.

The above definition of stability, designed to fit into our narrative of PDE stability in the sense of Lax, is of course norm dependent – here and throughout this paper, unless otherwise stated, we use the ℓ_2 norm. We commence from the one-dimensional case $d = 1$. Without loss of generality we may assume that $\Omega = [0, 1]$ and the grid is $0 = x_0^{(N)} < x_1^{(N)} < \cdots < x_N^{(N)} < x_{N+1}^{(N)} = 1$. The following proof is a mild extension of a result in (Iserles 2014).

Theorem 1 *Let*

1. *The grid be dense in* $[0, 1]$ *for* $N \gg 1$*:* $\sigma_N := \max_{m=0,\ldots,N}(x_{m+1} - x_m) = \mathcal{O}\big(N^{-1}\big)$*;*

2. $V : [0, 1] \to \mathbb{R}$ *be Lipschitz and* $\mathcal{V}_N \in \mathrm{M}_N[\mathbb{R}]$ *a diagonal matrix,* $\mathcal{V}_{m,m} = V(x_m)$*,* $m = 1, \ldots, N$*;*

3. *The differentiation matrix* $\mathcal{D}_N$ *be skew-symmetric, banded (i.e.,* $\mathcal{D}_{k,\ell} = 0$ *for* $|k - \ell| \ge r + 1$ *for some* $r \ge 1$*) and suppose that* $\max_{k,\ell} |\mathcal{D}_{k,\ell}| \le b^\star N$*, where* $b^\star > 0$ *is independent of* N.

Then the families of matrices $\{\mathcal{A}_N\} = \{\mathcal{V}_N\mathcal{D}_N\}$ *and* $\{\mathcal{B}_N\} = \{\mathcal{D}_N\mathcal{V}_N\}$ *are stable.*

Proof We suppress the subscript N in $\mathcal{V}_N$, $\mathcal{D}_N$, $\mathcal{A}_N$, $\mathcal{B}_N$, and we recall the inequality

$$\|\mathrm{e}^{t\mathcal{A}}\| \le \mathrm{e}^{t\mu[\mathcal{A}]}, \qquad t \ge 0, \tag{1.6}$$

where $\mu : \mathrm{M}_N[\mathbb{R}] \to \mathbb{R}$ is the *logarithmic norm:* $\mu[\mathcal{A}] = \lim_{\varepsilon\downarrow 0} \varepsilon^{-1}(\|I + \varepsilon A\| - 1)$ (Söderlind 2006). Since we are working in the ℓ_2 norm, it is well known that $\mu[\mathcal{A}]$ equals the rightmost eigenvalue (a.k.a. the *spectral abscissa*) of the symmetric matrix $\mathcal{C} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^\top)$.

Recalling that $\mathcal{V}$ is diagonal and $\mathcal{D}$ skew-symmetric and banded, letting $\mathcal{A} = \mathcal{V}\mathcal{D}$ we have

$$\mathcal{C}_{m,k} = \begin{cases} 0, & m = k \text{ or } |m - k| \ge r + 1, \\ \frac{1}{2}(\mathcal{V}_{m,m} - \mathcal{V}_{k,k})\mathcal{D}_{m,k}, & 1 \le |m - k| \le r. \end{cases}$$

Since the eigenvalues of a matrix are bounded by its infinity matrix norm we have

$$\mu[\mathcal{V}\mathcal{D}] \le \|\mathcal{C}\|_\infty = \max_{m=1,\ldots,N} \sum_{k=1}^{N} |\mathcal{C}_{m,k}| = \tfrac{1}{2} \sum_{|m-k|\le r} |\mathcal{V}_{m,m} - \mathcal{V}_{k,k}||\mathcal{D}_{m,k}|.$$

Recall that $\mathcal{V}_{m,m} = V(x_m^{(N)})$, $m = 1, \ldots, N$, that V is Lipschitz (i.e. $|V(x) - V(y)| \leq \lambda|x - y|$ for some $\lambda \geq 0$) and that $\sigma_N = \mathcal{O}\big(N^{-1}\big)$. Therefore

$$\begin{aligned}\mu[\mathcal{V}\mathcal{D}] &\leq \tfrac{1}{2} \sum_{|m-k|\leq r} |V(x_m^{(N)}) - V(x_k^{(N)})||\mathcal{D}_{m,k}| \leq \frac{\lambda b^\star N}{2} \sum_{|m-k|\leq r} |x_m^{(N)} - x_k^{(N)}| \\ &\leq \lambda b^\star N r \sigma_N.\end{aligned}$$

Since $\sigma_N = \mathcal{O}\big(N^{-1}\big)$, it follows that there exists $c_1 \geq 0$ such that $\mu[\mathcal{V}\mathcal{D}] \leq c_1$ and this, by virtue of (1.6), proves stability for $\mathcal{V}\mathcal{D}$.

The proof for $\mathcal{D}\mathcal{V}$ follows by an identical argument. □

The condition of $\mathcal{D}$ being banded is not strictly necessary for the proof, it is enough to assume that its coefficients decay sufficiently fast away from the diagonal. Yet, to keep the focus of the proof on its essential ingredients, we leave this easy generalisation to the reader. The requirement that the entries $\mathcal{D}_{m,k}$ are $\mathcal{O}(N)$ makes sense given that we are approximating the first derivative in an N-dimensional space.

It is an immediate consequence from the definition of the logarithmic norm and the triangle inequality that it is sub-additive,

$$\mu[\mathcal{A}_1 + \cdots + \mathcal{A}_d] \leq \sum_{\ell=1}^{d} \mu[\mathcal{A}_\ell].$$

Therefore, once each $\partial/\partial x_\ell$ is discretised by a skew-symmetric matrix, Theorem 1 implies stability. Consequently, skew-symmetry of the differentiation matrix is sufficient for the stability of finite-difference discretisation of the Liouville equation. Moreover, as long as the Laplacian is discretised stably (e.g. with a skew-symmetric matrix!), it is sufficient for the stability in the convection-diffusion case.

In the Fokker–Planck case the operator on the left, $\sum_{i=1}^{d} \partial[\mu_i(\boldsymbol{x})u]/\partial x_i$, can be discretised stably, in line with the theorem, with $\sum_{i=1}^{d} \mathcal{D}_i\mathcal{M}_i\boldsymbol{u}$, where $\mathcal{D}_i$, the discretisation of $\partial/\partial x_i$, is skew-symmetric and $\mathcal{M}_i$, the discretisation of multiplication by $\mu_i(\boldsymbol{x})$, is diagonal. Insofar as the right-hand side is concerned, let us assume for simplicity that the *diffusion tensor* $\mathcal{R}_{i,j}$ is independent of (i, j) and $\boldsymbol{x}$, $\mathcal{R}_{i,j}(\boldsymbol{x}) \equiv \rho \geq 0$. In that case the operator on the right-hand side is approximated by

$$\rho \sum_{i=1}^{d} \sum_{j=1}^{d} \mathcal{D}_i\mathcal{D}_j = \rho \left(\sum_{i=1}^{d} \mathcal{D}_i \right)^2,$$

where $\mathcal{D}_i$ is the differentiation matrix with respect to $\partial/\partial x_i$: once all the $\mathcal{D}_i$s are skew-symmetric, the above expression is negative semi-definite and stability follows.

Remarkably, the same feature of the discretisation process, skew-symmetry of the differentiation matrix, is sufficient for stability in a wide range of situations: the examples above are far from exhaustive. Although the conditions of Theorem 1 are 'just' sufficient for stability and we cannot exclude *a priori* the possibility of non-skew-symmetric differentiation matrices leading to stable schemes, this universality justifies the focus of this paper on skew-symmetric differentiation matrices.

Restricting henceforth our attention to univariate grids, the issue at hand is deceptively simple: given a grid $\{\boldsymbol{x}^{(N)}\}$, find a skew-symmetric $N \times N$ differentiation matrix $\mathcal{D}$ that approximates the first derivative to a given order $p \geq 1$. Assuming zero Dirichlet boundary conditions, order p is equivalent to the statement that

$$u'(x_m) = \sum_{k=1}^{N} \mathcal{D}_{m,k} u(x_k), \qquad m = 1, \ldots, N, \tag{1.7}$$

for every pth-degree polynomial u that vanishes at the endpoints. In the case of non-zero Dirichlet boundary conditions the relevant statement reads

$$u'(x_m) = \alpha_m u(x_0) + \sum_{k=1}^{N} \mathcal{D}_{m,k} u(x_k) + \gamma_m u(x_{N+1}), \qquad m = 1, \ldots, N, \tag{1.8}$$

where again u is a polynomial, $\deg u \leq p$, while $\{\alpha_m\}$ and $\{\gamma_m\}$ are constants. The coefficients $\mathcal{D}_{m,k}$ in (1.8) are the same as that in (1.7), and α_m, γ_m are readily obtained by letting $u(x) = x_{N+1} - x$ and $u(x) = x - x_0$, respectively.

Perhaps the most familiar differentiation matrix, the *central difference scheme*

$$\mathcal{D} = \begin{bmatrix} 0 & (N+1)/2 & 0 & \cdots & 0 \\ -(N+1)/2 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & (N+1)/2 \\ 0 & \cdots & 0 & -(N+1)/2 & 0 \end{bmatrix},$$

given on a uniform grid, $x_m = m/(N+1)$, is of order $p = 2$ and skew-symmetric. This example, though, is deceptive: it has been proven in (Iserles 2014) that this is as good as it gets: the highest order of a skew-symmetric differentiation matrix on a uniform grid is just two.[1]

An obvious remedy is to resort to a non-uniform grid. Mild perturbations of the uniform grid have resulted in (Iserles 2014) in third and fourth-order skew-symmetric differentiation matrices. In Section 2 we address this issue in considerably greater generality and demonstrate that a skew-symmetric matrix of order $p \geq 2$ can be constructed on the grid $\{x_m\}$ if and only if $2p-3$ polynomial conditions of the form $R_k(x_1, \ldots, x_N) = 0$, $k = 1, \ldots, 2p-3$, are satisfied by the grid points $x_1, \ldots, x_N$.

The mere existence of an order-p skew-symmetric differentiation matrix on a given grid is of interest, but we want more. Firstly, we want a *banded* skew-symmetric matrix – not just so as to be in line with Theorem 1 but also because practical numerics are considerably cheaper with banded matrices. Secondly, we wish for a constructive algorithm that automatically produces a skew-symmetric matrix of given order once the order conditions on $x_1, \ldots, x_N$ are satisfied. In Section 3 we address both problems in unison. We seek there a pth-order skew-symmetric matrix with bandwidth $2p-1$ (in

[1] Once we replace Dirichlet with periodic boundary conditions, the problem becomes trivial and it is exceedingly easy to present explicitly skew-symmetric circulant matrices which approximate the first derivative to any even order.

the example above $p = 2$). *A priori,* our result that the conditions $R_k(x_1, \ldots, x_N) = 0$, $k = 1, \ldots, 2p-3$ ensure the existence of a skew-symmetric pth-order differentiation matrix need not imply that such a *banded* matrix is possible. Specifically, an explicit construction of such matrix (hence fulfilling the second desideratum, a constructive algorithm!) seems to require $(p-1)p/2$ conditions. We prove that these conditions are not independent and that they reduce to exactly the above $2p-3$ polynomial identities $R_k(x_1, \ldots, x_N) = 0$, $k = 1, 2, \ldots, 2p-3$.

Sections 2 and 3 address two of the conditions on the differentiation matrix $\mathcal{D}$ in Theorem 1: skew-symmetry and bandedness. In Section 4 we consider the remaining condition, size: is it true for the skew-symmetric matrices of Section 3 that $\max_{k,\ell} |\mathcal{D}_{k,\ell}| \leq b^\star N$? The answer is positive for tridiagonal matrices, otherwise rather more tentative. Within the methodology underlying the proof of Theorem 1, quindiagonal differentiation matrices are 'unstable', or at least their logarithmic norm becomes unbounded for $N \gg 1$. Computer experiments, however, show that everything depends on the function V: the norm of the exponential is sometimes bounded (hence the method stable), for other functions V it is unbounded and further research is required.

In Section 5 we explore the conditions for the existence of grids consistent with order-p conditions on skew-symmetric differentiation matrices. We derive a set of $2p-3$ necessary conditions on a 'grid function' g, such that $x_m^{(N)}$ is an $\mathcal{O}\big(N^{-2}\big)$ perturbation of $g(m/(N+1))$, $m = 0, \ldots, N+1$ and prove that in the simplest possible case, $p = 2$, these conditions are also sufficient. In the latter case we also present an efficient algorithm for the computation of the $x_m^{(N)}$s.

This paper addresses an issue in PDE stability theory which, at this level of generality, has somehow escaped the attention of computational mathematicians in the last half-century. Needless to say, it is but an initial foray into a large and important area and our results, while solving some long-standing problems, pose many new questions. These questions and further thoughts on the interplay of skew-symmetry and stability are addressed in Section 6.

Skew-symmetry of a differentiation matrix is important not just in ensuring that a numerical method is stable but also in geometric numerical integration (Hairer, Lubich & Wanner 2006), specifically in numerical discretisation of differential equations which respects their first integrals (Kitson, McLachlan & Robidoux 2003) or unitarity (Bader, Iserles, Kropielnicka & Singh 2014). Note that in both these publications boundary conditions are presumed to be periodic and this makes the design of skew-symmetric differentiation matrices of an arbitrary order significantly easier. Indeed, the initial motivation for the work of this paper and for (Iserles 2014) is an attempt to design high-order unitarity-preserving methods for the semiclassical Schrödinger equation *á la* (Bader et al. 2014) while imposing zero Dirichlet boundary conditions.

2 Necessary conditions on non-uniform grids

Without loss of generality we consider a grid $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ on the interval $[0, 1]$, and assume zero Dirichlet boundary conditions. We recall that an

$N \times N$ matrix $\mathcal{D}$ is a pth-order *differentiation matrix* on this grid if

$$u'(x_m) = \sum_{k=1}^{N} \mathcal{D}_{m,k} u(x_k), \qquad m = 1, \ldots, N, \tag{2.1}$$

for every pth-degree polynomial u that vanishes at the endpoints.

Theorem 2 (necessary condition) *Consider a differentiation matrix $\mathcal{D}$ of order p corresponding to a grid $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$. A necessary condition for $\mathcal{D}$ to be skew-symmetric is that*

$$\sum_{k=1}^{N} f'(x_k) = 0 \qquad \textit{for} \qquad f(x) = x^2(1-x)^2 g(x) \tag{2.2}$$

and for all polynomials $g(x)$ of degree $2p-4$.

Proof We multiply (2.1) by $u(x_m)$ and sum up for $m = 1, \ldots, N$: since $\mathcal{D}$ is skew-symmetric, the outcome is

$$\sum_{m=1}^{N} u'(x_m) u(x_m) = \sum_{m=1}^{N} \sum_{k=1}^{N} u(x_m) \mathcal{D}_{m,k} u(x_k) = 0 \tag{2.3}$$

for an arbitrary polynomial $u(x)$ of degree $\leq p$ which vanishes at the endpoints. This proves the necessity of (2.2) for a square function $f(x) = u^2(x)$. For general $f(x) = x^2(1-x)^2 g(x)$ the statement follows from the fact that every polynomial $g(x)$ of degree $2p-4$ can be written as a linear combination of squares $v^2(x)$ with $\deg v \leq p-2$. This is a consequence of $2x = (x+1)^2 - x^2 - 1^2$. □

The polynomial $g(x)$ of degree $2p-4$ in (2.2) lives in the $(2p-3)$-dimensional linear space spanned by x^{s-1}, $s = 1, 2, \ldots, 2p-3$. The necessary condition of Theorem 2 can be written in this basis as

$$R_1 = R_2 = \cdots = R_{2p-3} = 0, \tag{2.4}$$

where

$$R_s = \sum_{k=1}^{N} x_k^s (1-x_k) \big[(s+1) - (s+3) x_k \big], \qquad s \in \mathbb{N}. \tag{2.5}$$

For order 2 the conditions (2.4) reduce to $\sum_{k=1}^{N} x_k(1-x_k)(1-2x_k) = 0$, and this is satisfied, for example, by any symmetric grid (i.e., $x_{N+1-m} = 1 - x_m$, $m = 1, \ldots, N$). Hence the stability of the simplest, second-order central-difference differentiation matrix. For $p = 3$ we may let $N = 2M$, $x_m = \rho m$, $x_{2M+1-m} = 1 - x_m = 1 - \rho m$, $m = 1, \ldots, M$, then, by virtue of symmetry, $R_1 = 0$ while $R_2 = R_3 = 0$ reduce to a single condition. Here $\rho = \mathcal{O}\big(N^{-1}\big)$ is a real zero of a certain cubic equation which has been already derived in (Iserles 2014) by a different argument. We will return to this example with greater detail in Section 4.

The fact that we need just $R_2 = 0$ in the above example is not a matter of serendipity.

Theorem 3 *Supposing that the grid is symmetric, $x_m + x_{N+1-m} = 1$ for $m = 1, \ldots, N$, the order conditions (2.2) (or equivalently (2.4)) reduce to*

$$R_2 = R_4 = \cdots = R_{2p-4} = 0. \tag{2.6}$$

Proof Reversing the order of summation, symmetry implies the condition (2.2) of Theorem 2 for polynomials satisfying $g(1-x) = g(x)$ and, in particular, for polynomials of the form $(x-1/2)^{2l}$. Considering the basis

$$\{x^{2l-1};\, l = 1, \ldots, p-2\} \cup \{(x-1/2)^{2l};\, l = 0, \ldots, p-2\}$$

for the linear space of polynomials of degree $2p-4$, it follows from Theorem 2 that we need impose order conditions only for even s and (2.6) follows. □

3 Sufficient conditions and banded matrices

In Theorem 2 we have established necessary order conditions for an existence of a pth-order skew-symmetric differentiation matrix on a given grid. Three issues arise and their analysis forms the body of this section. Firstly, are the conditions (2.2) sufficient? Secondly, bearing in mind Theorem 1, can they be realised by a banded matrix? Thirdly, assuming an affirmative answer to the first two questions, can we devise a constructive algorithm to derive a skew-symmetric band differentiation matrix of requisite order?

Theorem 4 (sufficient condition) *Consider a differentiation matrix $\mathcal{D}$ of order p corresponding to a grid $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$, and assume that $\mathcal{D}_{j,k} + \mathcal{D}_{k,j} = 0$ for all $1 \le \min\{j,k\} \le N-p+1$. A sufficient condition for $\mathcal{D}$ to be skew-symmetric is the condition (2.2) of Theorem 2.*

Proof Let $\tilde{u}$ be a polynomial of degree $\le p-2$ and set $u(x) = x(1-x)\tilde{u}(x)$. We then have

$$\begin{aligned} 0 = \sum_{m=1}^{N} u(x_m)u'(x_m) &= \sum_{m=1}^{N}\sum_{k=1}^{N} u(x_m)\mathcal{D}_{m,k}u(x_k) \\ &= \sum_{m=N-p+2}^{N}\sum_{k=N-p+2}^{N} u(x_m)\mathcal{D}_{m,k}u(x_k). \end{aligned} \tag{3.1}$$

The first equality follows from (2.2) with $f(x) = u^2(x)$, the second from the definition of order for a differentiation matrix, and the last because the first $N-p+1$ rows and columns of $\mathcal{D}$ are consistent with skew-symmetry.

Choose any $s \in \{N-p+2, \ldots, N\}$ and set

$$\tilde{u}(x) = \ell_s(x) = \prod_{\substack{j=N-p+2 \\ j\neq s}}^{N} \frac{x - x_j}{x_s - x_j},$$

the cardinal polynomial of Lagrangian interpolation at $x_{N-p+2}, \ldots, x_N$ such that $\tilde{u}(x_s) = 1$. Note that $\deg \tilde{u} = p-2$. Then

$$\sum_{m=N-p+2}^{N} \sum_{k=N-p+2}^{N} u(x_m)\mathcal{D}_{m,k}u(x_k) = x_s^2(1-x_s)^2\mathcal{D}_{s,s}$$

and we deduce from (3.1) that $\mathcal{D}_{k,k} = 0$ for $k = N-p+2, \ldots, N$. Next we choose $\tilde{u}(x) = \ell_q(x) + \ell_s(x)$ for distinct q and s in $\{N-p+2, \ldots, N\}$: again $\deg \tilde{u} = p-2$. It follows at once that

$$\sum_{m=N-p+2}^{N} \sum_{k=N-p+2}^{N} u(x_m)\mathcal{D}_{m,k}u(x_k) = x_q(1-x_q)x_s(1-x_s)(\mathcal{D}_{q,s} + \mathcal{D}_{s,q})$$

and we deduce from (3.1) that $\mathcal{D}_{q,s} + \mathcal{D}_{s,q} = 0$ for all q, s in $\{N-p+2, \ldots, N\}$. This demonstrates that the bottom $(p-2) \times (p-2)$ minor of $\mathcal{D}$ is skew symmetric. Since the remainder of $\mathcal{D}$ is consistent with skew-symmetry, the statement follows. □

Algorithm We present an algorithmic description for the construction of pth-order skew-symmetric differentiation matrices:

1. We commence the construction from the first row and set $\mathcal{D}_{1,1} = 0$. We have $N-1$ remaining parameters $\mathcal{D}_{1,k}$, $k = 2, \ldots, N$, and $p-1$ order conditions (2.1). Choose *arbitrarily* the $N-p$ entries $\mathcal{D}_{1,p+1}, \ldots, \mathcal{D}_{1,N}$ and treat the remaining ones, $\mathcal{D}_{1,2}, \ldots, \mathcal{D}_{1,p}$, as unknowns. The outcome is the Vandermonde linear algebraic system

   $$\sum_{k=2}^{p} \mathcal{D}_{1,k}x_k^s(1-x_k) = sx_1^{s-1} - (s+1)x_1^s - \sum_{k=p+1}^{N} \mathcal{D}_{1,k}x_k^s(1-x_k), \qquad s = 1, \ldots, p-1.$$

   It is nonsingular because the $\{x_k\}$ are distinct, hence it determines the rest of the first row of $\mathcal{D}$ uniquely. We let $\mathcal{D}_{k,1} = -\mathcal{D}_{1,k}$, $k = 2, \ldots, N$, so that the leading row and column are consistent with skew-symmetry.

2. We proceed next to the second row. $\mathcal{D}_{2,1}$ is already known, we let $\mathcal{D}_{2,2} = 0$ and choose arbitrarily the $N-2$ entries $\mathcal{D}_{2,p+2}, \ldots, \mathcal{D}_{2,N}$. This leaves us with $p-1$ entries which are uniquely determined by solving a non-singular Vandermonde system and which are subsequently extended to the second column of $\mathcal{D}$ by skew-symmetry.

3. In a similar manner, we sweep the first $N-p+1$ rows of $\mathcal{D}$ in progression. For every row s we already know $\mathcal{D}_{s,i}$ for $i \le s-1$, let $\mathcal{D}_{s,s} = 0$ and choose arbitrarily $\mathcal{D}_{s,p+s}, \ldots, \mathcal{D}_{s,N}$. This leaves us with a nonsingular Vandermonde system for the remaining $p-1$ elements of $\mathcal{D}_{s,k}$. We then define the sth column consistently with skew-symmetry.

4. We determine the remaining $(p-1)^2$ entries $\mathcal{D}_{j,k}$, $j,k = N-p+2, \ldots, N$ by applying the order conditions (2.1): we have exactly the right number of

equations and, in each row, a nonsingular Vandermonde system. Note that these entries range both to the right and to the left of the main diagonal. By virtue of Theorem 4, we are assured that the outcome, a pth-order differentiation matrix $\mathcal{D}$, is skew symmetric.

Note that instead of prescribing arbitrarily $\mathcal{D}_{s,p+s},\dots,\mathcal{D}_{s,N}$ in the sth row, one can prescribe any $N-p-s+1$ entries of the sth row lying right of the diagonal element. This construction principle therefore gives *all* skew-symmetric order-p differentiation matrices.

Corollary 1 (banded matrices) *For every $p\geq 2$ it is possible, subject to the conditions (2.2), to construct a skew-symmetric order-p differentiation matrix of bandwidth $2p-1$, i.e. such that $\mathcal{D}_{j,k}=0$ for $|j-k|\geq p$.*

Proof If in the above algorithm those entries of $\mathcal{D}$ which can be arbitrarily chosen, are all put to zero, we obtain a banded matrix of bandwidth $2p-1$. □

It is possible to construct explicitly banded matrices for orders 2 and 3. To this end, let

$$R_{s,m}=\sum_{k=1}^{m}x_k^s(1-x_k)\big[(s+1)-(s+3)x_k\big],\qquad s\in\mathbb{N},\quad m=1,\dots,N,$$

noting that, according to (2.5), $R_s=R_{s,N}$. The tridiagonal skew-symmetric matrix of order 2 has the form

$$\mathcal{D}_{m,m+1}=\frac{R_{1,m}}{2x_m(1-x_m)x_{m+1}(1-x_{m+1})},\qquad \mathcal{D}_{m+1,m}=-\mathcal{D}_{m,m+1}\tag{3.2}$$

for $m=1,\dots,N-1$, with the remaining entries nil. This, of course, must be accompanied by the condition $R_1=0$.

Likewise, stipulating $R_1=R_2=R_3=0$, the quindiagonal skew-symmetric matrix of order 3 has the form

$$\begin{aligned}\mathcal{D}_{m,m+1}&=\frac{R_{3,m}-2x_{m+1}R_{2,m}+x_{m+1}x_{m+2}R_{1,m}}{x_m(1-x_m)x_{m+1}(1-x_{m+1})(x_{m+1}-x_m)(x_{m+2}-x_{m+1})},\\ \mathcal{D}_{m,m+2}&=\frac{R_{3,m}-2x_mR_{2,m}+x_{m-1}x_mR_{1,m}}{x_m(1-x_m)x_{m+1}(1-x_{m+1})(x_m-x_{m-1})(x_{m+1}-x_m)}\end{aligned}\tag{3.3}$$

for $m=1,\dots,N-1$ and $m=1,\dots,N-2$ respectively, with the rest of the upper triangle filled with zeros and the lower triangle completed by skew-symmetry.

4 Size (sometimes) matters

To fit into the conditions of Theorem 1, a matrix $\mathcal{D}$ needs be skew-symmetric, banded and sufficiently small, consistently with the inequality

$$|\mathcal{D}_{k,\ell}|=|\mathcal{D}_{k,\ell}^{[N]}|\leq b^\star N,\qquad k,\ell=1,\dots,N,\tag{4.1}$$

where $\mathcal{D}^{[N]}$ is $N \times N$ and we need to consider all $N \gg 1$ over grids consistent with the first condition of that theorem. Theorem 1 is relevant to equations like convection–diffusion and Fokker–Planck, while size does not matter to stability in the solution of the diffusion equation. At turns out in the case of banded matrices from the last section, we can be assured of (4.1) only for $r = 1$ (tridiagonal differentiation matrices). Already for $r = 2$ the elements of $\mathcal{D}$ may increase too fast. Therefore we can be certain of stability, within the context of Theorem 1, only for order-2 tridiagonal differentiation matrices. However, further discussion makes it clear that the general picture is more complicated and stability of banded differentiation matrices of orders ≥ 3 requires further work.

To establish a connection between different grids, we assume the presence of a strictly-monotone function $g \in \mathrm{C}[0,1]$ such that $g(0) = 0$, $g(1) = 1$ and $x_m^{(N)} = g(\xi_m^{(N)})$ with $\xi_m^{(N)} = m/(N+1)$, $m = 0, \ldots, N+1$. Note that for a differentiable g condition 1 of Theorem 1 is satisfied with $\sigma_N \equiv \|g'\|_\infty / N$.

Second-order tridiagonal matrices. We commence from tridiagonal matrices of order 2 with entries given by (3.2): it follows at once from the definition of the Riemann integral that

$$R_{1,m} \approx 2(N+1) \int_0^{m/(N+1)} g(\tau)[1-g(\tau)][1-2g(\tau)]\,\mathrm{d}\tau$$

and (3.2) implies

$$\mathcal{D}_{m,m+1} \approx (N+1) \frac{\int_0^{\xi_m^{(N)}} g(\tau)[1-g(\tau)][1-2g(\tau)]\,\mathrm{d}\tau}{g(\xi_m^{(N)}) g(\xi_{m+1}^{(N)}) \left[1-g(\xi_m^{(N)})\right] \left[1-g(\xi_{m+1}^{(N)})\right]}.$$

Suppose that $\xi_n^{(N)} = m/(N+1) \to \xi \in [0, \frac{1}{2}]$ as $N \to \infty$. (If $\xi \in (\frac{1}{2}, 1]$ then all we need is to swap the role of m and $N+1-m$.) Then

$$\mathcal{D}_{m,m+1} \approx (N+1) \frac{\int_0^{\xi} g(\tau)[1-g(\tau)][1-2g(\tau)]\,\mathrm{d}\tau}{g^2(\xi)[1-g(\xi)]^2} = h(\xi)(N+1).$$

If $g(\xi) = g_1 \xi + \mathcal{O}(\xi^2)$, $g_1 > 0$, then $h(\xi) = 1/(2g_1) + \mathcal{O}(\xi)$ is bounded at the origin. Since h is clearly bounded in $(0, \frac{1}{2})$, it follows that $|\mathcal{D}_{m,m+1}|$ can be bounded by a multiple of N, uniformly in m and N, as required.

Third-order quindiagonal matrices. Similarly, we obtain (omitting the upper index N in $x_m^{(N)}$ and $\xi_n^{(N)}$)

$$\begin{aligned}
&R_{3,m} - 2x_{m+1}R_{2,m} + x_{m+1}x_{m+2}R_{1,m} \\
&\approx (N+1)\Bigg[\int_0^{\xi_m} g^3(\tau)[1-g(\tau)][4-6g(\tau)]\,\mathrm{d}\tau - 2g(\xi_m)\int_0^{\xi_m} g^2(\tau)[1-g(\tau)][3-5g(\tau)]\,\mathrm{d}\tau \\
&\quad + g^2(\xi_m)\int_0^{\xi_m} g(\tau)[1-g(\tau)][2-4g(\tau)]\,\mathrm{d}\tau\Bigg]
\end{aligned}$$

while

$$x_m(1-x_m)x_{m+1}(1-x_{m+1})(x_{m+1}-x_m)(x_{m+2}-x_{m+1}) \approx \frac{g^2(\xi_m)[1-g(\xi_m)]^2 g'^2(\xi_m)}{(N+1)^2}.$$

Therefore, by (3.3),

$$\begin{aligned}\mathcal{D}_{m,m+1} \approx \frac{(N+1)^3}{g^2(\xi_m)[1-g(\xi_m)]^2 g'^2(\xi_m)} \Bigg[&\int_0^{\xi_m} g^3(\tau)[1-g(\tau)][4-6g(\tau)]\,\mathrm{d}\tau \\ &- 2g(\xi_m)\int_0^{\xi_m} g^2(\tau)[1-g(\tau)][3-5g(\tau)]\,\mathrm{d}\tau \\ &+ g^2(\xi_m)\int_0^{\xi_m} g(\tau)[1-g(\tau)][2-4g(\tau)]\,\mathrm{d}\tau \Bigg].\end{aligned}$$

Hence, since the three integrals cannot vanish for all ξ_m, the size of $|\mathcal{D}_{m,m+1}|$ increases like $\mathcal{O}\big(N^3\big)$, much too fast!

The above calculation is not an artefact of rough estimates: this unwelcome growth of $|\mathcal{D}_{m,m+1}|$ (and of $|\mathcal{D}_{m,m+2}|$) occurs in specific examples. Let us commence from the uniform grid $x_m = m/(N+1)$, i.e. from $g(x) = x$: although it obeys (2.2) with just $p = 2$, this is a convenient point of departure because everything can be computed explicitly,

$$\begin{aligned}\mathcal{D}_{m,m+1} &= \frac{(N+1)(6N^2m - 12Nm^2 + 6m^3 + 6N^2 - 8Nm + m^2 + 2N - 5m - 2)}{6(N-m)(N-m+1)}, \\ \mathcal{D}_{m,m+2} &= -\frac{(N+1)(6N^2m - 12Nm^2 + 6m^3 + 8Nm - 7m^2 - 2N + 3m - 2)}{6(N-m)(N-m+1)}.\end{aligned}$$

In particular, $\mathcal{D}_{N-1,N} \approx -\frac{1}{12}N^3$, $\mathcal{D}_{N-2,N} \approx -\frac{1}{36}N^3$.

As we have just mentioned, the uniform grid $g(x) = x$ is compatible with order 2, as generically is the symmetric grid generated by $g(x) = ax + 3(1-a)x^2 - 2(1-a)x^3$ for any $a \in (0,3]$ – it is trivial to verify that g obeys all the necessary conditions. Of course, $a = 1$ corresponds to a uniform grid. However, simple calculation confirms that

$$\begin{aligned}a = a_N \sim 1 &- \frac{35}{3}\frac{1}{N^2} + \frac{70}{3}\frac{1}{N^3} + \frac{1820}{27}\frac{1}{N^4} - \frac{9800}{27}\frac{1}{N^5} - \frac{1600235}{2673}\frac{1}{N^6} + \frac{6600790}{891}\frac{1}{N^7} \\ &- \frac{393855490}{352741}\frac{1}{N^8} + \mathcal{O}\big(N^{-9}\big), \qquad N \gg 1,\end{aligned}$$

results in $R_1 = R_2 = R_3 = 0$, hence order-3 conditions. Although this renders g dependent on N, this does not change the essence of our analysis or its main conclusions. Originating in an $\mathcal{O}\big(N^{-2}\big)$ perturbation of the uniform grid, the order of increase of $|\mathcal{D}_{j,k}|$ is somewhat improved to $\mathcal{O}\big(N^2\big)$: this can be seen in a straightforward MATLAB computation and confirmed by an exact calculation for $a_N = 1 - \frac{35}{3}N^{-2}$ in MAPLE. Inasmuch as $\mathcal{O}\big(N^2\big)$ is better than $\mathcal{O}\big(N^3\big)$, it falls short of the linear growth required in Theorem 1.

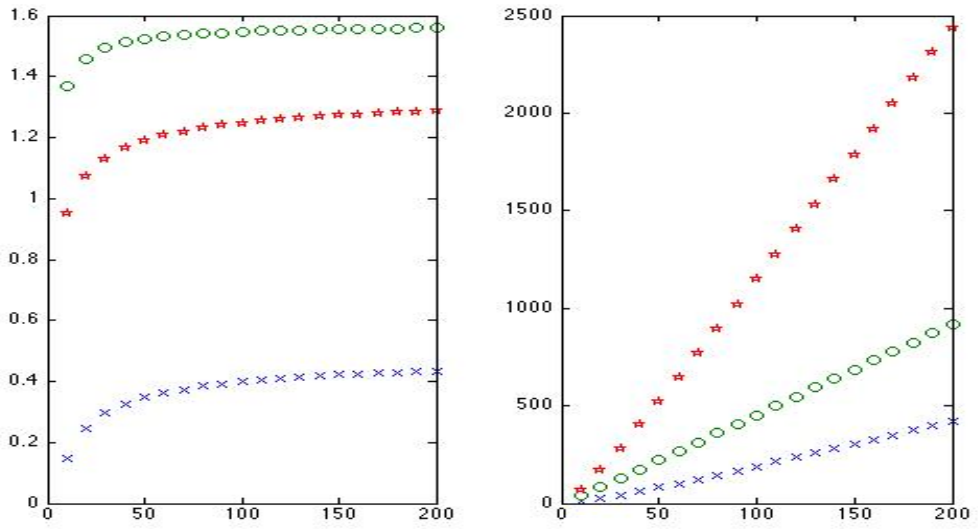

Figure 4.1: The logarithmic norm $\mu[\mathcal{VD}]$ as a function of N for three different functions V. On the left the results for a second-order tridiagonal differentiation matrix and on the right for a third-order quindiagonal one.

In Fig. 4.1 we display the logarithmic norm $\mu[\mathcal{VD}]$ for values of N between 10 and 200 and three functions V: $x^2(1-x)$ (crosses), e^x (pentagons) and $\cos \pi x$ (ovals). All the results were computed with the above symmetric, third-order grid. Recall that the entire point of the proof of Theorem 1 was in demonstrating that, for $N \gg 1$, $\mu[\mathcal{VD}]$ can be uniformly bounded. The results on the left correspond to tridiagonal matrices with $p = 2$, and we note that, indeed, logarithmic norms remain bounded – as a matter of fact, they rapidly tend to a limit. On the other hand, once we consider quindiagonal matrices with $p = 3$ the right, logarithmic norms grow linearly with N.

It is legitimate to query, however, how tight is (1.6), in other words how well does the exponential of the logarithmic norm bound from above the norm of a matrix exponential. The picture here is more mixed (and much more interesting!).

Thus, in Fig. 4.2 we have plotted the true norm of $\mathrm{e}^{\mathcal{VD}}$ under the same ground rules as in Fig. 4.1. The left plot, corresponding to a tridiagonal matrix, is unsurprising: the norms are uniformly bounded, tend to limits and everything is bounded fairly tightly by the logarithmic norm estimate (1.6). The surprise is on the right-hand side. For $V(x) = \cos \pi x$ the norm increases at an exponential speed (for $N = 10$ it equals 9.44×10^5, for $N = 100$ it is 8.14×10^{17} and for $N = 200$ the norm is 2.21×10^{33}) and has not been displayed in the figure. For $V(x) = x^2(1-x)$ the norm increases, albeit slowly, and it is premature even to guess its asymptotic behaviour. For $V(x) = \mathrm{e}^x$, however, the norm evidently tends to a bounded, fairly small limit, in the same ballpark as in the tridiagonal case. In other words, different functions V

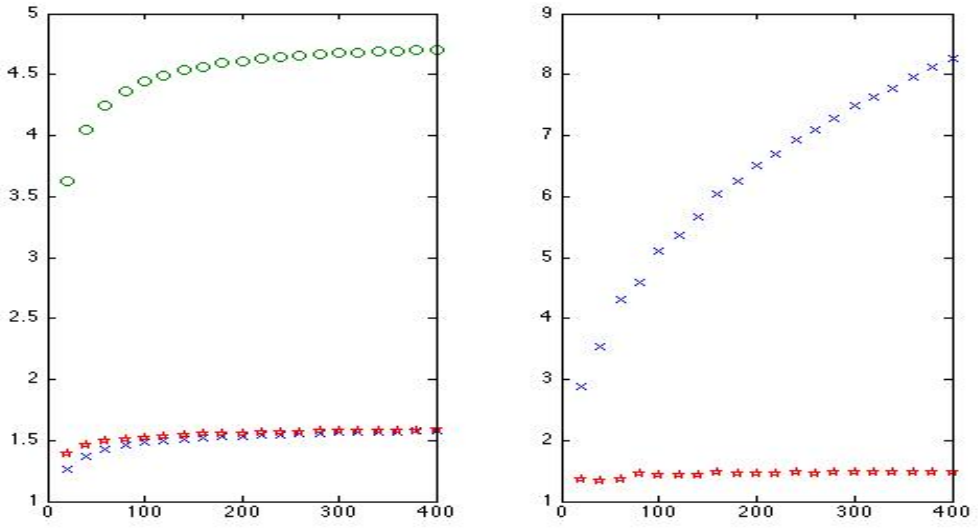

Figure 4.2: $\|\exp(\mathcal{V}\mathcal{D})\|$ as a function of N for three different functions V. On the left the results for a second-order tridiagonal differentiation matrix and on the right for a third-order quindiagonal one.

result in different asymptotic behaviour of $\|\mathrm{e}^{t\mathcal{V}\mathcal{D}}\|$, the logarithmic norm is a crude instrument and this entire issue calls for much more substantive investigation.

However, in place of the above grid we might use the construction from (Iserles 2014), with a *discontinuous,* N-dependent function g with a single $\mathcal{O}\big(N^{-1}\big)$ jump at $\frac{1}{2}$. Specifically (and correcting a misprint in (Iserles 2014)) we let $N = 2M$, set

$$x_m = \rho m, \quad x_{N+1-m} = 1 - \rho m, \qquad m = 0, \ldots, M,$$

and

$$\begin{aligned}
\mathcal{D}_{m,m+1} &= \frac{1}{6}\frac{[1-(2m+1)\rho][3-(2m+1)\rho]}{\rho(1-m\rho)[1-(m+1)\rho]}, \qquad m = 1, \ldots, M-2,\\
\mathcal{D}_{M-1,M} &= \frac{1}{12\rho}\cdot\frac{6-(26M-10)\rho+(35M-31)M\rho^2-(14M^2-19M+3)M\rho^3}{[(1-(M-1)\rho](1-M\rho)(1-2M\rho)},\\
\mathcal{D}_{M,M+1} &= \frac{M+1}{6M\rho}\cdot\frac{3-3(4M+1)\rho+(15M^2+7M+2)\rho^2-(6M^2+5M+1)M\rho^3}{(1-M\rho)^2(1-2M\rho)},\\
\mathcal{D}_{m,m+2} &= \frac{m+1}{12}\cdot\frac{2-(m+1)\rho}{(1-m\rho)[1-(m+2)\rho]}, \qquad m = 1, \ldots, M-2,\\
\mathcal{D}_{M-1,M+1} &= \frac{1}{12}\frac{(M+1)\rho(2-M\rho)}{[1-(M-1)\rho](1-M\rho)(1-2M\rho)},
\end{aligned}$$

with the remaining coefficients reflected by symmetry,

$$\mathcal{D}_{2M-m,2M-m+1} = \mathcal{D}_{m,m+1}, \qquad m = 1, \dots, M,$$
$$\mathcal{D}_{2M-m-1,2M-m+1} = \mathcal{D}_{m,m+2}, \qquad m = 1, \dots, M-1.$$

The choice of ρ consistent with order 3 is

$$\rho \approx \frac{1}{2M} - \frac{1}{4M^2} - \frac{5}{24M^3} + \mathcal{O}\big(M^{-4}\big)\,,$$

the unique real zero of the cubic

$$(2M+1)(3M^2+3M-1)\rho^3 - 15M(M+1)\rho^2 + 6(2M+1)\rho - 3.$$

Therefore

$$\mathcal{D}_{m,m+1} - \mathcal{D}_{m+1,m+2} = \frac{1}{3}\frac{1-\rho^2}{(1-m\rho)[1-(m+1)\rho][1-(m+2)\rho]} > 0,$$
$$\mathcal{D}_{m+1,m+3} - \mathcal{D}_{m,m+2} = \frac{1}{12}\frac{(1-\rho^2)[2-(2m+3)\rho]}{(1-m\rho)[1-(m+1)\rho][1-(m+2)\rho][1-(m+3)\rho]} > 0$$

for $m = 1, \dots, M-3$. It follows at once that all the $\mathcal{D}_{m,m+i}$s, $i = 1, 2$, are positive, the $\mathcal{D}_{m,m+1}$s decrease monotonically and the $\mathcal{D}_{m,m+2}$ increase monotonically in the first half of the range. Note that

$$\mathcal{D}_{1,2} \approx \frac{N}{2}, \qquad \mathcal{D}_{M-2,M-1} \approx \frac{16}{3}, \quad \mathcal{D}_{M-1,M} \approx \frac{35}{12}, \quad \mathcal{D}_{M,M+1} \approx \frac{11}{6}$$

and

$$\mathcal{D}_{1,3} = \frac{1}{3}\frac{1}{1-3\rho}, \qquad \mathcal{D}_{M-2,M} \approx \frac{N}{4} - \frac{7}{4}, \quad \mathcal{D}_{M-1,M+1} \approx \frac{N}{4} - \frac{13}{2}.$$

We deduce that the individual terms of the matrix are at most $\mathcal{O}(N)$ and all the conditions of Theorem 1 are satisfied. As a matter of record, the logarithmic norm $\mu[\mathcal{V}\mathcal{D}]$ and the exponential $\|\exp(\mathcal{V}\mathcal{D})\|$ are virtually indistinguishable from the left-hand side of Figs 4.1 and 4.2.

5 Grids and order conditions

It bears spelling out again the order-p conditions,

$$R_1 = R_2 = \cdots = R_{2p-3} = 0 \tag{5.1}$$

where

$$R_s = \sum_{k=1}^{N} \varphi_s'(x_k) \qquad \text{with} \qquad \varphi_s(x) = x^{s+1}(1-x)^2. \tag{5.2}$$

This is a system of $2p-3$ polynomial, separable equations in N variables: Since our interest is in stability for all sufficiently large N, we may assume $N \gg p$. Each $\varphi_s(x)$ is strictly convex in $(0,1)$, vanishes at the endpoints and $\varphi_s'(x^\star) = 0$ for $x^\star =$

$(s+1)/(s+3)$. In other words, $\varphi_s'(x_k) > 0$ for $x_k < x^\star$ and $\varphi_s'(x_k) < 0$ for $x_k > x^\star$ – the condition (5.1) for each s means that the x_ks need be distributed so that the increase of φ_s in $(0, x^\star)$ is somehow balanced by its decrease in $(x^\star, 1)$.

Our present concern is with the setting of the last section: We are given a strictly monotonically increasing, sufficiently smooth function g that maps $[0,1]$ to itself and let $x_m = g(m/(N+1))$, $m = 0, \dots, N$. In general, we can harbour little hope that the order conditions (5.1) hold since, in practical situation, the nature of the grid function g is determined by considerations like the variation of V, variation of the solution or presence of boundary and internal layers. Thus, we wish to explore the possibility of the existence of a *perturbed grid* $\{\tilde{x}_m\}_{m=0}^{N+1}$ such that $\tilde{R}_s = 0$, $s = 1, \dots, 2p-3$, where the x_ms are replaced by $\tilde{x}_m$s in $\tilde{R}_s$, and such that $\tilde{x}_m = x_m + \mathcal{O}(N^{-\alpha})$ for $m = 1, \dots, N$ and some $\alpha \geq 1$.

Lemma 5 *Consider a grid $\{x_m\}_{m=0}^{N+1}$ given by $x_m = g(m/(N+1))$, and let*

$$I_s[g] = \int_0^1 g^s(\tau)[1-g(\tau)][(s+1)-(s+3)g(\tau)]\,\mathrm{d}\tau, \qquad s \in \mathbb{N}. \tag{5.3}$$

A necessary condition for the existence of a perturbed grid $\tilde{x}_m = x_m + \mathcal{O}(N^{-\alpha})$ with $\alpha \geq 1$ to satisfy the order-p conditions is

$$I_1[g] = \cdots = I_{2p-3}[g] = 0. \tag{5.4}$$

Proof Using the Euler–Maclaurin formula

$$\sum_{k=0}^{N} f(k) = \int_0^N f(\tau)\,\mathrm{d}\tau + \tfrac{1}{2}[f(0)+f(N)] - \tfrac{1}{12}[f'(0)-f'(N)] + \cdots$$

(Abramowitz & Stegun 1964, p. 806), where $f \in \mathrm{C}^2[0,N]$, we have

$$\begin{aligned} R_s &= \sum_{m=0}^{N} g^s(\tfrac{m}{N+1})\left[1-g(\tfrac{m}{N+1})\right]\left[(s+1)-(s+3)g(\tfrac{m}{N+1})\right] \\ &= (N+1)I_s[g] - \tfrac{1}{12}\frac{1}{N+1}[g'(1)J_s[g](1) - g'(0)J_s[g](0)] + \mathcal{O}\big(N^{-2}\big), \end{aligned}$$

where

$$J[g](x) := s(s+1)g^{s-1}(x) - 2(s+1)(s+2)g^s(x) + (s+2)(s+3)g^{s+1}(x)$$

– the absence of the $\mathcal{O}(1)$ term is due to $g(0) = 0$, $g(1) = 1$. Suppose now that $\tilde{x}_m = x_m + c_m/N^\alpha$ where the c_ms are all $\mathcal{O}(1)$ in N – we can add higher order terms except that this renders the proof messier, yet conceptually identical. Therefore

$$\begin{aligned} \tilde{R}_s &= \sum_{m=1}^{N}\left[g(\tfrac{m}{N+1}) + \frac{c_m}{N^\alpha}\right]^s\left[1-g(\tfrac{m}{N+1}) - \frac{c_m}{N^\alpha}\right]\left[(s+1)-(s+3)g(\tfrac{m}{N+1}) - (s+3)\frac{c_m}{N^\alpha}\right] \\ &= R_s + \sum_{m=1}^{N}\left\{\frac{c_m}{N^\alpha}J_s[g](\tfrac{m}{N+1}) + \mathcal{O}\big(N^{-2\alpha}\big)\right\} = (N+1)I_s[g] + \mathcal{O}\Big(N^{-\min\{1,\alpha-1\}}\Big). \end{aligned}$$

In other words, and regardless of the choice of the perturbation $\{c_m\}$, we cannot force $\tilde{R}_s = 0$ in the case $I_s[g] \neq 0$. □

How restrictive is the condition $I_s[g] = 0$? It is obeyed *for every* $s \in \mathbb{N}$ for $g(x) = x$, the uniform grid, because $I_s[x] = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}x} x^{s+1}(1-x)^2 \,\mathrm{d}x = 0$. Moreover, trivially, $I_1[g] = 0$ whenever g is odd with respect to $x = \frac{1}{2}$, the case corresponding to a symmetric grid.

Lemma 6 *The identity (5.4) holds if and only if the inverse function* $h = g^{-1}$ *is orthogonal to* $\tilde{\mathrm{P}}_2, \tilde{\mathrm{P}}_3, \ldots, \tilde{\mathrm{P}}_{2p-2}$, *where* $\tilde{\mathrm{P}}_n$ *is the* n*th degree Legendre polynomial shifted to the interval* $[0,1]$, $\tilde{\mathrm{P}}_n(t) = \mathrm{P}_n(2t-1)$.

Proof Changing the variable $\eta = g(\tau)$, we have

$$I_s[g] = \int_0^1 \eta^s(1-\eta)[(s+1) - (s+3)\eta]h'(\eta)\,\mathrm{d}\eta$$

and integration by parts confirms that

$$I_s[g] = 0 \qquad \Leftrightarrow \qquad \int_0^1 \phi_s(\eta)h(\eta)\,\mathrm{d}\eta = 0, \tag{5.5}$$

where

$$\phi_s(\eta) = s(s+1)\eta^{s-1} - 2(s+1)(s+2)\eta^s + (s+2)(s+3)\eta^{s+1}, \qquad s \in \mathbb{N}.$$

Each ϕ_s is a polynomial of degree $s+1$, hence it is spanned by $\tilde{\mathrm{P}}_0, \tilde{\mathrm{P}}_1, \ldots, \tilde{\mathrm{P}}_{s+1}$. Moreover, it can be trivially checked that

$$\int_0^1 \phi_s(\eta)\,\mathrm{d}\eta = \int_0^1 \eta\phi_s(\eta)\,\mathrm{d}\eta = 0,$$

hence ϕ_s is orthogonal to $\tilde{\mathrm{P}}_0$ and $\tilde{\mathrm{P}}_1$. Consequently, the linear space spanned by $\{\phi_1, \ldots, \phi_{2p-3}\}$ is the same as that spanned by $\{\tilde{\mathrm{P}}_2, \ldots, \tilde{\mathrm{P}}_{2p-2}\}$. This implies that, by (5.5), the identity (5.4) is equivalent to the property that h is orthogonal to $\tilde{\mathrm{P}}_2, \ldots, \tilde{\mathrm{P}}_{2p-2}$. □

We have already shown that the uniform grid $g(x) = x$ yields $I_s[x] = 0$ for all $s \in \mathbb{N}$. Interestingly, it follows from the method of proof of Lemma 6 that it is the only smooth grid function with this feature. Specifically, suppose that $I_s[g] = 0$ for all $s \in \mathbb{N}$. Then h must be spanned by just $\tilde{\mathrm{P}}_0$ and $\tilde{\mathrm{P}}_1$ and the only linear function consistent with $h(0) = 1$ and $h(1) = 1$ is $h(x) = x$, hence $g(x) = x$.

Lemma 5 presents a necessary condition, $I_s[g] = 0$, $s = 1, \ldots, 2p-3$, for the existence of a perturbed grid. Is it sufficient? We cannot answer this question in its full generality and restrict our discussion to $p = 2$.

Theorem 7 *Let* $I_1[g] = 0$ *and* $x_m = g(m/(N+1))$, $m = 1, \ldots, N$. *Then there exists a grid* $0 < \tilde{x}_1 < \tilde{x}_2 < \cdots < \tilde{x}_N < 1$ *such that* $\tilde{R}_1 = 0$ *and* $\tilde{x}_m = x_m + \mathcal{O}\big(N^{-2}\big)$.

Proof We present an algorithm that constructs the grid $\{\tilde{x}_m\}$. Recall from the proof of Lemma 5 that $I_1[g] = 0$ implies that $R_1 = \mathcal{O}\big(N^{-1}\big)$. Let

$$F(x) = \int_0^x g(\xi)[1 - 6g(\xi) + 6g^2(\xi)]\,\mathrm{d}\xi, \qquad \xi \in [0,1].$$

Noting that $F \not\equiv 0$, choose any $x^\star \in (0,1)$ such that $F(x^\star) \neq 0$ and set $K = \lfloor (N+1)x^\star \rfloor$. We define

$$\tilde{x}_m = \begin{cases} (1-\alpha)x_m, & m = 1, \ldots, K, \\ x_m, & m = K+1, \ldots, N, \end{cases}$$

therefore

$$\tilde{R}_1 = R_1 - 2\alpha \sum_{m=1}^{K} x_m(1 - 6x_m + 6x_m^2) - 6\alpha^2 \sum_{m=1}^{K} x_m^2(1 - x_m) - 4\alpha^3 \sum_{m=1}^{K} x_m^3.$$

Any solution of this cubic renders $\tilde{R}_1$ equal to zero. Using Euler–Maclaurin,

$$\begin{aligned} \tilde{R}_1 = R_1 &- 2\alpha[KF(x^\star) + \mathcal{O}(1)] - 6\alpha^2 \left[K \int_0^{x^\star} g^2(\xi)[1 - g(\xi)]\,\mathrm{d}\xi + \mathcal{O}(1) \right] \\ &- 4\alpha^3 \left[K \int_0^{x^\star} g^3(\xi)\,\mathrm{d}\xi + \mathcal{O}(1) \right] \end{aligned}$$

and, bearing in mind that $R_1 = \mathcal{O}\big(N^{-1}\big)$, K is proportional to N, and $F(x^\star) \neq 0$, we deduce that the cubic equation has a solution

$$\alpha = \frac{R_1}{2KF(x^\star)} + \mathcal{O}\big(N^{-3}\big) = \mathcal{O}\big(N^{-2}\big)\,. \tag{5.6}$$

The theorem follows because

$$\tilde{x}_m - x_m = \begin{cases} -\alpha x_m = \mathcal{O}\big(N^{-2}\big)\,, & m = 1, \ldots, K, \\ 0, & m = K+1, \ldots, N. \end{cases}$$

□

The proof of the theorem reads like a constructive means to compute the grid in question but, to make it into a *practical* algorithm, we need to specify the optimal choice of $x^\star$. Clearly, the idea should be to maximise the size of the denominator in (5.6), to render $|\alpha|$ as small as possible. The function g being strictly monotone and $g'(0) > 0$, it follows at once that F' vanishes as the origin (where $F''(0) > 0$) and at two additional points in $(0,1)$, specifically at $h(\frac{1}{2} \pm \frac{\sqrt{3}}{6})$, where h is the inverse function of g: the first a maximum and the second a minimum. Ideally, we should choose one of these points, where $|F|$ is larger.

Actually, we can do even better: the entire construction is equally valid once we swap the endpoints, let $x^\star \in (0,1)$ be such that

$$F(x^\star) = \int_{x^\star}^1 [1 - g(\xi)][1 - 6g(\xi) + 6g^2(\xi)]\,\mathrm{d}\xi \neq 0,$$

let $K = \lceil (N+1)x^\star \rceil$,

$$\tilde{x}_m = \begin{cases} x_m, & m = 1, \dots, K-1, \\ (1-\alpha)x_m + \alpha, & m = K, \dots, N \end{cases}$$

and seek a zero of

$$\tilde{R}_1 = R_1 - 2\alpha \sum_{m=K}^{N} (1-x_m)(1-6x_m+x_m^2) - 6\alpha^2 \sum_{m=K}^{N} (1-x_m)^2(1-2x_m) - 6\alpha^3 \sum_{m=K}^{N} (1-x_m)^3.$$

Again, F has two critical points in $(0,1)$ – at exactly the same $x^\star$s as before! Hence, in principle, we can seek to minimise α out of a set of four possibilities. Note, incidentally, that an optimal α might well be negative and that, in general, the cubic seems to have three real solutions, two of which (needless to say) are not $\mathcal{O}(N^{-2})$.

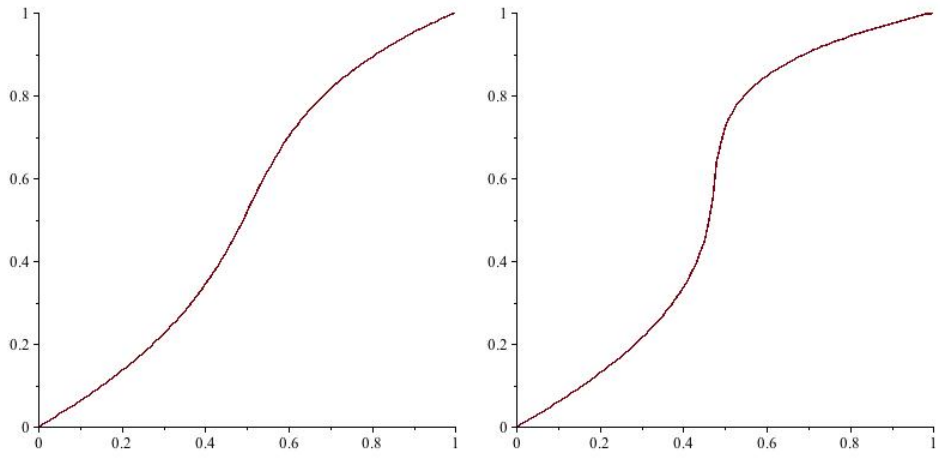

Figure 5.1: The functions g based upon $\tilde{h}(\eta) = \eta^4$ (on the left) and $\tilde{h}(\eta) = \eta^6$.

To flesh out numbers, we start from a strictly monotone function $\tilde{h} \in \mathrm{C}^2[0,1]$, $\tilde{h}(0) = 0$, $\tilde{h}(1) = 1$, and, in the spirit of the proof of Lemma 6, orthogonalise it with respect to $\tilde{\mathrm{P}}_2$,

$$h(\eta) = \tilde{h}(\eta) - [\tilde{\mathrm{P}}_2(\eta) - 1] \frac{\int_0^1 \tilde{h}(\kappa)\tilde{\mathrm{P}}_2(\kappa)\,\mathrm{d}\kappa}{\int_0^1 \tilde{\mathrm{P}}_2^2(\kappa)\,\mathrm{d}\kappa} = \tilde{h}(\eta) + 30\eta(1-\eta) \int_0^1 \tilde{h}(\kappa)(1-6\kappa+6\kappa^2)\,\mathrm{d}\kappa.$$

Of course, there is absolutely no guarantee that this h is monotone and this need be checked on a case-by-case basis. Moreover (and consistently with our former remark),

unless $\tilde{h}$ contains an even, non-constant component in its shifted Legendre expansion, the resulting h is symmetric. For example, letting $\tilde{h}(\eta) = \eta^3$ yields

$$h(\eta) = \tfrac{3}{2}\eta - \tfrac{3}{2}\eta^2 + \eta^3 = \tfrac{1}{2}\tilde{\mathrm{P}}_0(\eta) + \tfrac{9}{20}\tilde{\mathrm{P}}_1(\eta) + \tfrac{1}{20}\tilde{\mathrm{P}}_3(\eta)$$

– a monotone function with $h(\eta) + h(1-\eta) \equiv 1$, hence already associated with order $p \geq 2$. Moreover, $\tilde{h}(\eta) = \eta^4$ becomes

$$h(\eta) = \tfrac{12}{7}\eta - \tfrac{12}{7}\eta^2 + \eta^4 = \tfrac{17}{35}\tilde{\mathrm{P}}_0(\eta) + \tfrac{2}{5}\tilde{\mathrm{P}}_1(\eta) + \tfrac{1}{10}\tilde{\mathrm{P}}_3(\eta) + \tfrac{1}{70}\tilde{\mathrm{P}}_4(\eta),$$

which is monotone and fulfils all our conditions. (So does, say, $\tilde{h}(\eta) = \eta^6$, resulting in $h(\eta) = \frac{25}{14}\eta - \frac{25}{14}\eta^2 + \eta^6$, but not $\tilde{h}(\eta) = \eta^8$, which produces a non-monotone h.) All that remains now is to invert h and the outcome is a non-symmetric function g such that $I_1[g] = 0$. Although, by no stretch of imagination, this is a viable computational approach, at least it results in an existence proof of such a function g. Fig. 5.1 displays two such functions g.

To present a more detailed example, we commence from $\tilde{h}(\eta) = \eta \mathrm{e}^{3(\eta-1)}$, hence

$$h(\eta) = \eta \mathrm{e}^{-3}(\mathrm{e}^{3\eta} - 30\eta + 30),$$

a strictly monotone function. Both h and its inverse function g are displayed in Fig. 5.2.

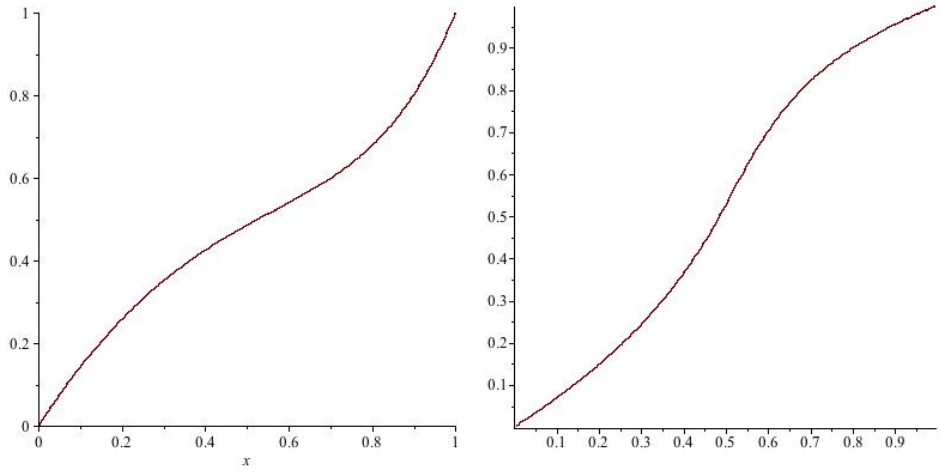

Figure 5.2: The functions h and g 'seeded' by $\tilde{h}(\eta) = \eta \mathrm{e}^{(3\eta-1)}$.

The two candidates for $x^\star$ are

$$x_1^\star = h(\tfrac{1}{2} - \tfrac{\sqrt{6}}{3}) \approx 0.268768817, \qquad x_2^\star = h(\tfrac{1}{2} + \tfrac{\sqrt{6}}{3}) \approx 0.667310687.$$

For each $x_j^\star$ we consider two options: summation on the left, denoted by $\boldsymbol{L}_j$, and summation on the right, denoted by $\boldsymbol{R}_j$. For $N = 200$ we have

Option	α	$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_\infty$
$\boldsymbol{L}_1$	-6.22333×10^{-5}	1.29×10^{-5}
$\boldsymbol{R}_1$	-9.78111×10^{-6}	7.71×10^{-6}
$\boldsymbol{L}_2$	$+9.96161 \times 10^{-6}$	7.80×10^{-6}
$\boldsymbol{R}_2$	$+5.07442 \times 10^{-5}$	1.08×10^{-5}

The differences, at least for this example, are fairly minor – indeed, taking the not-very-random $x^\star = \frac{1}{2}$ and the 'rightwise option' results in $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_\infty = 3.67 \times 10^{-5}$, which is not significantly larger. Thus, for all intents and purposes, the simplest option might well be to choose $x^\star$ at random, but the veracity of this statement would require further numerical experimentation.

The method of proof (and the algorithm) of Theorem 7 does not extend to $p = 3$. It might be seductive to take a symmetric grid g (hence, automatically, $I_1[g] = 0$) such that $I_3[g] = 0$ and repeat the same 'shift to one side' as in the proof to render $R_3 = 0$. This is certainly possible, except that the outcome is no longer a symmetric grid, hence for order $p = 3$ we require also $R_2 = 0$ and there is absolutely no reason why this should be the case – indeed, the procedure is likely to make R_1 nonzero.

6 Conclusions and pointers for future work

We have commenced in (Iserles 2014) the investigation of stability in the presence of variable coefficients. The work therein has been restricted to the convection-diffusion equation and symmetric grids. In the current paper we have ventured far beyond this narrow focus. Our current framework encompasses a much wider range of important linear partial differential equations of parabolic, hyperbolic and mixed type and, within the setting of finite-difference methods in one space dimension, it applies to all grids.

We have demonstrated in Theorem 1 that skew-symmetry, occasionally with additional conditions, implies stability, and this has motivated a detailed study of the relationship between order and grid for skew-symmetric differentiation matrices. The centrepiece of this investigation was the derivation of the order conditions (2.2). We have also proved that the very same order conditions suffice to ensure the existence of an pth-order method with the bandwidth $2p-1$, $p \geq 2$, except that the coefficients of such matrices might exhibit super-linear growth, thereby defying one of the side conditions of Theorem 1. Finally, we have discussed necessary and sufficient conditions for the existence of (possibly perturbed) grids consistent with skew-symmetric differentiation matrices of order $p \geq 2$.

In the remainder of this section we address a range of open problems following upon our work.

6.1 Order conditions on the grid

Open Problem 1 *Given $p \geq 3$ and a grid $\{x_m^{\rm old}\}_{m=0}^N$ in the interval $[0,1]$, does there exist another grid, $\{x_m^{\rm new}\}_{m=0}^N$, sufficiently near to the old grid (e.g., such that $x_m^{\rm new} = x_m^{\rm old} + \mathcal{O}\big(N^{-2}\big)$, $m = 1, \ldots, N$) which obeys the order conditions (5.1)?*

We already know from Theorem 7 the answer for $p = 2$ and, by virtue of Lemma 5, know that the necessary condition for all $p \geq 2$ is $I_s[g] = 0$, $s = 1, \ldots, 2p-3$. An

important aspect of Open Problem 1 is how well can we approximate any given grid using grid functions from the manifold

$$\{g \in \mathrm{C}^2 \,:\, g(0) = 0,\ g(1) = 1,\ g'(x) > 0 \text{ for } x \in [0,1],\ I_1[g] = \cdots = I_{2p-3}[g]\}.$$

As a matter of fact, it is entirely possible that the answer to Open Problem 1 is superfluous. Suppose that we are given a grid function g such that $I_s[g] = 0$, $s = 1, \ldots, 2p-3$, and we generate on the grid $x_m = g(m/(N+1))$ (e.g. by generalising the approach of Section 3) an order-p differentiation matrix $\mathcal{D}$. Such a matrix need no longer be skew-symmetric but, because $R_s = \mathcal{O}\big(N^{-1}\big)$, perhaps the very small departure from skew-symmetry might be forgivable in the context of Theorem 1 or in problems like the diffusion equation.

Open Problem 2 *Given a grid function g such that $I_s[g] = 0$, $s = 1, \ldots, 2p-3$, and letting $x_m = g(m/(N+1))$, $m = 0, \ldots, N$, does there exist a differentiation matrix $\mathcal{D}$ of order p such that $\mathcal{D} + \mathcal{D}^\top = \mathcal{O}\big(N^{-1}\big)$? Is it stable, e.g. in the sense of Theorem 1?*

The rationale underlying these two open problems is that, in general, the distribution of points on a grid may depend on a number of extraneous factors, e.g. the diffusion coefficient a in the diffusion equation, the potential V in the Liouville equation or the presence of boundary and internal layers in the convection-diffusion equation. In other words, we cannot choose the grid just for the convenience of designing a skew-symmetric differentiation matrix of requisite order. However, we may perturb an existing grid a little bit without distorting its essential shape, so as to obey the order conditions (5.1).

Of course, sheer existence of a grid consistent with the open problem falls short of the requirements of numerical mathematics: we need to compute it. Hence

Open Problem 3 *Provided that a grid sought in Open Problem 1 exists, compute it rapidly, i.e. at a computational cost significantly smaller than the cost of time-stepping the underlying PDE algorithm.*

We have described an easy-to-implement and cheap algorithm for $p = 2$ in Section 5, but the case $p \geq 3$ is open.

6.2 Size of the exponential

How large is the matrix exponential? This question admits a multitude of answers. Some, e.g. (Benzi & Razouk 2007/08, Iserles 2000), are concerned with the size of individual coefficients, hence are irrelevant to our narrative. Our concern is with uniform bounds of $\|\mathrm{e}^{t\mathcal{A}_N}\|_N$ in the Euclidean norm for an infinite family of $N \times N$ matrices, $N \to \infty$. This subject has already received a great deal of attention and is replete with beautiful and insightful results: from the *Kreiss Matrix Theorem,* namely that $\|\mathrm{e}^{t\mathcal{A}_N}\|_N \leq c$ if and only if $\|(\lambda\mathcal{I}_N - \mathcal{A}_N)^{-1}\|_N \leq c_1/(|\lambda| - 1)$ for all $\lambda \in \mathbb{C}$, $|\lambda| > 1$, where $\mathcal{I}_N$ is the $N \times N$ identity (Kreiss 1962, Strikwerda & Wade 1997) to the theory of ε-pseudospectra (Reddy & Trefethen 1992, Trefethen & Embree 2005), to estimates using the logarithmic norm *á la* (1.6). Yet, all said and done, it is clear that all these upper bounds, useful as they might be, are often much too conservative.

Although this excludes many bad and unsuitable methods (and it is always better to err on the side of caution), it is bound to exclude some good methods as well – we refer here to the discussion in Section 4 for an example. Approximating the operator $V(x)\cdot\partial/\partial x$ by the matrix product $\mathcal{V}\mathcal{D}$ to given order p, it is entirely possible that the first third-order, quindiagonal method from Section 3, say, might be stable for one V, unstable for another. In other words, the potential $V\in\mathrm{C}^1[0,1]$ might influence the choice of the grid not just according to a rough rule of a thumb of making the grid finer when $|V'|$ is larger, e.g.

$$(x_{m+1}-x_m)|V'(\tfrac{1}{2}(x_m+x_{m+1}))|\approx\text{const},$$

but also so as to produce a bounded exponential. This, of course, is by this stage no more than pure speculation.

Open Problem 4 *Explore further the norm of the matrix exponential, in particular in cases when current upper bounds are clearly excessive.*

Of course, it might be entirely possible to derive 'good' (i.e. growing not faster that $\mathcal{O}(N)$) skew-symmetric differentiation matrices like the second third-order method from Section 4 for $p\geq 4$: such methods can be applied stably, consistently with Theorem 1, with arbitrary potentials.

6.3 Skew-symmetry and time-stepping algorithms

The central assertion of this paper is that a single feature of a differentiation matrix, its skew-symmetry, is fundamental to its stability in a wide range of different linear PDEs of evolution with variable coefficients. The list is clearly non-exhaustive and it is of interest to investigate other linear PDEs with this feature. Even more important is to look into *nonlinear* PDEs of evolution which can be discretised in this manner. In that case it makes sense to couple the ideas of this paper with the use of different versions of operatorial splitting, e.g. the *Strang splitting* (Iserles 2008), *exponential integrators* (Hochbruck & Ostermann 2010) or IMEX-type methods (Kassam & Trefethen 2005). Consider, for example, the *reaction–diffusion equation*

$$\frac{\partial u}{\partial t}=\boldsymbol{\nabla}^{\top}a(\boldsymbol{x})\boldsymbol{\nabla}u+f(u), \tag{6.1}$$

with suitable initial and (for simplicity) zero Dirichlet boundary conditions. While we can use skew-symmetric matrices to discretise the Laplace–Beltrami operator $\mathcal{L}_1=\boldsymbol{\nabla}^{\top}a(\boldsymbol{x})\boldsymbol{\nabla}$, the nonlinear function $\mathcal{L}_2=f$ is outside the scope of our theory. However,

$$u(\,\cdot\,,t)=\mathrm{e}^{t(\mathcal{L}_1+\mathcal{L}_2)}u(\,\cdot\,,0)=\mathrm{e}^{\frac{1}{2}t\mathcal{L}_1}\mathrm{e}^{t\mathcal{L}_2}\mathrm{e}^{\frac{1}{2}t\mathcal{L}_1}u(\,\cdot\,,0)+\mathcal{O}\big(t^3\big)$$

(the Strang splitting), where $\mathrm{e}^{t\mathcal{L}}$ is formally the evolution operator of the equation $\partial u/\partial t=\mathcal{L}(u)$. This gives rise to a second-order (in Δt) time-stepping method that separates the solutions of $\partial u/\partial t=\boldsymbol{\nabla}^{\top}a\boldsymbol{\nabla}u$ and of $\partial u/\partial t=f(u)$ – the first can be dealt with by our approach and the second is an ODE. Many other equations fit into the same pattern as (6.1).

Potential relevance of splitting methods to our narrative is actually much greater: we can split, for example, *inside* the Laplace–Beltrami operator

$$\boldsymbol{\nabla}^\top a(\boldsymbol{x})\boldsymbol{\nabla} = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} a(\boldsymbol{x}) \frac{\partial u}{\partial x_i} = \sum_{i=1}^{d} \mathcal{L}^{[i]} u$$

(dimensional splitting) and follow with a multivariate version of Strang,

$$\mathrm{e}^{t(\mathcal{L}^{[1]}+\cdots+\mathcal{L}^{[d]})} = \mathrm{e}^{\frac{1}{2}t\mathcal{L}^{[1]}} \cdots \mathrm{e}^{\frac{1}{2}t\mathcal{L}^{[d-1]}} \mathrm{e}^{t\mathcal{L}^{[d]}} \mathrm{e}^{\frac{1}{2}t\mathcal{L}^{[d-1]}} \cdots \mathrm{e}^{\frac{1}{2}t\mathcal{L}^{[1]}} + \mathcal{O}\big(t^3\big)\,.$$

The main advantage of this approach is that each single one-dimensional evolution operator can be computed very cheaply and the cost increases *linearly* with the number of grid points.

Once (operatorial, dimensional or both) Strang splitting is used with the equation (6.1), the outcome is a second-order method. As long as we are allowed to use a uniform grid, there is no need for the work of this paper. However, once it becomes advantageous to use a nonuniform grid, it must satisfy the appropriate order conditions $R_1 = 0$.

The situation is, actually, more vexing, because approximating $\mathcal{L}_1$ by $\mathcal{DAD}$, where $\mathcal{D}$ is skew symmetric, is bound to double the bandwidth of $\mathcal{D}$ and introduce numerous zeros inside the band. This makes the methods more expensive and, even more importantly, introduces spurious oscillations in matrix exponentials. This unwelcome effect can be overcome by the use of *staggered grids,* resulting in the approximation $\mathcal{EAD}$, where $\mathcal{E} + \mathcal{D}^T = O$, but this belongs to another paper.

Within the context of parabolic operators like Laplace–Beltrami, no splittings of order ≥ 2 can coexist with stability (Sheng 1989). However, dimensional splitting is equally valuable in the context of hyperbolic operators, e.g. $-\boldsymbol{V}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}$, a situation when it is possible to design splittings of arbitrarily high order (McLachlan & Quispel 2002). In that case further investigation of the order conditions (2.2), implicit in Open Problems 1 and 2, becomes imperative.

Another approach, the *symmetric Zassenhaus splitting*, using commutators of differential operators, has been pioneered in (Bader et al. 2014) in the context of the semiclassical Schrödinger equation. In principle, similar approach might be relevant to many other hyperbolic PDEs. This is an alternative avenue leading towards stable splittings of high order, further underscoring the importance of these two open problems.

This brings us to the last major issue for further investigation. We have started this paper with full generality, considering all kinds of time-stepping methods for PDEs of evolution. After a while, we have restricted the field to 'nodal methods', whose unknowns are function values (and possibly derivatives) at the vertices of a grid. This excludes, for example, spectral methods, where the unknowns are expansion coefficients in an orthogonal basis of the underlying function space. Subsequently, we have lost even more generality, confining our attention just to finite difference methods in cubes, with tensor-product grids, a straightforward generalisation of the univariate case.

Open Problem 5 *Generalise the work of this paper to other time-stepping methods.*

The phrase "other time-stepping methods" refers to a wide range of different algorithms, each with its own complexities. Thus, the innocent phrase 'finite elements' hides a great deal of additional issues: different tessellations, conformal *vs* non-conformal elements, discontinuous Galerkin, spectral elements, unstructured meshes, *hp*-elements, ... The interplay between space and time discretisation, inclusive of the role of skew-symmetry in bringing about numerical stability, is likely to require a great deal of further specialised research.

It seems that for most of its history numerical analysis of PDEs has advanced along parallel tracks. Most of the community was concerned with the discretisation of the 'steady-state' part of the equation, and this has led to impressive advances, e.g. in finite element theory and in spectral methods. Another part of the community has been concerned with the time evolution, often adopting ideas from the theory of numerical ordinary differential equations (ODEs). It is clear, however, that space and time should not be discretised in isolation. We must fashion time-stepping methods not just by adopting known ODE solvers but by developing bespoke methods to cope with specific PDEs, like in (Hochbruck & Ostermann 2010) and (Shu & Osher 1988). Likewise – and this is the main message of this paper and, indeed, of the entire history of numerical stability – we need often to rethink our space discretisation to render time-stepping stable.

References

Abramowitz, M. & Stegun, I., eds (1964), *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC.

Bader, P., Iserles, A., Kropielnicka, K. & Singh, P. (2014), 'Effective approximation for the semiclassical Schrödinger equation', *Found. Comput. Maths* **14**, 689–720.

Benzi, B. & Razouk, N. (2007/08), 'Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices', *Electron. Trans. Numer. Anal.* **28**, 16–39.

Gustafsson, B., Kreiss, H.-O. & Sundström, A. (1972), 'Stability theory of difference approximations for mixed initial boundary value problems. II', *Maths Comp.* **26**, 649–686.

Hairer, E., Lubich, C. & Wanner, G. (2006), *Geometric Numerical Integration*, 2nd edn, Springer-Verlag, Berlin.

Hochbruck, M. & Ostermann, A. (2010), 'Exponential integrators', *Acta Numerica* **19**, 209–286.

Horn, R. A. & Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

Iserles, A. (2000), 'How large is the exponential of a banded matrix?', *J. New Zealand Maths Soc.* **29**, 177–192.

Iserles, A. (2008), *A First Course in the Numerical Analysis of Differential Equations*, 2nd edn, Cambridge University Press, Cambridge.

Iserles, A. (2014), 'On skew-symmetric differentiation matrices', *IMA J. Num. Anal.* **34**, 435–451.

Kassam, A.-K. & Trefethen, L. N. (2005), 'Fourth-order time-stepping for stiff PDEs', *SIAM J. Sci. Comput.* **26**, 1214–1233.

Kitson, A., McLachlan, R. I. & Robidoux, N. (2003), 'Skew-adjoint finite difference methods on nonuniform grids', *New Zealand J. Maths* **32**, 139–159.

Kreiss, H.-O. (1962), 'Über die stabilitätsdefinition für differenzengleichungen die partielle differentialgleichungen approximieren', *BIT* **2**, 153–181.

McLachlan, R. I. & Quispel, G. R. W. (2002), 'Splitting methods', *Acta Numerica* **11**, 341–434.

Reddy, S. & Trefethen, L. N. (1992), 'Stability of the method of lines', *Numer. Math.* **62**, 235–267.

Richtmyer, R. D. & Morton, K. W. (1967), *Difference Methods for Initial-Value Problems*, 2nd edn, Wiley-Interscience, New York.

Sheng, Q. (1989), 'Solving linear partial differential equations by exponential splitting', *IMA J. Numer. Anal.* **9**, 199–212.

Shu, C.-W. & Osher, S. (1988), 'Efficient implementation of essentially non-oscillatory shock-capturing schemes', *J. Comput. Phys.* **77**, 439–471.

Söderlind, G. (2006), 'The logarithmic norm. History and modern theory', *BIT* **46**, 631–652.

Strikwerda, J. C. & Wade, B. A. (1997), A survey of the Kreiss matrix theorem for power bounded families of matrices and its extensions, *in* 'Linear Operators', Banach Center Publ., pp. 339–360.

Trefethen, L. N. (1983), 'Group velocity interpretation of the stability theory of Gustafsson, Kreiss, and Sundström', *J. Comput. Phys.* **49**, 199–217.

Trefethen, L. N. & Embree, M. (2005), *Spectra and Pseudospectra. The Behavior of Nonnormal Matrices and Operators*, Princeton Univ. Press, Princeton, NJ.