J Sci Comput (2018) 75:906–940
https://doi.org/10.1007/s10915-017-0569-6

A Dual Consistent Finite Difference Method with Narrow Stencil Second Derivative Operators

Sofia Eriksson[1]

Received: 18 November 2016 / Revised: 24 August 2017 / Accepted: 25 September 2017 /
Published online: 16 October 2017

Abstract We study the numerical solutions of time-dependent systems of partial differential equations, focusing on the implementation of boundary conditions. The numerical method considered is a finite difference scheme constructed by high order summation by parts operators, combined with a boundary procedure using penalties (SBP–SAT). Recently it was shown that SBP–SAT finite difference methods can yield superconvergent functional output if the boundary conditions are imposed such that the discretization is dual consistent. We generalize these results so that they include a broader range of boundary conditions and penalty parameters. The results are also generalized to hold for narrow-stencil second derivative operators. The derivations are supported by numerical experiments.

Keywords Finite differences · Summation by parts · Simultaneous approximation term · Dual consistency · Superconvergence · Narrow stencil

Mathematics Subject Classification 65M06 · 65M20 · 65M12

1 Introduction

In this paper we consider a summation by parts (SBP) finite difference method, which is combined with a penalty technique denoted simultaneous approximation term (SAT) for the boundary conditions. The main advantages of the SBP–SAT finite difference methods are high accuracy, computational efficiency and provable stability. For a background on the history and the newer developments of SBP–SAT, see [6,19].

A discrete differential operator D_1 is said to be a SBP-operator if it can be factorized by the inverse of a positive definite matrix H and a difference operator Q, as specified later in Eq. (12). When H is diagonal, D_1 consists of a $2p$-order accurate central difference

✉ Sofia Eriksson
eriksson@mathematik.tu-darmstadt.de

[1] Department of Mathematics, Technische Universität Darmstadt, Dolivostr. 15, 64293 Darmstadt, Germany

approximation in the interior, but at the boundaries, the accuracy is limited to pth order. The global accuracy of the numerical solution can then be shown to be $p + 1$, see [18,19].

In many applications functionals are of interest, sometimes they are even more important than the primary solution itself (one example is lift or drag coefficients in computational fluid dynamics). It could be expected that functionals computed from the numerical solution would have the same order of accuracy as the solution itself. However, recently Hicken and Zingg [9] showed that when computing the numerical solution in a dual consistent way, the order of accuracy of the output functional is higher than that of the solution, in fact, the full $2p$ accuracy can be recovered. Related papers are [8,10] which includes interesting work on SBP operators as quadrature rules and error estimators for functional errors. Note that this kind of superconvergent behavior was already known for example for finite element and discontinuous Galerkin methods, but it had not been proven for finite difference schemes before, see [9]. Later Berg and Nordström [1–3] showed that the results hold also for time-dependent problems.

In [8,9] and [1] boundary conditions of Dirichlet type are considered (in [9] Neumann boundary conditions are included but are rewritten on first order form), and in [2,3] boundary conditions of far-field type are derived. In this paper, we generalize these results by deriving penalty parameters that yield dual consistency for all energy stable boundary conditions of Robin type (including the special cases Dirichlet and Neumann). In contrast to [2,3], where the boundary conditions were adapted to get the penalty in a certain form, we adapt the penalty after the boundary conditions instead. Furthermore, we extend the results such that they hold also for narrow-stencil second derivative operators (sometimes also denoted compact second derivative operators), where the term narrow is used to define explicit finite difference schemes with a minimal stencil width. In fact, the results even carry over to narrow-stencil second derivatives operators for variable coefficients (of the type considered for example in [12]).

To keep things simple we consider linear problems in one spatial dimension, however, note that this is not due to a limitation of the method. In [8,9] the extension to higher dimensions, curvilinear grids and non-linear problems are discussed and implemented for stationary problems and in [3] the theory is applied to the time-dependent Navier–Stokes and Euler equations in two dimensions.

The paper is organized as follows: In Sect. 2 we consider hyperbolic systems of partial differential equations and derive a family of SAT parameters which guarantees a stable and dual consistent discretization. Since higher order differential equations can always be rewritten as first order systems, this result directly leads to penalty parameters for parabolic problems, when using wide-stencil second derivative operators. Next, these parameters are generalized such that they hold also for narrow-stencil second derivative operators. This is all done in Sect. 3. In Sect. 4 a special aspect of the stability for the narrow operators is discussed. The derivations are then followed by examples and numerical simulations in Sect. 5 and a summary is given in Sect. 6.

1.1 Preliminaries

We consider time-dependent partial differential equations (PDE) as

$$\mathcal{U}_t + \mathcal{L}(\mathcal{U}) = \mathcal{F}, \qquad t \in [0, T], \qquad x \in \Omega, \tag{1}$$

where $\mathcal{L}$ represents a linear, spatial differential operator and $\mathcal{F}(x, t)$ is a forcing function. For simplicity, we will assume that the sought solution $\mathcal{U}(x, t)$ satisfies homogeneous initial and boundary conditions. To derive the dual equations we follow [1,2,9] and pose the problem

in a variational framework: Given a functional $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle$, where $\mathcal{G}(x,t)$ is a smooth weight function and where $\langle \mathcal{G}, \mathcal{U} \rangle = \int_\Omega \mathcal{G}^T \mathcal{U} \, \mathrm{d}x$ refers to the standard L^2 inner product, we seek a function $\mathcal{V}(x,t)$ such that $\mathcal{J}(\mathcal{U}) = \mathcal{J}^*(\mathcal{V}) = \langle \mathcal{V}, \mathcal{F} \rangle$. This defines the dual problem as

$$\mathcal{V}_\tau + \mathcal{L}^*(\mathcal{V}) = \mathcal{G}, \qquad \tau \in [0,T], \qquad x \in \Omega, \tag{2}$$

where $\mathcal{L}^*$ is the adjoint operator, given by $\langle \mathcal{V}, \mathcal{L}\mathcal{U} \rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U} \rangle$, and where $\mathcal{V}$ also satisfies homogeneous initial and boundary conditions. Note that the dual problem actually goes "backward" in time; the expression in (2) is obtained using the transformation $\tau = T - t$.

Let U and V be discrete vectors approximating $\mathcal{U}$ and $\mathcal{V}$, respectively, and let F and G be projections of $\mathcal{F}$ and $\mathcal{G}$ onto a spatial grid. We discretize (1) using a stable and consistent SBP–SAT scheme, leading to

$$U_t + LU = F, \qquad t \in [0,T]. \tag{3}$$

The SBP–SAT scheme has an associated matrix H which defines a discrete inner product, as $\langle G, U \rangle_H = G^T H U$ (when $\mathcal{U}$ is vector-valued, H must be replaced by $\overline{H}$, which is defined later in the paper). Now the discrete adjoint operator is given by $L^* = H^{-1} L^T H$, since this leads to $\langle V, LU \rangle_H = \langle L^* V, U \rangle_H$ which mimics the continuous relation above.

If L^* happens to be a consistent approximation of $\mathcal{L}^*$, then the discretization (3) is said to be *dual consistent* (if considering the stationary case) or *spatially dual consistent*, see [1,9] respectively. When (3) is a stable and dual consistent discretization of (1), then the linear functional $J(U) = \langle G, U \rangle_H$ is a $2p$-order accurate approximation of $\mathcal{J}(\mathcal{U})$, that is $J(U) = \mathcal{J}(\mathcal{U}) + \mathcal{O}(h^{2p})$, and we thus have superconvergent functional output. To obtain such high accuracy it is necessary to have compatible and sufficiently smooth data, see [9] for more details.

2 Hyperbolic Systems

We start by considering a hyperbolic system of PDEs of reaction-advection type, namely

$$\begin{aligned} \mathcal{U}_t + \mathcal{R}\mathcal{U} + \mathcal{A}\mathcal{U}_x &= \mathcal{F}, & x &\in [x_L, x_R], \\ \mathcal{B}_L \mathcal{U} &= g_L, & x &= x_L, \\ \mathcal{B}_R \mathcal{U} &= g_R, & x &= x_R, \end{aligned} \tag{4}$$

valid for $t \geq 0$ and augmented with initial data $\mathcal{U}(x,0) = \mathcal{U}_0(x)$. We let $\mathcal{R}$ and $\mathcal{A}$ be real-valued, symmetric $n \times n$ matrices with constant coefficients. Further, $\mathcal{R}$ is positive semi-definite, that is $\mathcal{R} \geq 0$. The operators $\mathcal{B}_L$ and $\mathcal{B}_R$ define the form of the boundary conditions and their properties are specified in (10) below. The forcing function $\mathcal{F}(x,t)$, the initial data $\mathcal{U}_0(x)$ and the boundary data $g_L(t)$ and $g_R(t)$ are assumed to be compatible and sufficiently smooth such that the solution $\mathcal{U}(x,t)$ exists. We will refer to (4) as our primal problem.

2.1 Well-Posedness Using the Energy Method

We call (4) well-posed if it has a unique solution and is stable. Existence is guaranteed by using the right number of boundary conditions, and uniqueness then follows from the stability, [7,15]. Next we show stability, using the energy method.

The PDE in the first row of (4) is multiplied by $\mathcal{U}^T$ from the left and integrated over the domain $\Omega = [x_L, x_R]$. Using integration by parts we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathcal{U}\|^2 + 2\langle \mathcal{U}, \mathcal{R}\mathcal{U}\rangle = 2\langle \mathcal{U}, \mathcal{F}\rangle + \mathrm{BT}_L + \mathrm{BT}_R \tag{5}$$

where $\|\mathcal{U}\|^2 = \langle \mathcal{U}, \mathcal{U}\rangle = \int_{x_L}^{x_R} \mathcal{U}^T\mathcal{U}\,\mathrm{d}x$ and where

$$\mathrm{BT}_L = \mathcal{U}^T\mathcal{A}\mathcal{U}\Big|_{x_L}, \qquad \mathrm{BT}_R = -\mathcal{U}^T\mathcal{A}\mathcal{U}\Big|_{x_R}.$$

To bound the growth of the solution, we must ensure that the boundary conditions make BT_L and BT_R non-positive for zero data. We consider the matrix $\mathcal{A}$ above and assume that we have found a factorization such that

$$\mathcal{A} = Z\Lambda Z^T, \qquad \Lambda = \begin{bmatrix} \Lambda_+ & & \\ & \Lambda_0 & \\ & & \Lambda_- \end{bmatrix}, \qquad Z = \begin{bmatrix} Z_+ & Z_0 & Z_- \end{bmatrix}, \tag{6}$$

where Z is non-singular. The parts of Λ are arranged such that $\Lambda_+ > 0$, $\Lambda_0 = 0$ and $\Lambda_- < 0$. According to Sylvester's law of inertia, the matrices $\mathcal{A}$ and Λ have the same number of positive (n_+), negative (n_-) and zero (n_0) eigenvalues (for a non-singular Z), where $n = n_+ + n_0 + n_-$. To bound the terms BT_L and BT_R, we have to give n_+ boundary conditions at $x = x_L$ and n_- boundary conditions at $x = x_R$. We note that

$$\mathcal{A} = Z_+\Lambda_+Z_+^T + Z_-\Lambda_-Z_-^T, \tag{7}$$

which gives

$$\begin{aligned} \mathrm{BT}_L &= \mathcal{U}^T\left(Z_+\Lambda_+Z_+^T + Z_-\Lambda_-Z_-^T\right)\mathcal{U}\Big|_{x_L}, \\ \mathrm{BT}_R &= -\mathcal{U}^T\left(Z_+\Lambda_+Z_+^T + Z_-\Lambda_-Z_-^T\right)\mathcal{U}\Big|_{x_R} \end{aligned}$$

where $Z_+^T\mathcal{U}$ represents the right-going variables (ingoing at the left boundary), and $Z_-^T\mathcal{U}$ represents the left-going variables (ingoing at the right boundary). The ingoing variables are given data in terms of known functions and outgoing variables, as

$$Z_+^T\mathcal{U}\Big|_{x_L} = \widetilde{g}_L - R_L Z_-^T\mathcal{U}\Big|_{x_L}, \qquad Z_-^T\mathcal{U}\Big|_{x_R} = \widetilde{g}_R - R_R Z_+^T\mathcal{U}\Big|_{x_R}, \tag{8}$$

where $\widetilde{g}_L$, $\widetilde{g}_R$ are the known data and where the matrices R_L and R_R must be sufficiently small, see below. Using the boundary conditions in (8), the boundary terms BT_L and BT_R become

$$\begin{aligned} \mathrm{BT}_L &= \mathcal{U}^T Z_-\mathcal{C}_L Z_-^T\mathcal{U}\Big|_{x_L} - 2\widetilde{g}_L^T\Lambda_+R_L Z_-^T\mathcal{U}\Big|_{x_L} + \widetilde{g}_L^T\Lambda_+\widetilde{g}_L \\ \mathrm{BT}_R &= \mathcal{U}^T Z_+\mathcal{C}_R Z_+^T\mathcal{U}\Big|_{x_R} + 2\widetilde{g}_R^T\Lambda_-R_R Z_+^T\mathcal{U}\Big|_{x_R} - \widetilde{g}_R^T\Lambda_-\widetilde{g}_R, \end{aligned} \tag{9}$$

where we have defined

$$\mathcal{C}_L = \Lambda_- + R_L^T\Lambda_+R_L, \qquad \mathcal{C}_R = -\Lambda_+ - R_R^T\Lambda_-R_R.$$

We note that if $\mathcal{C}_L \le 0$ and $\mathcal{C}_R \le 0$, the boundary terms in (9) will be non-positive for zero data. By integrating (5) in time we can now obtain a bound on $\|\mathcal{U}\|^2$. Since the boundary conditions have the form (8), we also know that the correct number of boundary conditions are specified at each boundary, which yields existence. Our problem is thus well-posed.

To relate the original boundary conditions in (4) to the ones in (8), we let

$$\mathcal{B}_L = P_L \left(Z_+^T + R_L Z_-^T \right), \qquad \mathcal{B}_R = P_R \left(Z_-^T + R_R Z_+^T \right), \tag{10}$$

where P_L and P_R are invertible scaling and/or permutation matrices given by the chosen $\mathcal{B}_{L,R}$ and $Z_\pm$. The data in (8) is identified as $\widetilde{g}_L = P_L^{-1} g_L$ and $\widetilde{g}_R = P_R^{-1} g_R$. We assume that the boundary conditions in (4) are properly chosen such that R_L and R_R are sufficiently small to make $\mathcal{C}_L, \mathcal{C}_R \leq 0$.

Remark 1 Note that the energy method is a sufficient but not necessary condition for stability and that it is rather restrictive with respect to the admissible boundary conditions. By rescaling the problem we could allow R_L and R_R to be larger, see [7,11]. We will not consider this complication but simply require that $\mathcal{C}_{L,R} \leq 0$.

Remark 2 In the homogeneous case, with boundary conditions such that $\mathcal{C}_{L,R} \leq 0$, the growth rate in (5) becomes $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathcal{U}\|^2 \leq 0$. Integrating this in time we obtain the energy estimate $\|\mathcal{U}\|^2 \leq \|\mathcal{U}_0\|^2$ and (4) is well-posed. Since (4) is an one-dimensional hyperbolic problem it is also possible to show strong well-posedness, i.e., that $\|\mathcal{U}\|$ is bounded by the data g_L, g_R, $\mathcal{F}$ and $\mathcal{U}_0$. See [7,11] for different definitions of well-posedness.

2.2 The Semi-discrete Problem

We discretize in space using $N+1$ equidistant grid points $x_i = x_L + hi$, where $h = (x_R - x_L)/N$ and $i = 0, 1, \ldots, N$. The semi-discrete scheme approximating (4) is written

$$\begin{aligned} U_t + (I_N \otimes \mathcal{R})U + (D_1 \otimes \mathcal{A})U = F + \left(H^{-1} e_0 \otimes \Sigma_0\right)(\mathcal{B}_L U_0 - g_L) \\ + \left(H^{-1} e_N \otimes \Sigma_N\right)(\mathcal{B}_R U_N - g_R), \end{aligned} \tag{11}$$

where $U = [U_0^T, U_1^T, \ldots, U_N^T]^T$ is a vector of length $n(N+1)$, such that $U_i(t) \approx \mathcal{U}(x_i, t)$, and where $F_i(t) = \mathcal{F}(x_i, t)$. The symbol $\otimes$ refers to the Kronecker product. The finite difference operator D_1 approximates $\partial/\partial x$ and satisfies the SBP-properties

$$D_1 = H^{-1} Q, \quad H = H^T > 0, \quad Q + Q^T = E_N - E_0 \tag{12}$$

where $E_0 = e_0 e_0^T$, $E_N = e_N e_N^T$, $e_0 = [1, 0, \ldots, 0]^T$ and $e_N = [0, \ldots, 0, 1]^T$. Note that $U_0 = (e_0^T \otimes I_n)U$ and $U_N = (e_N^T \otimes I_n)U$. By I_N and I_n we refer to identity matrices of size $N+1$ and n, respectively. The boundary conditions are imposed using the SAT technique which is a penalty method. The penalty parameters Σ_0 and Σ_N in (11) are at this point unknown, but are derived in the next subsections and presented in Theorem 1.

In this paper, we require that H is diagonal and in this case D_1 consists of a $2p$-order accurate central difference approximation in the interior and one-sided, p-order accurate approximations at the boundaries. Examples of SBP operators can be found in [13,17]. For more details about SBP–SAT, see [18] and references therein.

2.3 Numerical Stability Using the Energy Method

Just as in the continuous case we use the energy method to show stability. We multiply (11) by $U^T \overline{H}$ from the left, where $\overline{H} = H \otimes I_n$, and then add the transpose of the result. Thereafter using the SBP-properties in (12) we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_H^2 + 2U^T(H \otimes \mathcal{R})U = 2\langle U, F\rangle_H + \mathrm{BT}_L^{\mathrm{disc.}} + \mathrm{BT}_R^{\mathrm{disc.}},$$

where $\|U\|_H^2 = \langle U, U\rangle_H = U^T \overline{H} U$ is the discrete L^2-norm and where

$$\begin{aligned} \mathrm{BT}_L^{\mathrm{disc.}} &= U_0^T \left(\mathcal{A} + \Sigma_0 \mathcal{B}_L + \mathcal{B}_L^T \Sigma_0^T\right) U_0 - U_0^T \Sigma_0 g_L - g_L^T \Sigma_0^T U_0, \\ \mathrm{BT}_R^{\mathrm{disc.}} &= U_N^T \left(-\mathcal{A} + \Sigma_N \mathcal{B}_R + \mathcal{B}_R^T \Sigma_N^T\right) U_N - U_N^T \Sigma_N g_R - g_R^T \Sigma_N^T U_N. \end{aligned} \tag{13}$$

We define $C_0 = \mathcal{A} + \Sigma_0 \mathcal{B}_L + \mathcal{B}_L^T \Sigma_0^T$ and $C_N = -\mathcal{A} + \Sigma_N \mathcal{B}_R + \mathcal{B}_R^T \Sigma_N^T$. For stability $\mathrm{BT}_L^{\mathrm{disc.}}$ and $\mathrm{BT}_R^{\mathrm{disc.}}$ must be non-positive for zero boundary data, i.e., $C_0 \leq 0$ and $C_N \leq 0$. We make the following ansatz for the penalty parameters:

$$\Sigma_0 = (Z_+ \Pi_0 + Z_- \Gamma_0) P_L^{-1}, \qquad \Sigma_N = (Z_+ \Gamma_N + Z_- \Pi_N) P_R^{-1}, \tag{14}$$

where the matrices Π_0, Γ_0, Γ_N and Π_N will be determined over the next few pages. Taking the left boundary as example and using (7), (10) and (14) we obtain

$$C_0 = \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} \Delta_+ + \Pi_0 + \Pi_0^T & \Pi_0 R_L + \Gamma_0^T \\ \Gamma_0 + R_L^T \Pi_0^T & \Delta_- + \Gamma_0 R_L + R_L^T \Gamma_0^T \end{bmatrix} \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}. \tag{15}$$

2.4 The Dual Problem

Given the functional $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U}\rangle$, the dual problem of (4) is

$$\begin{aligned} \mathcal{V}_\tau + \mathcal{R}\mathcal{V} - \mathcal{A}\mathcal{V}_x &= \mathcal{G}, && x \in [x_L, x_R], \\ \widetilde{\mathcal{B}_L}\mathcal{V} &= \widetilde{g_L}, && x = x_L, \\ \widetilde{\mathcal{B}_R}\mathcal{V} &= \widetilde{g_R}, && x = x_R, \end{aligned} \tag{16}$$

which holds for $\tau \geq 0$ and is complemented with the initial data $\mathcal{V}(x, 0) = \mathcal{V}_0(x)$. Note that we have used the transformation $\tau = T - t$ mentioned in Sect. 1.1, such that $\mathcal{V} = \mathcal{V}(x, \tau)$. The boundary operators in (16) have the form

$$\widetilde{\mathcal{B}_L} = \widetilde{P_L}\left(Z_-^T + \widetilde{R_L} Z_+^T\right), \qquad \widetilde{\mathcal{B}_R} = \widetilde{P_R}\left(Z_+^T + \widetilde{R_R} Z_-^T\right), \tag{17}$$

where $\widetilde{P_L}$ and $\widetilde{P_R}$ are arbitrary invertible matrices and where $\widetilde{R_L}$ and $\widetilde{R_R}$ depend on the primal boundary conditions as

$$\widetilde{R_L} = -\Delta_-^{-1} R_L^T \Delta_+, \qquad \widetilde{R_R} = -\Delta_+^{-1} R_R^T \Delta_-. \tag{18}$$

The claim that (16), (17) and (18) describes the dual problem is motivated below: Using the notation in (1) and (2) we identify the spatial operators of (4) and (16) as

$$\mathcal{L} = \mathcal{R} + \mathcal{A}\frac{\partial}{\partial x}, \qquad \mathcal{L}^* = \mathcal{R} - \mathcal{A}\frac{\partial}{\partial x}, \tag{19}$$

respectively. For (16) to be the dual problem of (4), $\mathcal{L}$ and $\mathcal{L}^*$ must fulfill the relation $\langle \mathcal{V}, \mathcal{L}\mathcal{U}\rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U}\rangle$. Using integration by parts we obtain

$$\langle \mathcal{V}, \mathcal{L}\mathcal{U}\rangle = \langle \mathcal{L}^*\mathcal{V}, \mathcal{U}\rangle + [\mathcal{V}^T \mathcal{A}\mathcal{U}]_{x_L}^{x_R}$$

and we see that $\mathcal{V}^T \mathcal{A}\mathcal{U}$ must be zero at both boundaries (the boundary conditions for the dual problem are defined as the minimal set of homogeneous conditions such that all boundary terms vanish after that the homogeneous boundary conditions for the primal problem have been applied, see [1]). Using the boundary conditions of the primal problem, (8), followed by the dual boundary conditions, (16), (17), yields (for zero data)

$$\mathcal{V}^T \mathcal{A}\mathcal{U}\Big|_{x_L} = -\mathcal{V}^T Z_+ \left(\Delta_+ R_L + \widetilde{R_L}^T \Delta_-\right) Z_-^T \mathcal{U}\Big|_{x_L}$$
$$\mathcal{V}^T \mathcal{A}\mathcal{U}\Big|_{x_R} = -\mathcal{V}^T Z_- \left(\widetilde{R_R}^T \Delta_+ + \Delta_- R_R\right) Z_+^T \mathcal{U}\Big|_{x_R}$$

and if (18) holds, then $\mathcal{V}^T \mathcal{A}\mathcal{U} = 0$ at both boundaries and the above claim is confirmed.

Remark 3 The functional of interest can also include outgoing solution terms from the boundary, as $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U}\rangle + \Phi\mathcal{U}|_{x_L} + \Psi\mathcal{U}|_{x_R}$, where Φ and Ψ have the form $\Phi = \Phi_+ Z_+^T + \Phi_- Z_-^T$ and $\Psi = \Psi_+ Z_+^T + \Psi_- Z_-^T$. This specifies the boundary data in (16) to $\widetilde{g_L} = -\widetilde{P_L}\Delta_-^{-1}(\Phi_- - \Phi_+ R_L)^T$ and $\widetilde{g_R} = \widetilde{P_R}\Delta_+^{-1}(\Psi_+ - \Psi_- R_R)^T$, compare with [9]. When $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U}\rangle$ then the boundary data in (16) is actually zero.

2.4.1 Well-Posedness of the Dual Problem

The growth rate for the dual problem is given by

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\|\mathcal{V}\|^2 + 2\langle \mathcal{V}, \mathcal{R}\mathcal{V}\rangle = \mathrm{BT}_L^{\mathrm{dual}} + \mathrm{BT}_R^{\mathrm{dual}}$$

where the boundary terms (after that the homogeneous boundary conditions have been applied) are

$$\mathrm{BT}_L^{\mathrm{dual}} = \mathcal{V}^T Z_+ \widetilde{\mathcal{C}_L}\, Z_+^T \mathcal{V}\Big|_{x_L}, \qquad \mathrm{BT}_R^{\mathrm{dual}} = \mathcal{V}^T Z_- \widetilde{\mathcal{C}_R}\, Z_-^T \mathcal{V}\Big|_{x_R}$$

and where $\widetilde{\mathcal{C}_L} = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} R_L^T \Delta_+$ and $\widetilde{\mathcal{C}_R} = \Delta_- + \Delta_- R_R \Delta_+^{-1} R_R^T \Delta_-$. For well-posedness of the dual problem $\widetilde{\mathcal{C}_L} \le 0$ and $\widetilde{\mathcal{C}_R} \le 0$ are necessary.

Recall that the primal problem is well-posed if $\mathcal{C}_L, \mathcal{C}_R \le 0$. The dual demand $\widetilde{\mathcal{C}_L} \le 0$ is directly fulfilled if $\mathcal{C}_L \le 0$ and $\widetilde{\mathcal{C}_R} \le 0$ follows from $\mathcal{C}_R \le 0$. When R_L, R_R are square, invertible matrices, this is trivial. For general R_L, R_R it can be shown with the help of the determinant relation in Lemma 1 in "Appendix B". We conclude that the dual problem (16) with (17), (18) is well-posed if the primal problem (4) with (10) is well-posed.

Remark 4 In [2,3] the dual consistent schemes are constructed by first designing the boundary conditions (for incompletely parabolic problems) such that both the primal and the dual problem are well-posed. Their different approach can partly be explained by their wish to have the boundary conditions in the special form $H_{L,R}U \mp BU_x = G_{L,R}$. Looking e.g. at Eq. (30) in [2], we note that after applying the boundary conditions, $U^T M_L U \ge 0$ is needed for stability. However, if B is singular, replacing BU_x by $\pm H_{L,R}U$ does not guarantee that all conditions have been completely used, and u and p in $U = [p, u]^T$ in $U^T M_L U$ can be linearly dependent. Therefore the demand $M_L \ge 0$ after Eq. (31) in [2] is unnecessarily strong and gives some extra restrictions on the boundary conditions.

2.4.2 Discretization of the Dual Problem

The semi-discrete scheme approximating the dual problem (16) is written

$$\begin{aligned} V_\tau + (I_N \otimes \mathcal{R})V - (D_1 \otimes \mathcal{A})V = G &+ \left(H^{-1}e_0 \otimes \widetilde{\Sigma_0}\right)\left(\widetilde{\mathcal{B}_L}V_0 - \widetilde{g_L}\right) \\ &+ \left(H^{-1}e_N \otimes \widetilde{\Sigma_N}\right)\left(\widetilde{\mathcal{B}_R}V_N - \widetilde{g_R}\right), \end{aligned} \tag{20}$$

where $V_i(\tau)$ represents $\mathcal{V}(x_i, \tau)$. The SAT parameters $\widetilde{\Sigma_0}$ and $\widetilde{\Sigma_N}$ are yet unknown.

2.5 Dual Consistency

The semi-discrete scheme (11) is rewritten as $U_t + LU = \text{RHS}$, where

$$L = (I_N \otimes \mathcal{R}) + (D_1 \otimes \mathcal{A}) - \left(H^{-1}E_0 \otimes \Sigma_0 \mathcal{B}_L\right) - \left(H^{-1}E_N \otimes \Sigma_N \mathcal{B}_R\right)$$

and where RHS only depends on known data. In contrast to its continuous counterpart $\mathcal{L}$, L includes the boundary conditions explicitly. According to [2], the discrete adjoint operator is given by $L^* = \overline{H}^{-1} L^T \overline{H}$, which, using (12), leads to

$$\begin{aligned} L^* = &(I_N \otimes \mathcal{R}) - (D_1 \otimes \mathcal{A}) - \left(H^{-1}E_0 \otimes \mathcal{B}_L^T \Sigma_0^T + \mathcal{A}\right) \\ &- \left(H^{-1}E_N \otimes \mathcal{B}_R^T \Sigma_N^T - \mathcal{A}\right). \end{aligned} \tag{21}$$

If L^* is a consistent approximation of $\mathcal{L}^*$ in (19), then the scheme (11) is dual consistent. Looking at (20), we see that L^* must have the form

$$\begin{aligned} L^*_{\text{goal}} = &(I_N \otimes \mathcal{R}) - (D_1 \otimes \mathcal{A}) - \left(H^{-1}E_0 \otimes \widetilde{\Sigma_0}\widetilde{\mathcal{B}_L}\right) \\ &- \left(H^{-1}E_N \otimes \widetilde{\Sigma_N}\widetilde{\mathcal{B}_R}\right). \end{aligned} \tag{22}$$

Thus we have dual consistency if the expressions in (21) and (22) are equal. This gives us the following requirements:

$$\mathcal{B}_L^T \Sigma_0^T + \mathcal{A} - \widetilde{\Sigma_0}\widetilde{\mathcal{B}_L} = 0, \qquad \mathcal{B}_R^T \Sigma_N^T - \mathcal{A} - \widetilde{\Sigma_N}\widetilde{\mathcal{B}_R} = 0.$$

Similarly to the penalty parameters (14) for the primal problem, we make the ansatz

$$\widetilde{\Sigma_0} = \left(Z_+ \widetilde{\Gamma_0} + Z_- \widetilde{\Pi_0}\right) \widetilde{P_L}^{-1}, \qquad \widetilde{\Sigma_N} = \left(Z_+ \widetilde{\Pi_N} + Z_- \widetilde{\Gamma_N}\right) \widetilde{P_R}^{-1} \tag{23}$$

for the penalty parameters of the dual problem. The matrices $\widetilde{\Gamma_0}$, $\widetilde{\Pi_0}$, $\widetilde{\Pi_N}$ and $\widetilde{\Gamma_N}$ are at this stage unknown. We consider the left boundary and use (14) and (23), together with (7), (10) and (17), to write

$$\mathcal{B}_L^T \Sigma_0^T + \mathcal{A} - \widetilde{\Sigma_0}\widetilde{\mathcal{B}_L} = \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} \Delta_+ + \Pi_0^T - \widetilde{\Gamma_0}\widetilde{R_L} & \Gamma_0^T - \widetilde{\Gamma_0} \\ R_L^T \Pi_0^T - \widetilde{\Pi_0}\widetilde{R_L} & \Delta_- + R_L^T \Gamma_0^T - \widetilde{\Pi_0} \end{bmatrix} \begin{bmatrix} Z_+^T \\ Z_-^T \end{bmatrix}$$

which is zero if and only if the four entries of the matrix are zero. These four demands are rearranged to the more convenient form

$$\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0 \tag{24a}$$

$$\widetilde{R_L} = -\Delta_-^{-1} R_L^T \Delta_+ \tag{24b}$$

$$\widetilde{\Gamma_0} = \Gamma_0^T \tag{24c}$$

$$\widetilde{\Pi_0} = \Delta_- - \Delta_- \widetilde{R_L} \Delta_+^{-1} \widetilde{\Gamma_0}. \tag{24d}$$

Note that (24a) only depends on parameters from the primal problem, while (24d) only depends on parameters from the dual problem. Interestingly enough, (24b) is nothing but the duality demand (18) for the continuous problem. The demand (24c) relates the penalty of the dual problem to the primal penalty.

Unless we actually want to solve the dual problem, it is enough to consider the first demand, (24a). We repeat the above derivation also for the right boundary and get the following result: The penalty parameters Σ_0 and Σ_N in (14) with

$$\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1} \Gamma_0, \qquad \Pi_N = \Delta_- - \Delta_- R_R \Delta_+^{-1} \Gamma_N, \tag{25}$$

makes the discretization (11) dual consistent.

Remark 5 If the discrete primal problem (11) is dual consistent there is no need to check if the discrete dual problem (20) is stable—in [8] it is stated that stability of the primal problem implies stability of the dual problem, because the system matrix for the dual problem is the transpose of the system matrix for the primal problem—that is the primal and dual discrete problems have exactly the same growth rates for zero data.

2.6 Penalty Parameters for the Hyperbolic Problem

Consider the ansatz for the left penalty parameter, $\Sigma_0 = (Z_+\Pi_0 + Z_-\Gamma_0)P_L^{-1}$, which is given in (14). From a stability point of view, we must choose Π_0 and Γ_0 such that C_0 in (15) becomes non-positive. In addition, for dual consistency the constraint in (25) must be fulfilled. By inserting $\Pi_0 = -\Delta_+ - \Delta_+ R_L \Delta_-^{-1}\Gamma_0$ from (25) into C_0 we obtain, after some rearrangements, the expression

$$C_0 = \begin{bmatrix} P_L^{-1}\mathcal{B}_L \\ Z_-^T \end{bmatrix}^T \begin{bmatrix} -\Delta_+ - \Delta_+ R_L \Delta_-^{-1}\Gamma_0 - (\Delta_+ R_L \Delta_-^{-1}\Gamma_0)^T & \Gamma_0^T \Delta_-^{-1}\mathcal{C}_L \\ \mathcal{C}_L \Delta_-^{-1}\Gamma_0 & \mathcal{C}_L \end{bmatrix} \begin{bmatrix} P_L^{-1}\mathcal{B}_L \\ Z_-^T \end{bmatrix}.$$

The most obvious strategy to make $C_0 \leq 0$ is to cancel the off-diagonal entries by putting $\Gamma_0 = 0$, but note that other choices exist. To single out the optimal (in a certain sense) candidate, we use another approach. With (7), (10) and $\widetilde{g}_L = P_L^{-1} g_L$, the left boundary term in (13) can be rearranged as

$$\begin{aligned} \mathrm{BT}_L^{\mathrm{disc.}} = {} & U_0^T Z_- \mathcal{C}_L Z_-^T U_0 - 2\widetilde{g}_L^T \Delta_+ R_L Z_-^T U_0 + \widetilde{g}_L^T \Delta_+ \widetilde{g}_L \\ & - (\mathcal{B}_L U_0 - g_L)^T P_L^{-T} \Delta_+ P_L^{-1} (\mathcal{B}_L U_0 - g_L) \\ & + 2 (\mathcal{B}_L U_0 - g_L)^T \left(\Sigma_0 + Z_+ \Delta_+ P_L^{-1}\right)^T U_0, \end{aligned} \tag{26}$$

where we see that the first row corresponds exactly to the continuous boundary term BT_L in (9). The second row is a damping term that is quadratically proportional to the solution's deviation from data at the boundary, $\mathcal{B}_L U_0 - g_L$. The term in the last row is only linearly proportional to this deviation, so we would prefer it to be zero. This is possible if the penalty parameter is chosen exactly as $\Sigma_0 = -Z_+ \Delta_+ P_L^{-1}$. Luckily this choice fulfills both the stability requirement and the duality constraint. We repeat the above derivation also for the right boundary and summarize our findings in Theorem 1.

Theorem 1 *Consider the problem* (4) *with an associated factorization* (6) *where* Z *is non-singular. With the particular choice of penalty parameters*

$$\Sigma_0 = -Z_+ \Delta_+ P_L^{-1}, \qquad \Sigma_N = Z_- \Delta_- P_R^{-1}, \tag{27}$$

the scheme (11) *is a stable and dual consistent discretization of* (4)*. The matrices* P_L *and* P_R *are specified through* (10)*.*

Proof Comparing with (14), we note that Σ_0 in (27) is obtained using $\Pi_0 = -\Delta_+$ and $\Gamma_0 = 0$. These values fulfill the left duality constraint in (25). Inserting $\Gamma_0 = 0$ into C_0 above, we obtain $C_0 = Z_- \mathcal{C}_L Z_-^T - \mathcal{B}_L^T P_L^{-T} \Delta_+ P_L^{-1} \mathcal{B}_L$, which is negative semi-definite if the continuous problem is well-posed (in the $\mathcal{C}_L \leq 0$ sense). Thus the stability demand $C_0 \leq 0$ is fulfilled. The same is done for the right boundary, completing the proof. □

Remark 6 The seemingly very specific choice of penalty parameters in Theorem 1 is, in fact, a family of penalty parameters, depending on the factorization used. Note that it is not necessary to use the same factorization for the left and the right boundary.

Remark 7 If characteristic boundary conditions (in the sense $R_L, R_R = 0$) are used, the scheme (11) together with the SATs from Theorem 1 simplifies to

$$U_t + (I_N \otimes \mathcal{R})U + (D_1 \otimes \mathcal{A})U = F + (H^{-1}E_0 \otimes -\mathcal{A}_+)U + (H^{-1}E_N \otimes \mathcal{A}_-)U$$

in the homogeneous case, where $\mathcal{A}_+ = Z_+ \Delta_+ Z_+^T$ and $\mathcal{A}_- = Z_- \Delta_- Z_-^T$. When the factorization refers to the eigendecomposition, this corresponds to the SAT used for the characteristic boundary conditions of the nonlinear Euler equations in [9].

Remark 8 Assume that we are interested in the functional mentioned in Remark 3,

$$\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U} \rangle + \Phi \mathcal{U}|_{x_L} + \Psi \mathcal{U}|_{x_R}, \tag{28}$$

which includes boundary terms. We approximate it with the discrete functional

$$J(U) = G^T \overline{H} U + \Phi U_0 + \Psi U_N - \Phi_+ P_L^{-1}(\mathcal{B}_L U_0 - g_L) - \Psi_- P_R^{-1}(\mathcal{B}_R U_N - g_R). \tag{29}$$

Note that we have added correction terms which are proportional to the boundary condition deviations, where Φ_+ and Ψ_- are specified through $\Phi = \Phi_+ Z_+^T + \Phi_- Z_-^T$ and $\Psi = \Psi_+ Z_+^T + \Psi_- Z_-^T$. These penalty-like correction terms, which are derived following techniques from [9], make the discrete functional superconvergent.

3 Parabolic Systems

Consider the parabolic (or incompletely parabolic) system of partial differential equations

$$\begin{aligned} \mathcal{U}_t + \mathcal{A}\mathcal{U}_x - \mathcal{E}\mathcal{U}_{xx} &= \mathcal{F}, \qquad x \in [x_L, x_R], \\ \mathcal{H}_L \mathcal{U} + \mathcal{G}_L \mathcal{U}_x &= g_L, \qquad x = x_L, \\ \mathcal{H}_R \mathcal{U} + \mathcal{G}_R \mathcal{U}_x &= g_R, \qquad x = x_R, \end{aligned} \tag{30}$$

for $t \geq 0$, augmented with the initial condition $\mathcal{U}(x, 0) = \mathcal{U}_0(x)$. The matrices $\mathcal{A}$ and $\mathcal{E} \geq 0$ are symmetric $n \times n$ matrices, and we assume that $\mathcal{G}_L$ and $\mathcal{G}_R$ scales as $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R \mathcal{E}$, respectively. Treating $\mathcal{U}_x$ as a separate variable, we can rewrite (30) as a first order system (as was also done in [1,9]), arriving at

$$\begin{aligned} \overline{\mathcal{I}}\,\overline{\mathcal{U}}_t + \overline{\mathcal{R}}\,\overline{\mathcal{U}} + \overline{\mathcal{A}}\,\overline{\mathcal{U}}_x &= \overline{\mathcal{F}}, \quad x \in [x_L, x_R], \\ \overline{\mathcal{B}}_L \overline{\mathcal{U}} &= g_L, \quad x = x_L, \\ \overline{\mathcal{B}}_R \overline{\mathcal{U}} &= g_R, \quad x = x_R, \end{aligned} \tag{31}$$

where

$$\overline{\mathcal{I}} = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}, \qquad \overline{\mathcal{R}} = \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{E} \end{bmatrix}, \qquad \overline{\mathcal{U}} = \begin{bmatrix} \mathcal{U} \\ \mathcal{U}_x \end{bmatrix}, \qquad \overline{\mathcal{F}} = \begin{bmatrix} \mathcal{F} \\ 0 \end{bmatrix}$$

and

$$\overline{\mathcal{A}} = \begin{bmatrix} \mathcal{A} & -\mathcal{E} \\ -\mathcal{E} & 0 \end{bmatrix}, \qquad \overline{\mathcal{B}}_L = \begin{bmatrix} \mathcal{H}_L & \mathcal{G}_L \end{bmatrix}, \qquad \overline{\mathcal{B}}_R = \begin{bmatrix} \mathcal{H}_R & \mathcal{G}_R \end{bmatrix}. \tag{32}$$

The system (31) has almost the same form as (4) since $\overline{\mathcal{R}} \geq 0$ and $\bar{\mathcal{A}}$ are symmetric $m \times m$ matrices, where $m = 2n$. Thus we can use the results from the hyperbolic case.

Remark 9 In [2,3] the operators corresponding to $\mathcal{H}_L$, $\mathcal{G}_L$, $\mathcal{H}_R$ and $\mathcal{G}_R$ are square $n \times n$ matrices and their ranks are changed to suit the number of boundary conditions. We adapt the matrix dimensions instead. Both approaches have their respective advantages.

3.1 Discretization Using Wide-Stencil Second Derivative Operators

To discretize the parabolic problem, we first consider the reformulated problem (31), and use the results from the hyperbolic section. Then we rearrange the terms such that we get an equivalent scheme but in a form corresponding to (30). These steps, which are done in "Appendix A", lead to

$$\begin{aligned} U_t + (D_1 \otimes \mathcal{A})U - (D_1^2 \otimes \mathcal{E})U = F &+ \overline{H}^{-1}(e_0 \otimes \widehat{\mu}_0 + D_1^T e_0 \otimes \widehat{\nu}_0)\widehat{\xi}_0 \\ &+ \overline{H}^{-1}(e_N \otimes \widehat{\mu}_N + D_1^T e_N \otimes \widehat{\nu}_N)\widehat{\xi}_N \end{aligned} \tag{33}$$

where

$$\widehat{\xi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\overline{D}U)_0 - g_L, \qquad \widehat{\xi}_N = \mathcal{H}_R U_N + \mathcal{G}_R(\overline{D}U)_N - g_R, \tag{34}$$

and $\overline{H} = (H \otimes I_n)$ and $\overline{D} = (D_1 \otimes I_n)$. The penalty parameters in (33) are

$$\begin{aligned} \widehat{\mu}_0 &= (-\overline{Z}_1 + \widehat{q}\,\overline{Z}_2)\overline{\Delta}_+\widehat{\Xi}_L^{-1}, & \widehat{\nu}_0 &= \overline{Z}_2\overline{\Delta}_+\widehat{\Xi}_L^{-1}, \\ \widehat{\mu}_N &= (\overline{Z}_3 + \widehat{q}\,\overline{Z}_4)\overline{\Delta}_-\widehat{\Xi}_R^{-1}, & \widehat{\nu}_N &= -\overline{Z}_4\overline{\Delta}_-\widehat{\Xi}_R^{-1}, \end{aligned} \tag{35}$$

where $\widehat{\Xi}_L = \overline{P}_L + \widehat{q}\,\mathcal{K}_L\overline{Z}_2\overline{\Delta}_+$ and $\widehat{\Xi}_R = \overline{P}_R - \widehat{q}\,\mathcal{K}_R\overline{Z}_4\overline{\Delta}_-$ and where the matrices $\overline{Z}_{1,2,3,4}$ are defined through

$$\overline{Z}_+ = \begin{bmatrix} \overline{Z}_1 \\ \overline{Z}_2 \end{bmatrix}, \qquad \overline{Z}_- = \begin{bmatrix} \overline{Z}_3 \\ \overline{Z}_4 \end{bmatrix}. \tag{36}$$

As before, $\overline{\Delta}_\pm$, $\overline{Z}_\pm$ and $\overline{P}_L$, $\overline{P}_R$ are described in (6) and (10), respectively, but are now obtained using $\bar{\mathcal{A}}$ and $\overline{\mathcal{B}}_L$, $\overline{\mathcal{B}}_R$ from (32). Finally, the quantity $\widehat{q}$ in (35) is given by

$$\widehat{q} = e_0^T H^{-1} e_0 = e_N^T H^{-1} e_N. \tag{37}$$

The matrix H is positive definite and proportional to the grid size h, and thus $\widehat{q}$ is a positive scalar proportional to $1/h$.

Remark 10 Given well-posedness (i.e., in the sense $\overline{\mathcal{C}}_L = \overline{\Delta}_- + \overline{R}_L^T\overline{\Delta}_+\overline{R}_L \leq 0$ and $\overline{\mathcal{C}}_R = -\overline{\Delta}_+ - \overline{R}_R^T\overline{\Delta}_-\overline{R}_R \leq 0$) and that $\widehat{q} \geq 0$, one can show that $\widehat{\Xi}_L$ and $\widehat{\Xi}_R$ are non-singular. This is done in "Appendix B".

3.2 Discretization Using Narrow-Stencil Second Derivative Operators

In [1], it was suggested that dual consistency might require wide-stencil second derivative operators, but next we will show that this is not necessary. The semi-discrete scheme approximating (30) is now written, analogously to (33), as

$$\begin{aligned} U_t + (D_1 \otimes \mathcal{A})U - (D_2 \otimes \mathcal{E})U = F &+ \overline{H}^{-1}\left(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0\right)\xi_0 \\ &+ \overline{H}^{-1}\left(e_N \otimes \mu_N + S^T e_N \otimes \nu_N\right)\xi_N. \end{aligned} \tag{38}$$

The operator D_2, which approximates the second derivative operator, is no longer limited to the previous form D_1^2, where the first derivative is used twice. However, D_2 must still fulfill the SBP relations

$$D_2 = H^{-1}(-A_S + (E_N - E_0)S), \qquad A_S = A_S^T = S^T M S \geq 0, \tag{39}$$

where the first and last row of the matrix S are consistent difference stencils, see e.g., [13]. For dual consistency, A_S must be symmetric. Note that when having narrow-stencil operators the interior of S is not uniquely defined and neither is the matrix M above (this is discussed in Sect. 4). Furthermore, in (38) we have

$$\xi_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\bar{S}U)_0 - g_L, \qquad \xi_N = \mathcal{H}_R U_N + \mathcal{G}_R(\bar{S}U)_N - g_R, \tag{40}$$

where

$$\bar{S} = S \otimes I_n, \qquad (\bar{S}U)_0 = (e_0^T S \otimes I_n)U, \qquad (\bar{S}U)_N = (e_N^T S \otimes I_n)U.$$

We also define

$$q \equiv q_0 + |q_c| = q_N + |q_c| \tag{41}$$

where

$$q_0 = e_0^T M^{-1} e_0, \qquad q_N = e_N^T M^{-1} e_N, \qquad q_c = e_0^T M^{-1} e_N = e_N^T M^{-1} e_0, \tag{42}$$

where M is a part of D_2 as stated in (39). In Sect. 4 we provide q for various D_2 matrices. The penalty parameters μ_0, ν_0, μ_N and ν_N in (38) are now given by:

Theorem 2 *Consider the problem (30) with $\mathcal{G}_L = \mathcal{K}_L\mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R\mathcal{E}$. Furthermore, let $\bar{\mathcal{A}}$, which is specified in (32), be factorized as $\bar{\mathcal{A}} = \bar{Z}\,\bar{\Lambda}\bar{Z}^T$ as described in (6). Then the particular choice of penalty parameters*

$$\begin{aligned} \mu_0 &= \left(-\bar{Z}_1 + q\bar{Z}_2\right)\bar{\Lambda}_+\Xi_L^{-1}, & \nu_0 &= \bar{Z}_2\bar{\Lambda}_+\Xi_L^{-1}, \\ \mu_N &= \left(\bar{Z}_3 + q\bar{Z}_4\right)\bar{\Lambda}_-\Xi_R^{-1}, & \nu_N &= -\bar{Z}_4\bar{\Lambda}_-\Xi_R^{-1}, \end{aligned} \tag{43}$$

where $\Xi_L = \bar{P}_L + q\,\mathcal{K}_L\bar{Z}_2\bar{\Lambda}_+$ and $\Xi_R = \bar{P}_R - q\,\mathcal{K}_R\bar{Z}_4\bar{\Lambda}_-$, makes the scheme in (38) stable and dual consistent. The matrices $\bar{Z}_{1,2,3,4}$ are given in (36), $\bar{P}_L$, $\bar{P}_R$ are obtained from (10) (using $\bar{\mathcal{B}}_L$, $\bar{\mathcal{B}}_R$ in (32)) and q is defined in (41). The matrices Ξ_L and Ξ_R are non-singular for well-posed problems, see "Appendix B".

Note that q in (41) is a generalization of $\widehat{q}$ in (37), and that the penalty parameters in (43) and (35) are identical if $q = \widehat{q}$. Hence the narrow-stencil scheme (38) is a generalization of the wide-stencil scheme in (33), since the schemes are identical if we choose $D_2 = D_1^2$, $S = D_1$ and $M = H$. In the rest of this section we will justify these generalizations and prove Theorem 2 by showing that the penalties given in (43) indeed make the scheme (38) stable and dual consistent.

3.3 Stability When Using Narrow-Stencil Second Derivative Operators

We multiply the scheme (38) by $U^T\bar{H}$ from the left and add the transpose of the result. Thereafter using the SBP-properties in (12) and (39) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_H^2 + 2U^T(S^T M S \otimes \mathcal{E})U = 2\langle U, F\rangle_H + \mathrm{BT}_L^{\mathrm{disc.}} + \mathrm{BT}_R^{\mathrm{disc.}}, \tag{44}$$

where

$$\begin{aligned} \mathrm{BT}_L^{\mathrm{disc.}} &= U_0^T \mathcal{A} U_0 \; - 2U_0^T \mathcal{E}(\bar{S}U)_0 \; + 2\left(U_0^T \mu_0 \; + (\bar{S}U)_0^T \nu_0 \right) \xi_0 \\ \mathrm{BT}_R^{\mathrm{disc.}} &= -U_N^T \mathcal{A} U_N + 2U_N^T \mathcal{E}(\bar{S}U)_N + 2\left(U_N^T \mu_N + (\bar{S}U)_N^T \nu_N\right) \xi_N \end{aligned} \tag{45}$$

where $\xi_{0,N}$ are given in (40). If $\mathrm{BT}_L^{\mathrm{disc.}}$ and $\mathrm{BT}_R^{\mathrm{disc.}}$ are non-positive for zero data the scheme is stable. This can be achieved if μ_0, ν_0, μ_N and ν_N are chosen freely, but the scheme should also be dual consistent. It turns out that in some cases these requirements are impossible to combine, for example when having Dirichlet boundary conditions. We therefore need an alternative way to show stability.

First, we assume that the penalty parameters ν_0 and ν_N scales with $\mathcal{E}$. Let

$$\nu_0 = -\mathcal{E}\kappa_0, \qquad\qquad \nu_N = -\mathcal{E}\kappa_N. \tag{46}$$

Next, we take a look at the wide case (which is partly presented in "Appendix A"). Using a wide counterpart to (46), $\widehat{\nu}_0 = -\mathcal{E}\widehat{\kappa}_0$ and $\widehat{\nu}_N = -\mathcal{E}\widehat{\kappa}_N$, and the later relations in (71) and (72), we can rewrite (67b) as

$$\widehat{W} = \overline{D}U + (H^{-1}e_0 \otimes \widehat{\kappa}_0)\widehat{\xi}_0 + (H^{-1}e_N \otimes \widehat{\kappa}_N)\widehat{\xi}_N.$$

We return to the narrow-stencil scheme (38). Inspired by the wide case, we define

$$W \equiv \bar{S}U + (M^{-1}e_0 \otimes \kappa_0)\xi_0 + (M^{-1}e_N \otimes \kappa_N)\xi_N. \tag{47}$$

From (47) we compute

$$\begin{aligned} W^T(M \otimes \mathcal{E})W = U^T(S^T M S \otimes \mathcal{E})U + \left(2(\bar{S}U)_0 \; + q_0\kappa_0\xi_0 \; + q_c\kappa_N\xi_N\right)^T \mathcal{E}\kappa_0\,\xi_0 \\ + \left(2(\bar{S}U)_N + q_N\kappa_N\xi_N + q_c\kappa_0\xi_0\right)^T \mathcal{E}\kappa_N\xi_N \end{aligned}$$

where q_0, q_N and q_c are given in (42). In the general case, q_c can be non-zero. Since we want to treat the two boundaries separately, we use Young's inequality, $q_c(\xi_N^T\kappa_N^T\mathcal{E}\kappa_0\xi_0 + \xi_0^T\kappa_0^T\mathcal{E}\kappa_N\xi_N) \le |q_c|\left(\xi_0^T\kappa_0^T\mathcal{E}\kappa_0\xi_0 + \xi_N^T\kappa_N^T\mathcal{E}\kappa_N\xi_N\right)$, which yields

$$\begin{aligned} W^T(M \otimes \mathcal{E})W \le U^T(S^T M S \otimes \mathcal{E})U + \left(2(\bar{S}U)_0 \; + q\kappa_0\,\xi_0\right)^T \mathcal{E}\kappa_0\,\xi_0 \\ + \left(2(\bar{S}U)_N + q\kappa_N\xi_N\right)^T \mathcal{E}\kappa_N\xi_N \end{aligned} \tag{48}$$

where $q = q_0 + |q_c| = q_N + |q_c|$, as stated in (41). Further, we note that multiplying (47) by $(e_0^T \otimes I_n)$ and $(e_N^T \otimes I_n)$, respectively, yields the relations $W_0 = (\bar{S}U)_0 + q_0\kappa_0\xi_0 + q_c\kappa_N\xi_N$ and $W_N = (\bar{S}U)_N + q_c\kappa_0\xi_0 + q_N\kappa_N\xi_N$. Instead of using those, which contain unwanted terms from the other boundary, we define

$$\widetilde{W}_0 \equiv (\bar{S}U)_0 + q\kappa_0\xi_0 \qquad\qquad \widetilde{W}_N \equiv (\bar{S}U)_N + q\kappa_N\xi_N. \tag{49}$$

Inserting the relation (48) into (44), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|U\|_H^2 + 2W^T(M \otimes \mathcal{E})W \le 2\langle U, F\rangle_H + \widetilde{\mathrm{BT}}_L^{\mathrm{disc.}} + \widetilde{\mathrm{BT}}_R^{\mathrm{disc.}} \tag{50}$$

where (45) and (49) together with (46) yields

$$\begin{aligned} \widetilde{\mathrm{BT}}_L^{\mathrm{disc.}} &= U_0^T\mathcal{A}U_0 - 2U_0^T\mathcal{E}\widetilde{W}_0 + 2(U_0^T(\mu_0 - q\nu_0) - \widetilde{W}_0^T\nu_0)\xi_0 \\ \widetilde{\mathrm{BT}}_R^{\mathrm{disc.}} &= -U_N^T\mathcal{A}U_N + 2U_N^T\mathcal{E}\widetilde{W}_N + 2(U_N^T(\mu_N + q\nu_N) - \widetilde{W}_N^T\nu_N)\xi_N. \end{aligned} \tag{51}$$

If the penalty parameters make $\widetilde{\mathrm{BT}}_L^{\mathrm{disc.}} \le 0$ and $\widetilde{\mathrm{BT}}_R^{\mathrm{disc.}} \le 0$ for zero data, (38) is stable.

Again taking the left boundary as an example, we define $\widetilde{U}_0 = [U_0^T, \widetilde{W}_0^T]^T$ and write the first part of $\widetilde{\mathrm{BT}}_L^{\mathrm{disc.}}$ in (51) as

$$U_0^T \mathcal{A} U_0 - 2U_0^T \mathcal{E} \widetilde{W}_0 = \widetilde{U}_0^T \bar{\mathcal{A}} \widetilde{U}_0. \tag{52}$$

Next, using the relations (32), (49) and (40), recalling the assumptions $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$ and $\nu_0 = -\mathcal{E}\kappa_0$, and thereafter using (43) from Theorem 2, we obtain

$$\bar{\mathcal{B}}_L \widetilde{U}_0 - g_L = \overline{P}_L (\overline{P}_L + q\, \mathcal{K}_L \bar{Z}_2 \bar{\Delta}_+)^{-1} \xi_0. \tag{53}$$

From (43) we also get

$$\mu_0 - q\nu_0 = -\bar{Z}_1 \bar{\Delta}_+ (\overline{P}_L + q\, \mathcal{K}_L \bar{Z}_2 \bar{\Delta}_+)^{-1}, \quad -\nu_0 = -\bar{Z}_2 \bar{\Delta}_+ (\overline{P}_L + q\, \mathcal{K}_L \bar{Z}_2 \bar{\Delta}_+)^{-1}$$

such that the second part of $\widetilde{\mathrm{BT}}_L^{\mathrm{disc.}}$ in (51) becomes

$$2\left(U_0^T(\mu_0 - q\nu_0) - \widetilde{W}_0^T \nu_0\right)\xi_0 = 2\widetilde{U}_0^T \overline{\Sigma}_0 (\bar{\mathcal{B}}_L \widetilde{U}_0 - g_L) \tag{54}$$

where the relations (36) and (53) have been used, and where $\overline{\Sigma}_0 = -\bar{Z}_+ \bar{\Delta}_+ \overline{P}_L^{-1}$. Now we can, by inserting (52) and (54) into (51), write

$$\widetilde{\mathrm{BT}}_L^{\mathrm{disc.}} = \widetilde{U}_0^T \bar{\mathcal{A}} \widetilde{U}_0 + 2\widetilde{U}_0^T \overline{\Sigma}_0 \left(\bar{\mathcal{B}}_L \widetilde{U}_0 - g_L\right)$$

which has exactly the same form as $\mathrm{BT}_L^{\mathrm{disc.}}$ in (13). We thus know that $\widetilde{\mathrm{BT}}_L^{\mathrm{disc.}} \leq 0$ for zero data, since $\overline{\Sigma}_0$ is computed just as in the hyperbolic case. The same procedure can, of course, be repeated for the right boundary. We conclude that the scheme (38) with the penalty parameters (43) is stable.

3.4 Dual Consistency for Narrow-Stencil Second Derivative Operators

The dual problem of (30) is

$$\begin{aligned} \mathcal{V}_\tau - \mathcal{A}\mathcal{V}_x - \mathcal{E}\mathcal{V}_{xx} &= \mathcal{G}, && x \in [x_L, x_R], \\ \widetilde{\mathcal{H}}_L \mathcal{V} + \widetilde{\mathcal{G}}_L \mathcal{V}_x &= \widetilde{g_L}, && x = x_L, \\ \widetilde{\mathcal{H}}_R \mathcal{V} + \widetilde{\mathcal{G}}_R \mathcal{V}_x &= \widetilde{g_R}, && x = x_R, \end{aligned} \tag{55}$$

for $\tau \geq 0$ and with $\mathcal{V}(x,0) = \mathcal{V}_0(x)$. The spatial operator in (30) and its dual are thus

$$\mathcal{L} = \mathcal{A}\frac{\partial}{\partial x} - \mathcal{E}\frac{\partial^2}{\partial x^2}, \qquad \mathcal{L}^* = -\mathcal{A}\frac{\partial}{\partial x} - \mathcal{E}\frac{\partial^2}{\partial x^2}. \tag{56}$$

The semi-discrete approximation of (55) is

$$\begin{aligned} V_\tau - (D_1 \otimes \mathcal{A})V - (D_2 \otimes \mathcal{E})V = G &+ \overline{H}^{-1}\left(e_0 \otimes \widetilde{\mu_0} + S^T e_0 \otimes \widetilde{\nu_0}\right)\widetilde{\xi_0} \\ &+ \overline{H}^{-1}\left(e_N \otimes \widetilde{\mu_N} + S^T e_N \otimes \widetilde{\nu_N}\right)\widetilde{\xi_N}, \end{aligned} \tag{57}$$

where

$$\widetilde{\xi_0} = \widetilde{\mathcal{H}}_L V_0 + \widetilde{\mathcal{G}}_L (\bar{S}V)_0 - \widetilde{g_L}, \qquad \widetilde{\xi_N} = \widetilde{\mathcal{H}}_R V_N + \widetilde{\mathcal{G}}_R (\bar{S}V)_N - \widetilde{g_R}.$$

From (38) we see that the discrete operator, corresponding to $\mathcal{L}$ in (56), is

$$\begin{aligned} L = &(D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) \\ &- \overline{H}^{-1} \left(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0 \right) \left(e_0^T \otimes \mathcal{H}_L + e_0^T S \otimes \mathcal{G}_L \right) \\ &- \overline{H}^{-1} \left(e_N \otimes \mu_N + S^T e_N \otimes \nu_N \right) \left(e_N^T \otimes \mathcal{H}_R + e_N^T S \otimes \mathcal{G}_R \right). \end{aligned} \tag{58}$$

Using the relations in (12) and (39), we obtain

$$\begin{aligned} L^* = \overline{H}^{-1} L^T \overline{H} = &- (D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) \\ &- \overline{H}^{-1} \left(e_0 e_0^T \otimes \mathcal{A} \right) + \overline{H}^{-1} \left(\left(S^T e_0 e_0^T - e_0 e_0^T S \right) \otimes \mathcal{E} \right) \\ &+ \overline{H}^{-1} \left(e_N e_N^T \otimes \mathcal{A} \right) - \overline{H}^{-1} \left(\left(S^T e_N e_N^T - e_N e_N^T S \right) \otimes \mathcal{E} \right) \\ &- \overline{H}^{-1} \left(e_0 \otimes \mathcal{H}_L^T + S^T e_0 \otimes \mathcal{G}_L^T \right) \left(e_0^T \otimes \mu_0^T + e_0^T S \otimes \nu_0^T \right) \\ &- \overline{H}^{-1} \left(e_N \otimes \mathcal{H}_R^T + S^T e_N \otimes \mathcal{G}_R^T \right) \left(e_N^T \otimes \mu_N^T + e_N^T S \otimes \nu_N^T \right). \end{aligned}$$

However, from (57) we see that for dual consistency L^* must have the form

$$\begin{aligned} L^*_{\text{goal}} = &- (D_1 \otimes \mathcal{A}) - (D_2 \otimes \mathcal{E}) \\ &- \overline{H}^{-1} \left(e_0 \otimes \widetilde{\mu_0} + S^T e_0 \otimes \widetilde{\nu_0} \right) \left(e_0^T \otimes \widetilde{\mathcal{H}_L} + e_0^T S \otimes \widetilde{\mathcal{G}_L} \right) \\ &- \overline{H}^{-1} \left(e_N \otimes \widetilde{\mu_N} + S^T e_N \otimes \widetilde{\nu_N} \right) \left(e_N^T \otimes \widetilde{\mathcal{H}_R} + e_N^T S \otimes \widetilde{\mathcal{G}_R} \right). \end{aligned}$$

Demanding that $L^* = L^*_{\text{goal}}$, gives us the duality constraints

$$\begin{aligned} \begin{bmatrix} \mathcal{H}_L^T \mu_0^T + \mathcal{A} & \mathcal{H}_L^T \nu_0^T + \mathcal{E} \\ \mathcal{G}_L^T \mu_0^T - \mathcal{E} & \mathcal{G}_L^T \nu_0^T \end{bmatrix} &= \begin{bmatrix} \widetilde{\mu_0} \widetilde{\mathcal{H}_L} & \widetilde{\mu_0} \widetilde{\mathcal{G}_L} \\ \widetilde{\nu_0} \widetilde{\mathcal{H}_L} & \widetilde{\nu_0} \widetilde{\mathcal{G}_L} \end{bmatrix} \\ \begin{bmatrix} \mathcal{H}_R^T \mu_N^T - \mathcal{A} & \mathcal{H}_R^T \nu_N^T - \mathcal{E} \\ \mathcal{G}_R^T \mu_N^T + \mathcal{E} & \mathcal{G}_R^T \nu_N^T \end{bmatrix} &= \begin{bmatrix} \widetilde{\mu_N} \widetilde{\mathcal{H}_R} & \widetilde{\mu_N} \widetilde{\mathcal{G}_R} \\ \widetilde{\nu_N} \widetilde{\mathcal{H}_R} & \widetilde{\nu_N} \widetilde{\mathcal{G}_R} \end{bmatrix}. \end{aligned} \tag{59}$$

The duality constraints in (59) do not depend explicitly on the grid size h. Moreover, we already know that for the wide case, the penalty parameters in (35)—even though they contain the h-dependent constant $\widehat{q}$—gives dual consistency. Since the generalized penalty parameters in (43) have exactly the same form (the only difference is that they depend on *another* h-dependent constant, q) they will also yield dual consistency. We have thus shown that the penalty parameters in Theorem 2 indeed makes the scheme (38) stable and dual consistent.

Remark 11 The SAT parameters in Theorem 2 are probably a subset of all parameters giving stability and dual consistency since the duality constraint (59) could be used in combination with some other stability proof than the one presented here.

4 Computing q

We want to compute $q = q_0 + |q_c| = q_N + |q_c|$ as stated in (41) and are thus looking for q_0, q_N and q_c specified in (42). For wide second derivative operators, M is equal to H, and is thus well-defined. When using narrow second derivative operators, M is defined in

Table 1 The q-values (scaled with h) for various second derivative operators

Order	Type	qh	Comment
2,0	wide	2	
4,1	wide	$\frac{48}{17} \approx 2.8235$	
6,2	wide	$\frac{43200}{13649} \approx 3.1651$	
8,3	wide	$\frac{5080320}{1498139} \approx 3.3911$	
2,0	narrow	1	See Eq. (73)
2,1	narrow	2.5	
4,2	narrow	3.986391480987749	$(N = 8)$
6,3	narrow	5.322804652661742	$(N = 12)$
8,4	narrow	633.69326893357	$(N = 16)$

(39) through $A_S = S^T M S$. However, only the first and last row of S are clearly specified. In for example [4,5,13], the interior of S is chosen to be the identity matrix, and S is then invertible. A_S is singular (since $A_S = (E_N - E_0)S - H D_2$, where D_2 and the first and last row of S are consistent difference operators) and thus an invertible S implies that M is singular.

If M and S are defined such that M is singular and S not (which is often the case), we use the following strategy to find q: The relation $A_S = S^T M S$ leads to $M^{-1} = S A_S^{-1} S^T$, but since A_S is singular we define the perturbed matrix $\widetilde{A}_S \equiv A_S + \delta E_0$ and compute $\widetilde{M}^{-1} = S \widetilde{A}_S^{-1} S^T$ instead. This is motivated by the following proposition:

Proposition 1 *Define $\widetilde{A}_S \equiv A_S + \delta E_0$, where A_S is a part of a consistent second derivative operator D_2 fulfilling the relations in (39), $\delta \neq 0$ is a scalar parameter and E_0 is an all-zero matrix except for the element $(E_0)_{0,0} = 1$. The inverse of $\widetilde{A}_S$ is $\widetilde{A}_S^{-1} = J/\delta + K_0$, where J is an all-ones matrix and K_0 is a matrix that does not depend on δ. A consequence of this structure is that the corners of $\widetilde{M}^{-1} = S \widetilde{A}_S^{-1} S^T$ are independent of δ, such that*

$$q_0 = e_0^T \widetilde{M}^{-1} e_0, \qquad q_N = e_N^T \widetilde{M}^{-1} e_N, \qquad q_c = e_0^T \widetilde{M}^{-1} e_N = e_N^T \widetilde{M}^{-1} e_0. \tag{60}$$

Proposition 1 is proved in "Appendix C". In Table 1 we provide the value of q for all second derivative operators considered in this paper. The wide-stencil operators are given by $D_2 = D_1^2$, where D_1 has the order of accuracy $(2, 1), (4, 2), (6, 3)$ or $(8, 4)$, paired as (interior order, boundary order). For these operators, the q values are obtained directly from the matrix H. For the narrow-stencil operators, the q values are computed according to Proposition 1. All examples in Table 1, except the narrow $(2, 0)$ order operator, refers to operators given in [13]. Note that for some of the narrow-stencil operators q varies slightly with N, see "Appendix C".

Remark 12 The SBP operators with interior order 6 or 8 have free parameters, and if those parameters are chosen differently than in [13], that will affect q.

Remark 13 The quantity q has nothing to do with dual consistency, but indicates how the penalty should be chosen to give energy stability. As an example, consider solving the scalar problem presented below in (62) with Dirichlet boundary conditions, using the scheme (64). Using the same technique as in Sect. 3.3, we find that the stability demands for the (left) penalty parameter μ_0, in three special cases of ν_0, are

Dual consistent (see Eq. (65))	$\nu_0 = -\varepsilon$	$\mu_0 \leq -a/2 - \varepsilon q$
Method 1 (dual inconsistent)	$\nu_0 = 0$	$\mu_0 \leq -a/2 - \varepsilon q/4$
Method 2 (dual inconsistent)	$\nu_0 = \varepsilon$	$\mu_0 \leq -a/2.$

The two latter approaches are frequently used but they do not yield dual consistency.

5 Examples and Numerical Experiments

In this section, we give a few concrete examples of the derived penalty parameters and perform some numerical simulations. We demonstrate that these penalty parameters give superconvergent functional output not only for the wide second derivative operators but also for the narrow ones. The following procedure is used:

(i) Consider a continuous problem formulated as (30), where $\mathcal{G}_L = \mathcal{K}_L\mathcal{E}$ and $\mathcal{G}_R = \mathcal{K}_R\mathcal{E}$ are required. Identify $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}_L, \bar{\mathcal{B}}_R$ according to (32).

(ii) Factorize $\bar{\mathcal{A}}$ as $\bar{\mathcal{A}} = \bar{Z}\,\bar{\Delta}\,\bar{Z}^T$, according to (6), where $\bar{Z}$ must be non-singular.

(iii) Compute $\bar{P}_L$ and $\bar{P}_R$. From (10) we see that $\bar{P}_L$ is the first $m_+ \times m_+$ part of $\bar{\mathcal{B}}_L\bar{Z}^{-T}$ and that $\bar{P}_R$ is the last $m_- \times m_-$ part of $\bar{\mathcal{B}}_R\bar{Z}^{-T}$, as

$$\bar{\mathcal{B}}_L\bar{Z}^{-T} = \left[\bar{P}_L\ 0_{m_+,m_0}\ \bar{P}_L\bar{R}_L\right], \qquad \bar{\mathcal{B}}_R\bar{Z}^{-T} = \left[\bar{P}_R\bar{R}_R\ 0_{m_-,m_0}\ \bar{P}_R\right]. \tag{61}$$

(iv) The problem (30) is discretized in space using the scheme (38). Rearranging the terms in the scheme yields $U_t + LU = \text{RHS}$, where L is given in (58), and where

$$\text{RHS} = F - \bar{H}^{-1}(e_0 \otimes \mu_0 + S^T e_0 \otimes \nu_0)g_L - \bar{H}^{-1}(e_N \otimes \mu_N + S^T e_N \otimes \nu_N)g_R.$$

The penalty parameters μ_0, ν_0, μ_N and ν_N are specified in Theorem 2.

(v) If $\mathcal{U}_t = 0$, we have a stationary problem and the linear system $LU = \text{RHS}$ must be solved. For the time-dependent cases, we use the method of lines and discretize $U_t + LU = \text{RHS}$ in time using a suitable solver for ordinary differential equations.

Remark 14 When we have a hyperbolic problem, step (i) is omitted and step (iv) is modified such that the scheme (11) is used with penalty parameters given in Theorem 1.

In the simulations, we are interested in the functional error $\mathtt{E} = J(U) - \mathcal{J}(\mathcal{U})$, where $\mathcal{J}(\mathcal{U}) = \langle \mathcal{G}, \mathcal{U}\rangle$, $J(U) = \langle G, U\rangle_H$ and $G_i(t) = \mathcal{G}(x_i, t)$, but of course also in the solution error $\mathtt{e}$, where $\mathtt{e}_i(t) = U_i(t) - \mathcal{U}(x_i, t)$. We also investigate the spectra of L, that is the eigenvalues λ_j of L, with $j = 1, 2, \ldots, n(N+1)$. Here we are in particular interested in the spectral radius $\rho = \max_j(|\lambda_j|)$ and in $\eta = \min_j(\Re(\lambda_j))$. For time-dependent problems $\rho\Delta t \lesssim C$ is a crude estimate of the stability regions of explicit Runge–Kutta schemes, and thus ρ can be seen as a measure of stiffness. The eigenvalue with the smallest real part, η, determines how fast a time-dependent solution converges to a steady-state solution, see [14]. Ideally, the penalties are chosen such that the errors and ρ are kept small while η is maximized. For steady problems or when using implicit time solvers, other properties (e.g., the condition number) might be of greater interest.

We start by investigating the scalar case in some detail, then give an example of a system with a solid wall type of boundary condition.

5.1 The Scalar Case

Consider the scalar advection-diffusion equation,

$$\begin{aligned} \mathcal{U}_t + a\mathcal{U}_x - \varepsilon\mathcal{U}_{xx} &= \mathcal{F}, \quad x \in [0,1], \\ \alpha_L\mathcal{U} + \beta_L\mathcal{U}_x &= g_L, \quad x = 0, \\ \alpha_R\mathcal{U} + \beta_R\mathcal{U}_x &= g_R, \quad x = 1, \end{aligned} \tag{62}$$

valid for $t \geq 0$, with initial condition $\mathcal{U}(x,0) = \mathcal{U}_0(x)$ and where $\varepsilon > 0$. Using (32) yields

$$\bar{\mathcal{A}} = \begin{bmatrix} a & -\varepsilon \\ -\varepsilon & 0 \end{bmatrix}, \qquad \bar{\mathcal{B}}_L = \begin{bmatrix} \alpha_L & \beta_L \end{bmatrix}, \qquad \bar{\mathcal{B}}_R = \begin{bmatrix} \alpha_R & \beta_R \end{bmatrix}.$$

In this case, the factorization of the matrix $\bar{\mathcal{A}}$ can be parameterized as

$$\bar{\mathcal{A}} = \bar{Z}\,\bar{\Delta}\,\bar{Z}^T = \begin{bmatrix} \frac{a+\omega}{2s_1} & \frac{a-\omega}{2s_2} \\ \frac{-\varepsilon}{s_1} & \frac{-\varepsilon}{s_2} \end{bmatrix} \begin{bmatrix} \frac{s_1^2}{\omega} & 0 \\ 0 & -\frac{s_2^2}{\omega} \end{bmatrix} \begin{bmatrix} \frac{a+\omega}{2s_1} & \frac{a-\omega}{2s_2} \\ \frac{-\varepsilon}{s_1} & \frac{-\varepsilon}{s_2} \end{bmatrix}^T, \tag{63}$$

with $\omega > 0$. In particular, if $\omega = \sqrt{a^2+4\varepsilon^2}$ and if $s_{1,2}^2 = \omega(\omega \pm a)/2$, then the above factorization is the eigendecomposition of $\bar{\mathcal{A}}$. The discrete scheme mimicking (62) is

$$\begin{aligned} U_t + aD_1U - \varepsilon D_2U = F &+ H^{-1}(\mu_0 e_0 + \nu_0 S^T e_0)\left(\alpha_L U_0 + \beta_L (SU)_0 - g_L\right) \\ &+ H^{-1}(\mu_N e_N + \nu_N S^T e_N)\left(\alpha_R U_N + \beta_R (SU)_N - g_R\right). \end{aligned} \tag{64}$$

To compute the penalty parameters, we need $\overline{P}_L$ and $\overline{P}_R$. They are given by (61), as $\overline{P}_L = \frac{s_1}{\omega}\left(\alpha_L + \beta_L\frac{a-\omega}{2\varepsilon}\right)$ and $\overline{P}_R = -\frac{s_2}{\omega}\left(\alpha_R + \beta_R\frac{a+\omega}{2\varepsilon}\right)$. Theorem 2 now yields

$$\begin{aligned} \mu_0 &= \frac{-\frac{a+\omega}{2} - q\varepsilon}{\alpha_L + \beta_L\frac{a-\omega}{2\varepsilon} - q\beta_L}, & \nu_0 &= \frac{-\varepsilon}{\alpha_L + \beta_L\frac{a-\omega}{2\varepsilon} - q\beta_L}, \\ \mu_N &= \frac{\frac{a-\omega}{2} - q\varepsilon}{\alpha_R + \beta_R\frac{a+\omega}{2\varepsilon} + q\beta_R}, & \nu_N &= \frac{\varepsilon}{\alpha_R + \beta_R\frac{a+\omega}{2\varepsilon} + q\beta_R}. \end{aligned} \tag{65}$$

Formally $0 < \omega < \infty$ is necessary (since in the limits $\bar{Z}$ becomes singular), but as long as the number of imposed boundary condition does not change or the penalty parameters go to infinity, we can allow $0 \leq \omega \leq \infty$.

Remark 15 For Dirichlet boundary conditions we have $\alpha_{L,R} = 1$ and $\beta_{L,R} = 0$. Translating the penalty parameters for the advection-diffusion case in [1] to the form used here, it can be seen that they are exactly the same. If we instead have $\alpha_L = \frac{|a|+a}{2}$, $\beta_L = -\varepsilon$ at the left boundary and $\alpha_R = \frac{|a|-a}{2}$, $\beta_R = \varepsilon$ at the right boundary, we have boundary conditions of a low-reflecting far-field type. In the limit $\omega \to \infty$, we obtain $\mu_0 = -1$, $\nu_0 = 0$, $\mu_N = -1$ and $\nu_N = 0$. This particular choice corresponds to the penalty $\Sigma = -I$ used in [2,3] for systems with boundary conditions of far-field type.

Remark 16 If $\varepsilon = 0$ in (62) we get the transport equation, and then only one boundary condition should be given instead of two. That means that the derivation of the penalty parameters must be redone accordingly. See [1], where this case is covered.

Remark 17 The results can be extended to the case of varying coefficients. Consider the scalar diffusion problem $\mathcal{U}_t - (\varepsilon\mathcal{U}_x)_x = \mathcal{F}$, where $\varepsilon(x) > 0$, with Dirichlet boundary

conditions at $x = 0$ and $x = 1$. Following [12], we define a narrow-stencil operator mimicking $\partial/\partial x(\varepsilon\partial/\partial x)$ as

$$D_2^{(\varepsilon)} = H^{-1}\left(-A_S^{(\varepsilon)} + (\varepsilon(1)E_N - \varepsilon(0)E_0)S\right)$$

where $A_S^{(\varepsilon)}$ is symmetric and positive semi-definite. It is assumed that $D_2^{(\varepsilon)} = \varepsilon D_2$ holds when ε is constant. The discrete problem becomes

$$\begin{aligned} U_t - D_2^{(\varepsilon)}U = F &+ H^{-1}\left(\mu_0 e_0 + \nu_0 S^T e_0\right)(U_0 - g_L) \\ &+ H^{-1}\left(\mu_N e_N + \nu_N S^T e_N\right)(U_N - g_R)\,. \end{aligned}$$

The continuous problem is self-adjoint, so for dual consistency $L^* = H^{-1}L^T H = L$ is needed, which is fulfilled if $\nu_0 = -\varepsilon(0)$ and $\nu_N = \varepsilon(1)$. Moreover, using $A_S^{(\varepsilon)} \geq \varepsilon_{\min} A_S$, where $\varepsilon_{\min} = \min_{x\in[0,1]} \varepsilon(x)$, it can be shown that the discretization will be stable if we choose $\mu_0 \leq -\frac{q}{\varepsilon_{\min}}\varepsilon(0)^2$ and $\mu_N \leq -\frac{q}{\varepsilon_{\min}}\varepsilon(1)^2$. The superconvergence for functionals has been confirmed numerically and the resulting "best" choices of μ_0 and μ_N are similar to what we obtain in the constant case considered below.

5.1.1 The Heat Equation with Dirichlet Boundary Conditions

We consider the heat equation with Dirichlet boundary conditions, i.e., problem (62) with $a = 0$, $\alpha_{L,R} = 1$ and $\beta_{L,R} = 0$, which we solve using the scheme (64), with the penalty parameters given by (65). To isolate the errors originating from the spatial discretization, we first look at the steady problem. Thus we let $\mathcal{U}_t = 0$ and solve $-\mathcal{U}_{xx} = \mathcal{F}(x)$ numerically. The resulting quantities ρ and η, the solution error $\|e\|_H$ and the functional error $|\mathrm{E}|$ are given (as functions of the factorization parameter ω) in Fig. 1. The spectral radius ρ grows with ω, so we do not want $\omega \to \infty$. On the other hand, the decay rate η shrinks with ω so $\omega \to 0$ should also be avoided. The errors tend to decrease with increasing ω (the errors naturally varies slightly depending on the choice of $\mathcal{F}$ and $\mathcal{G}$, but the example in Fig. 1 shows a typical behavior). Thus the demand for accuracy is conflicting with the demand of keeping ρ small (the aim to maximize η is met before the aim to minimize the errors and is therefore not a limiting factor in this case). Empirically we have found that a good compromise, which gives small errors without increasing the spectral radius dramatically, is obtained using $\omega \approx q\varepsilon$.

From this example, we make an observation. If we use the eigenfactorization, $\omega = \sqrt{a^2 + 4\varepsilon^2} = 2$. However, in Fig. 1 we see that that choice is not especially good, since the errors then become much larger than if using $\omega = q\varepsilon$ (i.e., $\omega \approx 200$ and $\omega \approx 340$, respectively). In some cases, the difference in accuracy is so severe that the choice of factorization parameter ω affects the convergence rate. For the narrow operator with the order $(2, 0)$, the errors behave as $\|e\|_H \sim h^{3/2}$ when using $\omega \sim 1$, whereas we obtain the expected $\|e\|_H \sim h^2$ when using $\omega \sim 1/h$. Similar behaviors are observed also for narrow operators of higher order, see below.

In Fig. 2a the errors $\|e\|_H$ for the operators with interior order 6 are shown. For the narrow scheme, the convergence rate is 4.5 when using $\omega = 2\varepsilon$ and 5.5 when using $\omega = q\varepsilon$. For the wide scheme, the order is 4 in both cases, but the error constant changes. In the 8th order case, Fig. 2b, the convergence rates are not affected, but in the narrow case the errors are around 2500 times smaller when using $\omega = q\varepsilon$. In this example, the functional errors are not as sensitive to ω as the solution errors. In the 6th order case, the convergence rates are slightly better than the predicted $2p = 6$, both for the wide and the narrow schemes, see Fig. 3a. For

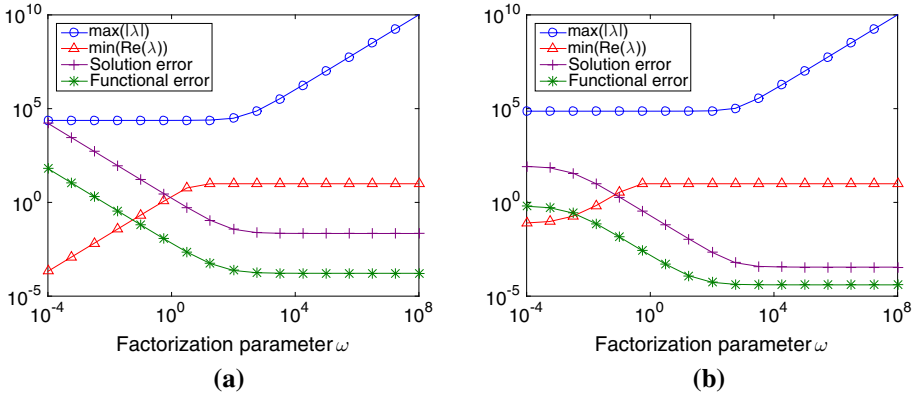

Fig. 1 Properties of L and errors when solving $-\mathcal{U}_{xx} = \mathcal{F}(x)$ with Dirichlet boundary conditions. The number of grid points is $N = 64$, the second derivative operator is either wide or narrow and is 6th order accurate in the interior. Here $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$. **a** Wide-stencil operator. **b** Narrow-stencil operator

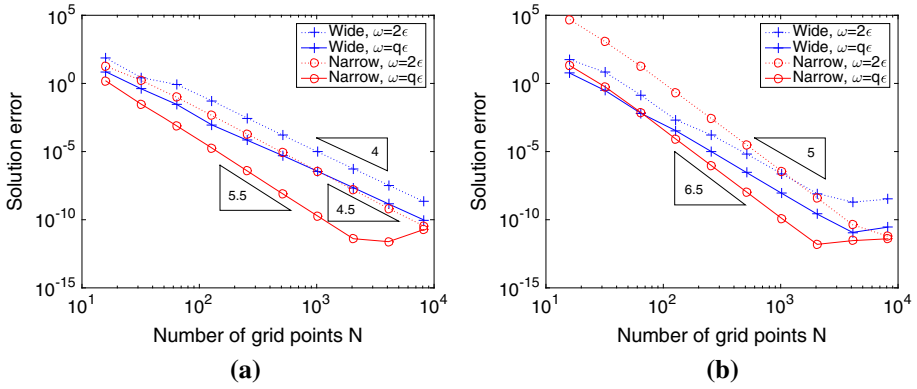

Fig. 2 The error $\|e\|_H$, for $-\mathcal{U}_{xx} = \mathcal{F}(x)$. The exact solution is $\mathcal{U} = \cos(30x)$. **a** Interior order 6. **b** Interior order 8

the 8th order case, see Fig. 3b, the convergence rates are in all cases higher than $2p = 8$. Thus the derived SAT parameters actually produce superconvergent functionals, also for the narrow operators.

Remark 18 For the time-dependent case, i.e., the actual heat equation, the superconvergence is confirmed both when using Dirichlet and additionally when using Neumann boundary conditions. This is presented in "Appendix D".

5.1.2 The Advection–Diffusion Equation with Dirichlet Boundary Conditions

For simplicity we consider steady problems again, this time $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx} + \mathcal{F}$. That is, we solve (62) using the scheme (64), both with omitted time derivatives. The penalty parameters for Dirichlet boundary conditions are given in (65) with $\alpha_{L,R} = 1$ and $\beta_{L,R} = 0$.

First, we take a look at an interesting special case, namely when $\mathcal{F} = 0$. Then the exact solution is $\mathcal{U}(x) = c_1 + c_2 \exp(ax/\varepsilon)$, where the constants c_1 and c_2 are determined by

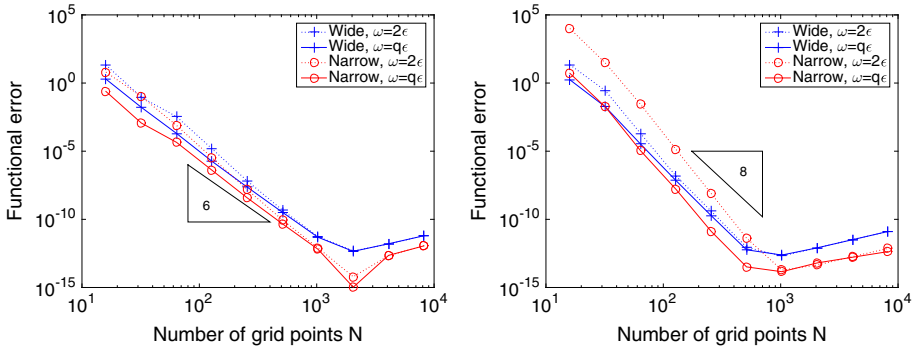

Fig. 3 The functional error $|\mathrm{E}|$, using the weight function $\mathcal{G}(x) = \cos(30x)$. **a** Interior order 6. **b** Interior order 8

the boundary conditions. For $\varepsilon \ll |a|$ the exact solution forms a thin boundary layer at the outflow boundary, which for insufficient resolution usually leads to oscillations in the numerical solution. This can be handled by upwinding or artificial diffusion (see e.g., [16]). Here we will instead use the free parameter ω in the penalty to minimize the oscillating modes (the so-called π-modes).

We start with the wide second derivative operators. The ansatz $U_i = k^i$, inserted into the interior of the scheme (64), gives (for the second order case) a numerical solution

$$U_i = \widetilde{c}_1 + \widetilde{c}_2(-1)^i + \widetilde{c}_3 k_3^i + \widetilde{c}_4 k_4^i, \qquad k_{3,4} = \frac{ha}{\varepsilon} \pm \sqrt{\frac{h^2a^2}{\varepsilon^2} + 1}.$$

Thus there exist two modes with alternating signs, $\widetilde{c}_2(-1)^i$ and $\widetilde{c}_4 k_4^i$. However, one can show that the choice $\omega = |a|$ leads to $\widetilde{c}_2 = 0$ and to $\widetilde{c}_4$ being small enough compared to $\widetilde{c}_3$ such that U_i is monotone. Empirically we have seen that this nice behavior holds also for the wide schemes with higher order of accuracy. In Fig. 4 the result using the scheme with interior order 8 is shown. The solution obtained using $\omega = |a|$ shows no oscillations in the interior, even though the grid is very coarse (the small over-shoot that can be observed close to the boundary layer when $\omega = |a|$ has nothing to do with the π-mode since it is in the zone where the stencil is modified due to the boundary closure). Moreover, this particular choice of factorization gives functional errors almost at machine precision (although it should be noted that this is a special case since $\mathcal{F}(x) = 0$ and $\mathcal{G}(x) = 1$).

For the narrow-stencil schemes, the existence of spurious oscillating modes depends on the resolution. In the second order case, the interior solution is

$$U_i = \widetilde{\widetilde{c}}_1 + \widetilde{\widetilde{c}}_2 \left(\frac{1 + ah/(2\varepsilon)}{1 - ah/(2\varepsilon)} \right)^i,$$

which has an oscillating component if $|a|h/(2\varepsilon) > 1$. With very particular choices of the penalty parameters this component can be canceled (for the operators with order $(2, 0)$ and $(2, 1)$ it is achieved using $\omega = |a|/(1 - \frac{2\varepsilon}{|a|h})$ and $\omega = |a|(1 - \frac{\varepsilon}{|a|h})/(1 - \frac{2\varepsilon}{|a|h})^2$, respectively) such that the numerical solution becomes constant. As soon as $|a|h/(2\varepsilon) < 1$, this mode should not be canceled anymore, but how to do the transition between the unresolved case and the resolved case is not obvious. For the higher order schemes the ω which cancels the oscillating modes are even more complicated and in some cases negative (i.e., useless). In

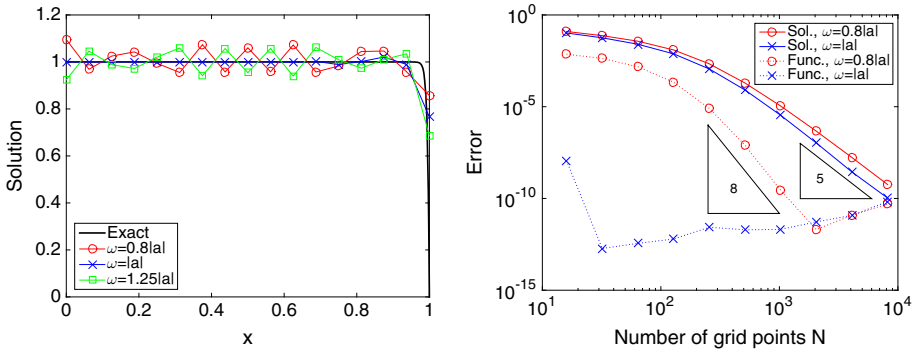

Fig. 4 We solve $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx}$ with $a = 1$, $\varepsilon = 0.005$ using the wide-stencil scheme with interior order 8. **a** The solutions, when the number of grid points is $N = 16$. **b** The solution errors $\|e\|_H$ and functional errors $|\mathrm{E}|$. The weight function is $\mathcal{G}(x) = 1$

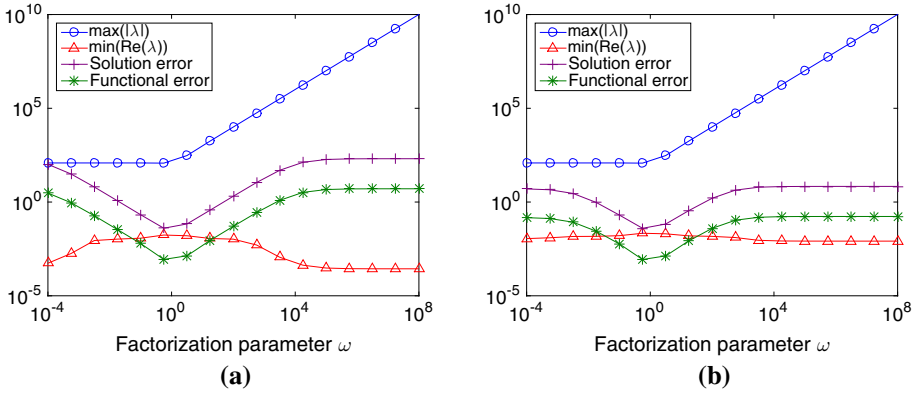

Fig. 5 We solve $a\mathcal{U}_x = \varepsilon\mathcal{U}_{xx} + \mathcal{F}(x)$ with Dirichlet boundary conditions and with $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$. The number of grid points is $N = 64$, the interior order is 6. **a** Wide-stencil operator. **b** Narrow-stencil operator

short, these particular, canceling choices of ω are not worth the effort. Instead, we recommend to use $\omega \approx |a| + q\varepsilon$ for the narrow-stencil operators, see below.

The above results were obtained under the assumption $\mathcal{F} = 0$. Next, we use a forcing function $\mathcal{F}$ such that the exact solution is $\mathcal{U}(x) = \cos(30x)$. The resulting errors, together with ρ and η, are shown in Fig. 5 for $a = 1$ and $\varepsilon = 10^{-6}$. Clearly, $\omega \approx |a|$ is still a good choice since the errors are small, ρ is not too large and η is maximal. For $\varepsilon \gg |a|h$ the curves are more similar to those in Fig. 1, and $\omega \approx |a| + q\varepsilon$ will be a better choice. In the transition region $\varepsilon \sim |a|h$ we sometimes observe order reduction. This can be seen in Figs. 6 and 7 for the schemes with an interior order of accuracy 6. Figure 6 shows the convergence rates when $\varepsilon = 0.1$, which is large enough for the numerical solution to be well resolved. For the narrow scheme, we see an improved convergence rate for the solution error if $\omega = |a| + q\varepsilon$ is used. The functional output converges with $2p = 6$ for all schemes. Figure 7 shows the convergence rates when ε is decreased to 10^{-4}, such that the numerical solution is badly resolved. For all schemes, except the wide scheme with the particular choice $\omega = |a|$, we see a pre-asymptotic order reduction of the functional.

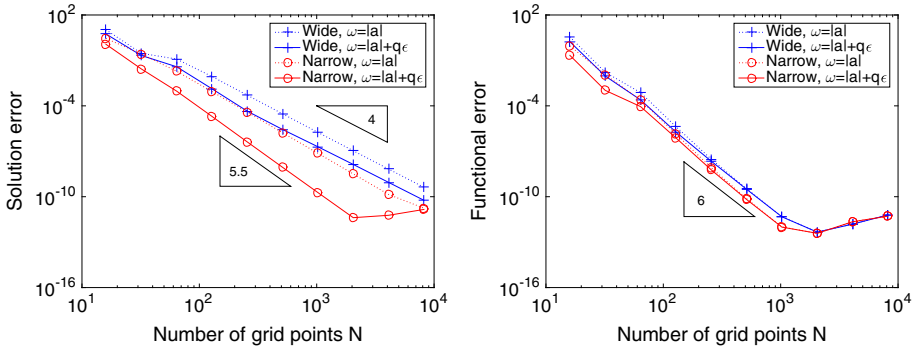

Fig. 6 The inner order of accuracy is 6, $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$, $a = 1$ and $\varepsilon = 0.1$

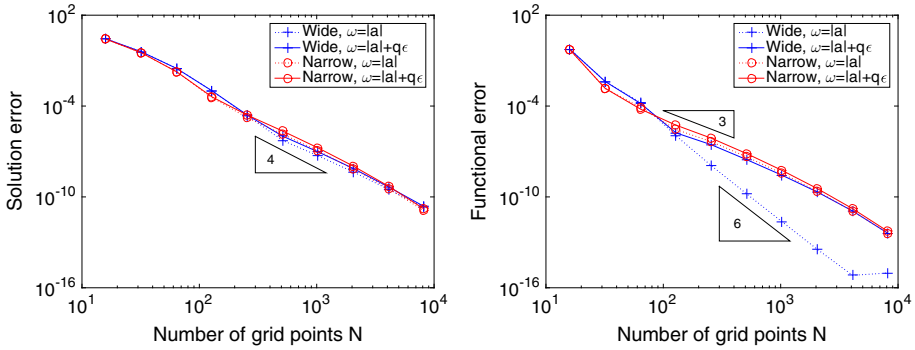

Fig. 7 The inner order of accuracy is 6, $\mathcal{U}(x) = \mathcal{G}(x) = \cos(30x)$, $a = 1$ and $\varepsilon = 10^{-4}$

We conclude that the penalties in Theorem 2 yields superconvergent functionals for the advection-diffusion equation with Dirichlet boundary conditions—in the asymptotic limit. In the special case when having the wide scheme with $\omega = |a|$ we even get superconvergent functionals in the troublesome transition region.

Remark 19 In this case the adjoint problem is $-a\mathcal{V}_x = \varepsilon\mathcal{V}_{xx} + \mathcal{G}$ with (homogeneous) Dirichlet boundary conditions. For a general $\mathcal{G}$, the solution has a thin boundary layer when $\varepsilon \ll |a|$ and the discrete solution will be non-smooth for coarse grids (except for the wide scheme with $\omega = |a|$). If the functional coincidentally is chosen such that $\mathcal{V}$ is "nice", then the dual problem can be resolved even for coarse meshes. When we instead of $\mathcal{G}(x) = \cos(30x)$ use for example $\mathcal{G}(x) = -a\ell\pi\cos(\ell\pi x) + \varepsilon\ell^2\pi^2\sin(\ell\pi x)$, which has the solution $\mathcal{V} = \sin(\ell\pi x)$ if ℓ is an integer, then the functional in Fig. 7 converges with a clear 6th order of accuracy for all four schemes.

5.1.3 Functionals Including Boundary Terms

Consider solving the heat equation with one Dirichlet boundary condition and one Neumann condition, i.e., (62) with $a = 0$, $\alpha_L = 1$, $\beta_L = 0$, $\alpha_R = 0$ and $\beta_R = 1$. If we are using the wide scheme, we can discretize the above problem as a system, by approximating $\mathcal{U}$ by U and $\mathcal{U}_x$ by $\widehat{W}$ as in (67). We use the factorization suggested in (63) to compute the penalty parameters σ_0, τ_0, σ_N and τ_N and obtain the wide semi-discrete scheme

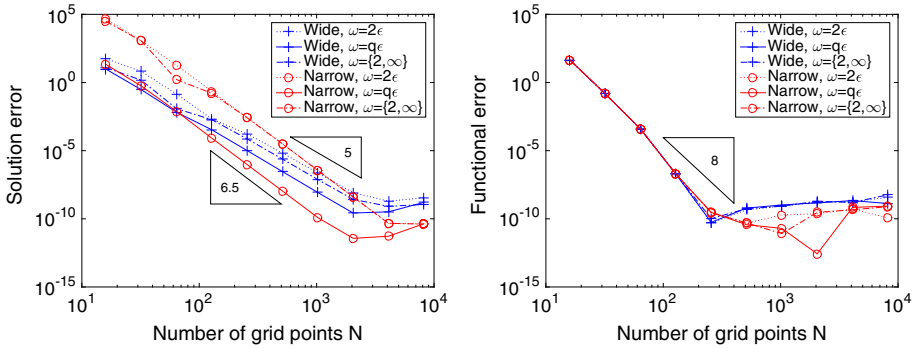

Fig. 8 Errors when solving $-\mathcal{U}_{xx} = \mathcal{F}(x)$ using the schemes with interior order 8. The exact solution is $\mathcal{U} = \cos(30x)$ and the functional is $\mathcal{J}(\mathcal{U}) = \mathcal{U}|_{x=1} + \varepsilon \mathcal{U}_x|_{x=0}$. **a** Solution error $\|\mathbb{e}\|_H$. **b** Functional error $|\mathbb{E}|$

$$U_t - \varepsilon D_1 \widehat{W} = F - H^{-1} e_0 \frac{\omega}{2}(U_0 - g_L) - H^{-1} e_N \varepsilon (\widehat{W}_N - g_R),$$
$$\widehat{W} = D_1 U + H^{-1} e_0 (U_0 - g_L) - H^{-1} e_N \frac{2\varepsilon}{\omega}(\widehat{W}_N - g_R).$$

Now assume that the functional of interest is

$$\mathcal{J}(\mathcal{U}) = \int_0^1 \mathcal{G}_1 \mathcal{U}\, \mathrm{d}x + \int_0^1 \mathcal{G}_2 \mathcal{U}_x\, \mathrm{d}x + \alpha \mathcal{U}|_{x=1} + \beta\varepsilon \mathcal{U}_x|_{x=0}.$$

With $\bar{\mathcal{U}} = \left[\mathcal{U}^T,\, \mathcal{U}_x^T\right]^T$ we identify $\Phi = \left[0,\, \beta\varepsilon\right]$ and $\Psi = \left[\alpha,\, 0\right]$ in (28). Next, using (29) with $\overline{P}_L = s_1/\omega$, $\overline{P}_R = -s_2/(2\varepsilon)$ and $Z_\pm$ given in (63), we obtain

$$J(U, \widehat{W}) = G_1^T H U + G_2^T H \widehat{W} + \alpha U_N + \beta\varepsilon \widehat{W}_0 - \frac{2\varepsilon\alpha}{\omega}(\widehat{W}_N - g_R) + \frac{\beta\omega}{2}(U_0 - g_L).$$

This is the time-dependent and constant coefficient version of (3.11)-(3.14) in [9] (their penalty corresponds to using $\omega = 2$ at the left boundary and $\omega = \infty$ at the right boundary).

We choose $\mathcal{G}_1 = \mathcal{G}_2 = 0$ and $\beta = \alpha = 1$ such that $\mathcal{J}(\mathcal{U}) = \mathcal{U}|_{x=1} + \varepsilon \mathcal{U}_x|_{x=0}$ (we consider the steady case, solving $-\mathcal{U}_{xx} = \mathcal{F}(x)$). Although the penalty-like correction terms are derived for the wide scheme, they work for the narrow-stencil scheme as well, with $\widehat{W}_{0,N}$ replaced by $\widetilde{W}_{0,N}$ from (49). This can be seen in Fig. 8 where we show the result when using the schemes with 8th order inner accuracy and where we by $\omega = \{2, \infty\}$ refer to the penalty choice used in [9].

Remark 20 When using the wide scheme, the $\int_0^1 \mathcal{G}_2 \mathcal{U}_x\, \mathrm{d}x$-part of the functional is approximated by $G_2^T H \widehat{W}$. However, in the narrow case the corresponding W in (47) is not uniquely defined (since M and S are not unique), only $\widetilde{W}_{0,N}$ are. How this kind of functionals best should be approximated is not investigated here, but one workaround is to simply consider $[\mathcal{G}_2 \mathcal{U}]_0^1 - \int_0^1 (\mathcal{G}_2)_x \mathcal{U}\, \mathrm{d}x$ instead of $\int_0^1 \mathcal{G}_2 \mathcal{U}_x\, \mathrm{d}x$.

5.1.4 Reflections From the Scalar Case

From what we have seen from the numerical experiments so far, the best choice of the factorization parameter ω is not only dependent on the continuous problem at hand (i.e., the

parameters a and ε and the type of boundary conditions), but also on numerical quantities, such as the grid resolution and if the stencils are wide or narrow. In some cases the factorization has almost no impact, sometimes it makes the system at hand extremely ill-conditioned or even changes the order of accuracy of the scheme.

In the scalar case it is rather straightforward to optimize with respect to the single factorization parameter ω, but for systems this task becomes non-trivial and one might have to settle for the factorizations at hand. Nevertheless, we note that the eigendecomposition is not necessarily the best factorization and that it could be worth searching for other options. With that being said, next we consider a system and use nothing but the eigendecomposition for constructing the penalty parameters.

5.2 A Fluid Dynamics System with Solid Wall Boundary Conditions

The symmetrized, compressible Navier–Stokes equations in one dimension (with $\Omega = [0, 1]$) with frozen coefficients is given by (30), with

$$\mathcal{A} = \begin{bmatrix} \bar{u} & a & 0 \\ a & \bar{u} & b \\ 0 & b & \bar{u} \end{bmatrix}, \qquad \mathcal{E} = \varepsilon \begin{bmatrix} 0 & 0 & 0 \\ 0 & \varphi & 0 \\ 0 & 0 & \psi \end{bmatrix}, \qquad \mathcal{U} = \begin{bmatrix} \varrho \\ u \\ T \end{bmatrix},$$

where the constants $\bar{u}, a, b, \varepsilon, \varphi$ and ψ denote suitable physical quantities and where ϱ, u and T are scaled perturbations in density, velocity and temperature. Let $\bar{u} < 0$ and $\varepsilon, \varphi, \psi > 0$. In this case, two boundary conditions should be given at the left boundary and three at the right boundary. We impose solid wall boundary conditions (a perfectly insulated wall) at the left boundary, that is $u(0, t) = T_x(0, t) = 0$. At the right boundary, we impose free stream boundary conditions of Dirichlet type, as $\mathcal{U}(1, t) = \mathcal{U}_\infty$. These boundary conditions give a well-posed problem. The boundary operators are

$$\mathcal{H}_L = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{G}_L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathcal{H}_R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathcal{G}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

These boundary conditions can not be rearranged to the far-field form and therefore the penalty used in [2,3] can not be applied. We identify $\bar{\mathcal{A}}$, $\bar{\mathcal{B}}_L$ and $\bar{\mathcal{B}}_R$ according to (32), and factorize $\bar{\mathcal{A}}$ using the eigendecomposition. The dual consistent penalty parameters are now described in (43), with

$$\mathcal{K}_L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/(\varepsilon\psi) \end{bmatrix}, \qquad \mathcal{K}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

As a comparison, we use the alternative penalty parameters (cf. Method 2 in Remark 13)

$$\tilde{\mu}_0 = \begin{bmatrix} -a & 0 \\ 0 & 0 \\ -b & \varepsilon\psi \end{bmatrix}, \quad \tilde{\nu}_0 = \begin{bmatrix} 0 & 0 \\ \varepsilon\varphi & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\mu}_N = \begin{bmatrix} \bar{u} & a & 0 \\ 0 & \bar{u} & 0 \\ 0 & b & \bar{u} \end{bmatrix}, \quad \tilde{\nu}_N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\varepsilon\varphi & 0 \\ 0 & 0 & -\varepsilon\psi \end{bmatrix}$$

which give stability (they are chosen such that the boundary terms in (45) are non-positive for zero data) but they do not fulfill the demands for dual consistency.

In the numerical simulations we use the exact solution $\varrho = \cos(7x)$, $u = \sin(13x)$ and $T = \cos(30x)$ and as weight functions we use $\mathcal{G}(x) = [1,\ 0,\ 0]^T$, $\mathcal{G}(x) = [0,\ 1,\ 0]^T$ and $\mathcal{G}(x) = [0,\ 0,\ 1]^T$ (such that one functional output is obtained for each variable). Figure 9 shows the resulting errors when using the schemes with interior order 6. In the wide case, the

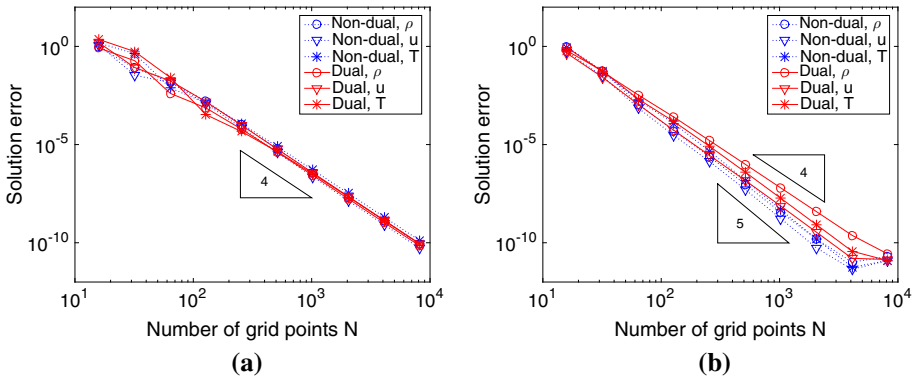

Fig. 9 Solution errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 0.01$. The interior order is 6. **a** Wide-stencil operator. **b** Narrow-stencil operator

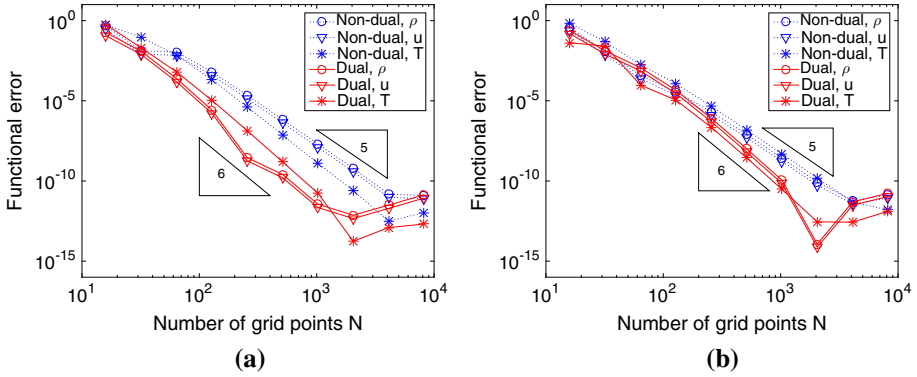

Fig. 10 Functional errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 0.01$. The interior order is 6. **a** Wide-stencil operator. **b** Narrow-stencil operator

solutions do not differ much. In the narrow case, the dual consistent solution converges one half order slower than the dual inconsistent one (order 4 for ϱ and 4.5 for u, T compared to 4.5 for ϱ and 5 for u, T), but the result is still as good as in the wide case. Moreover, recall that in the scalar case the order could be improved by choosing another factorization than the eigendecomposition, see Fig. 6a. In Fig. 10 we see that the functionals convergence with the expected 6th order for both the dual consistent schemes, whereas the dual inconsistent schemes yield 5th order.

The diffusion parameter is decreased from $\varepsilon = 0.01$ to $\varepsilon = 10^{-6}$ and the resulting errors are shown in Figs. 11 and 12. Now the solution errors obtained using the dual consistent schemes are slightly better than the ones obtained using the dual inconsistent schemes, but the difference is small, see Fig. 11.

For the functional errors the difference is more pronounced, see Fig. 12. In the wide case, the dual consistent scheme produces a perfect convergence rate of almost 7. This behavior was observed already in the scalar case, when the factorization parameter was chosen exactly as $\omega = |a|$ (which for small amounts of diffusion is very close to the eigendecomposition). For the narrow-stencil schemes the dual consistent scheme still produces smaller errors than the

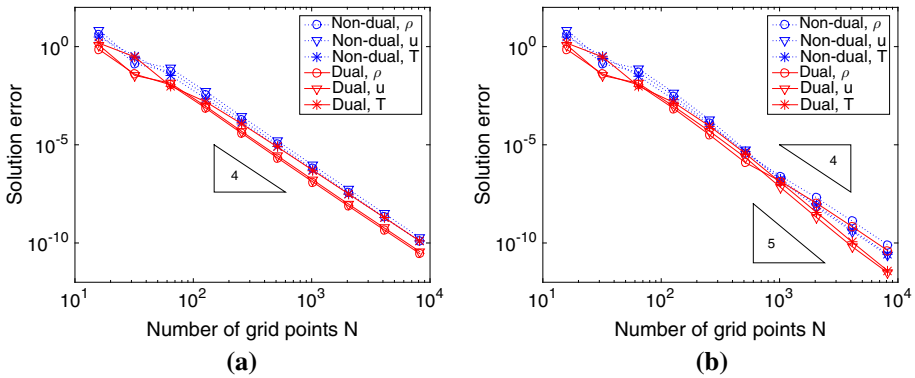

Fig. 11 Solution errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 10^{-6}$. The interior order is 6. **a** Wide-stencil operator. **b** Narrow-stencil operator

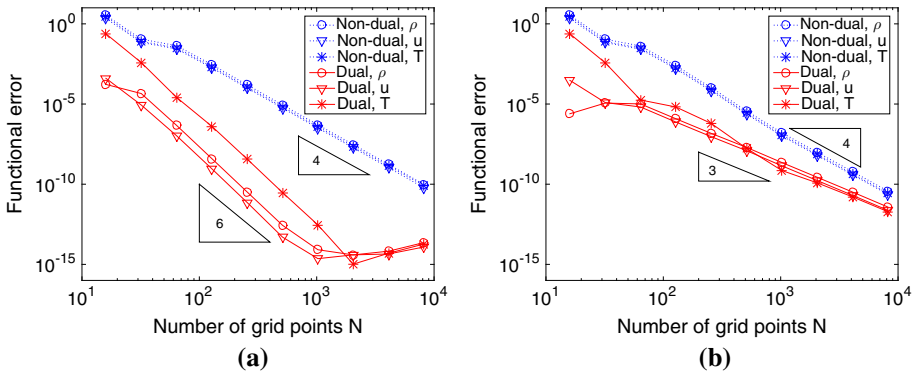

Fig. 12 Functional errors, for $\bar{u} = -0.5$, $a = 0.8$, $b = 0.6$, $\varphi = 1$, $\psi = 2$, $\varepsilon = 10^{-6}$. The interior order is 6. **a** Wide-stencil operator. **b** Narrow-stencil operator

dual inconsistent scheme, but the order is reduced to 3 (a pre-asymptotic low-order tendency seen already in Fig. 7 in the scalar case).

Extrapolating from the scalar case, we assume that it could be worth searching for better penalty parameters for the narrow-stencil schemes when having diffusion dominated problems. However, for convection dominated problems the wide scheme with a factorization close to the eigendecomposition is hard to beat.

6 Concluding Remarks

We use a finite difference method based on summation by parts operators, combined with a penalty method for the boundary conditions (SBP–SAT). Diagonal-norm SBP operators have $2p$-order accurate interior stencils and p-order accurate boundary closures, which limits the global accuracy of the solution to $p + 1$ (or $p + 2$ for parabolic problems under certain conditions). Recently, it has been shown that SBP–SAT schemes can give functional estimates that are $\mathcal{O}(h^{2p})$. To achieve this superconvergence, the SAT parameters must be carefully chosen to ensure that the discretization is dual consistent.

We first look at hyperbolic systems and derive stability requirements and duality constraints for the SATs. Then we present a recipe to choose these SAT parameters such that both these (independent) demands are fulfilled. When wide-stencil second derivative operators are used, the results automatically extend to parabolic problems. We generalize the recipe such that it holds also for narrow-stencil second derivative operators.

The $2p$ order convergence of SBP–SAT functional estimates is confirmed numerically for a variety of scalar examples, as well as for an incompletely parabolic system. For low-diffusion advection-diffusion problems, the superconvergence is sometimes seen first asymptotically. Generally speaking, the narrow-stencil schemes are better for diffusion dominated problems whereas the wide schemes are preferable for advection dominated problems.

In most cases the derived dual consistent SAT parameters have some remaining degree of freedom. The free parameters can be used to improve the accuracy of the primary solution or to tune numerical quantities such as spectral radius, decay rate or condition numbers. Optimal choices of the SAT parameters are suggested for the scalar problems, however, to do the same for systems is considered a task for the future.

Acknowledgements The author would like to sincerely thank the anonymous referees for their valuable comments and suggestions.

Appendix A: Reformulation of the First Order Form Discretization

We derive the scheme (33) with penalty parameters (35), using the hyperbolic results.
Step 1: Consider the problem (31), which is a first order system. We represent the solution $\bar{\mathcal{U}}$ by a discrete solution vector $\overline{U} = [\overline{U}_0^T, \overline{U}_1^T, \dots, \overline{U}_N^T]^T$, where $\overline{U}_i(t) \approx \bar{\mathcal{U}}(x_i, t)$ and discretize (31) exactly as was done in (11) for the hyperbolic case, that is as

$$\begin{aligned}(I_N \otimes \overline{\mathcal{I}})\overline{U}_t + (I_N \otimes \overline{\mathcal{R}})\overline{U} + (D_1 \otimes \bar{\mathcal{A}})\overline{U} = \overline{F} + (H^{-1}e_0 \otimes \overline{\Sigma}_0)(\overline{\mathcal{B}}_L\overline{U}_0 - g_L) \\ + (H^{-1}e_N \otimes \overline{\Sigma}_N)(\overline{\mathcal{B}}_R\overline{U}_N - g_R).\end{aligned} \tag{66}$$

As proposed in Theorem 1, we let $\overline{\Sigma}_0 = -\overline{Z}_+\overline{\Lambda}_+\overline{P}_L^{-1}$ and $\overline{\Sigma}_N = \overline{Z}_-\overline{\Lambda}_-\overline{P}_R^{-1}$.
Step 2: We discretize (30) directly by approximating $\mathcal{U}$ by U and $\mathcal{U}_x$ by $\widehat{W}$. We obtain

$$\begin{aligned}U_t + (D_1 \otimes \mathcal{A})U - (D_1 \otimes \mathcal{E})\widehat{W} = F + \left(H^{-1}e_0 \otimes \sigma_0\right)\left(\mathcal{H}_L U_0 + \mathcal{G}_L\widehat{W}_0 - g_L\right) \\ + \left(H^{-1}e_N \otimes \sigma_N\right)\left(\mathcal{H}_R U_N + \mathcal{G}_R\widehat{W}_N - g_R\right),\end{aligned} \tag{67a}$$

$$\begin{aligned}(I_N \otimes \mathcal{E})\widehat{W} - (D_1 \otimes \mathcal{E})U = \left(H^{-1}e_0 \otimes \tau_0\right)\left(\mathcal{H}_L U_0 + \mathcal{G}_L\widehat{W}_0 - g_L\right) \\ + \left(H^{-1}e_N \otimes \tau_N\right)\left(\mathcal{H}_R U_N + \mathcal{G}_R\widehat{W}_N - g_R\right).\end{aligned} \tag{67b}$$

If $\overline{\Sigma}_0 = [\sigma_0^T, \tau_0^T]^T$ and $\overline{\Sigma}_N = [\sigma_N^T, \tau_N^T]^T$, then (67) is a permutation of (66).
Step 3: The scheme in (67) is a system of differential algebraic equations, so we would like to cancel the variable $\widehat{W}$ and get a system of ordinary differential equations instead. Multiplying (67b) by $\overline{D} = (D_1 \otimes I_n)$ and adding the result to (67a), yields

$$\begin{aligned}U_t + (D_1 \otimes \mathcal{A})U - (D_1^2 \otimes \mathcal{E})U = F + \left(H^{-1}e_0 \otimes \sigma_0 + D_1H^{-1}e_0 \otimes \tau_0\right)\widehat{\chi}_0 \\ + \left(H^{-1}e_N \otimes \sigma_N + D_1H^{-1}e_N \otimes \tau_N\right)\widehat{\chi}_N,\end{aligned}$$

where

$$\widehat{\chi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L\widehat{W}_0 - g_L, \qquad \widehat{\chi}_N = \mathcal{H}_R U_N + \mathcal{G}_R\widehat{W}_N - g_R. \tag{68}$$

Next, using the properties in (12), together with the fact that H is diagonal, we compute

$$D_1 H^{-1} e_0 = H^{-1}\left(-\widehat{q}\, I_N - D_1^T\right) e_0, \qquad D_1 H^{-1} e_N = H^{-1}\left(\widehat{q}\, I_N - D_1^T\right) e_N,$$

where $\widehat{q}$ is the scalar $\widehat{q} = e_0^T H^{-1} e_0 = e_N^T H^{-1} e_N$ given in (37). This yields

$$\begin{aligned} U_t + (D_1 \otimes \mathcal{A})U - \left(D_1^2 \otimes \mathcal{E}\right) U = F + \overline{H}^{-1}\left(e_0 \otimes (\sigma_0 - \widehat{q}\,\tau_0) - D_1^T e_0 \otimes \tau_0\right)\widehat{\chi}_0 \\ + \overline{H}^{-1}\left(e_N \otimes (\sigma_N + \widehat{q}\,\tau_N) - D_1^T e_N \otimes \tau_N\right)\widehat{\chi}_N, \end{aligned} \tag{69}$$

where $\overline{H} = (H \otimes I_n)$. However, the boundary condition deviations $\widehat{\chi}_0$ and $\widehat{\chi}_N$ still contain $\widehat{W}$, so we multiply (67b) by $(e_0^T \otimes I_n)$ and $(e_N^T \otimes I_n)$, respectively, to get

$$\mathcal{E}\widehat{W}_0 - \mathcal{E}(\overline{D}U)_0 = \widehat{q}\,\tau_0 \widehat{\chi}_0, \qquad \mathcal{E}\widehat{W}_N - \mathcal{E}(\overline{D}U)_N = \widehat{q}\,\tau_N \widehat{\chi}_N. \tag{70}$$

Next, we need boundary condition deviations without $\widehat{W}$, and define

$$\widehat{\xi}_0 = \mathcal{H}_L U_0 + \mathcal{G}_L(\overline{D}U)_0 - g_L, \qquad \widehat{\xi}_N = \mathcal{H}_R U_N + \mathcal{G}_R(\overline{D}U)_N - g_R.$$

Recall that $\mathcal{G}_{L,R} = \mathcal{K}_{L,R}\mathcal{E}$. Using (70), we can now relate $\widehat{\xi}_{0,N}$ above to $\widehat{\chi}_{0,N}$ in (68) as

$$\widehat{\xi}_0 = (I_{m_+} - \widehat{q}\,\mathcal{K}_L\tau_0)\widehat{\chi}_0, \qquad \widehat{\xi}_N = (I_{m_-} - \widehat{q}\,\mathcal{K}_R\tau_N)\widehat{\chi}_N, \tag{71}$$

where I_{m_+} and I_{m_-} are identity matrices of sizes corresponding to the number of positive (m_+) and negative (m_-) eigenvalues of $\bar{\mathcal{A}}$, respectively. Inserting $\widehat{\chi}_{0,N}$ from (71) into (69) allows us to finally write the scheme without any $\widehat{W}$ terms and we obtain (33), with

$$\begin{aligned} \widehat{\mu}_0 &= (\sigma_0 - \widehat{q}\,\tau_0)(I_{m_+} - \widehat{q}\,\mathcal{K}_L\tau_0)^{-1}, & \widehat{\nu}_0 &= -\tau_0(I_{m_+} - \widehat{q}\,\mathcal{K}_L\tau_0)^{-1}, \\ \widehat{\mu}_N &= (\sigma_N + \widehat{q}\,\tau_N)(I_{m_-} - \widehat{q}\,\mathcal{K}_R\tau_N)^{-1}, & \widehat{\nu}_N &= -\tau_N(I_{m_-} - \widehat{q}\,\mathcal{K}_R\tau_N)^{-1}. \end{aligned} \tag{72}$$

From Step 1 and 2 we know that

$$\begin{bmatrix} \sigma_0 \\ \tau_0 \end{bmatrix} = -\begin{bmatrix} \overline{Z}_1\overline{\Delta}_+\overline{P}_L^{-1} \\ \overline{Z}_2\overline{\Delta}_+\overline{P}_L^{-1} \end{bmatrix}, \qquad \begin{bmatrix} \sigma_N \\ \tau_N \end{bmatrix} = \begin{bmatrix} \overline{Z}_3\overline{\Delta}_-\overline{P}_R^{-1} \\ \overline{Z}_4\overline{\Delta}_-\overline{P}_R^{-1} \end{bmatrix},$$

where $\overline{Z}_{1,2,3,4}$ are given in (36). Inserting the above relation into (72), we obtain the penalty parameters presented in (35).

Appendix B: Validity of the Derivations and Penalty Parameters in Appendix A

The following Lemma will prove useful:

Lemma 1 (Determinant theorem) *For any matrices A and B of size $m \times n$ and $n \times m$, respectively,* $\det(I_m + AB) = \det(I_n + BA)$ *holds. This lemma is a generalization of the "matrix determinant lemma" and sometimes referred to as "Sylvester's determinant theorem".*

Proof Consider the product of block matrices below:

$$\begin{pmatrix} I_m & 0 \\ B & I_n \end{pmatrix}\begin{pmatrix} I_m + AB & A \\ 0 & I_n \end{pmatrix}\begin{pmatrix} I_m & 0 \\ -B & I_n \end{pmatrix} = \begin{pmatrix} I_m & A \\ 0 & I_n + BA \end{pmatrix}.$$

Using the multiplicativity of determinants, the determinant rule for block triangular matrices and the fact that $\det(I_m) = \det(I_n) = 1$, we see that the determinant of the left hand side is $\det(I_m + AB)$ and that the determinant of the right hand side is $\det(I_n + BA)$. □

When (67) in "Appendix A" is rewritten such that all dependence of $\widehat{W}$ is removed, we rely on the assumption that we can extract $(I_N \otimes \mathcal{E})\widehat{W}$ from (67b) and insert it into (67a). Intuitively we expect this to be possible, since (67) is in fact (although indirectly) a consistent approximation of (30). To investigate this more carefully, we multiply (67b) by $(H \otimes I_n)$ and move all $\widehat{W}$ dependent parts to the left hand side (recall that $\mathcal{G}_{L,R} = \mathcal{K}_{L,R}\mathcal{E}$). This yields

$$\begin{aligned}((H \otimes I_n) - (E_0 \otimes \tau_0 \mathcal{K}_L) - (E_N \otimes \tau_N \mathcal{K}_R))\,(I_N \otimes \mathcal{E})\widehat{W} = (Q \otimes \mathcal{E})U \\ + (e_0 \otimes \tau_0)(\mathcal{H}_L U_0 - g_L) \\ + (e_N \otimes \tau_N)(\mathcal{H}_R U_N - g_R).\end{aligned}$$

We see that we can solve for $(I_N \otimes \mathcal{E})\widehat{W}$ if the matrices $H_{0,0} I_n - \tau_0 \mathcal{K}_L$ and $H_{N,N} I_n - \tau_N \mathcal{K}_R$ are non-singular. From "Appendix A" we know that

$$\tau_0 = -\overline{Z}_2 \overline{\Delta}_+ \overline{P}_L^{-1}, \qquad \tau_N = \overline{Z}_4 \overline{\Delta}_- \overline{P}_R^{-1},$$

that is, we need $I_n + \widehat{q}\,\overline{Z}_2 \overline{\Delta}_+ \overline{P}_L^{-1} \mathcal{K}_L$ and $I_n - \widehat{q}\,\overline{Z}_4 \overline{\Delta}_- \overline{P}_R^{-1} \mathcal{K}_R$ to be non-singular (note that $\widehat{q} = 1/H_{0,0} = 1/H_{N,N}$ for diagonal matrices H). According to Lemma 1 above, we have

$$\begin{aligned}\det\left(I_n + \widehat{q}\,\overline{Z}_2 \overline{\Delta}_+ \overline{P}_L^{-1} \mathcal{K}_L\right) &= \det\left(I_{m_+} + \widehat{q}\,\overline{P}_L^{-1} \mathcal{K}_L \overline{Z}_2 \overline{\Delta}_+\right) = \det\left(\overline{P}_L^{-1}\right) \det\left(\widehat{\Xi}_L\right) \\ \det\left(I_n - \widehat{q}\,\overline{Z}_4 \overline{\Delta}_- \overline{P}_R^{-1} \mathcal{K}_R\right) &= \det\left(I_{m_-} - \widehat{q}\,\overline{P}_R^{-1} \mathcal{K}_R \overline{Z}_4 \overline{\Delta}_-\right) = \det\left(\overline{P}_R^{-1}\right) \det\left(\widehat{\Xi}_R\right).\end{aligned}$$

That is, we can solve for $(I_N \otimes \mathcal{E})\widehat{W}$ in (67b) if the matrices $\widehat{\Xi}_{L,R}$ are non-singular, where $\widehat{\Xi}_{L,R}$ are nothing else than the matrices that shows up in the penalty parameters in (35). So, we are thus interested in the regularity of the matrices

$$\widehat{\Xi}_L = \overline{P}_L + \widehat{q}\,\mathcal{K}_L \overline{Z}_2 \overline{\Delta}_+, \qquad \widehat{\Xi}_R = \overline{P}_R - \widehat{q}\,\mathcal{K}_R \overline{Z}_4 \overline{\Delta}_-,$$

which are inverted in (35)—or if $\widehat{q}$ is replaced by q we consider $\Xi_{L,R}$ from (43). Below we show that $\widehat{\Xi}_L$ is non-singular for well-posed problems. First, using (6), (32) and (36) we obtain

$$\left[-\mathcal{E}\ 0_{n,n}\right] \overline{Z}^{-T} = \left[\overline{Z}_2 \overline{\Delta}_+\ 0_{n,m_0}\ \overline{Z}_4 \overline{\Delta}_-\right]$$

and realize that the matrices $\overline{Z}_2$ and $\overline{Z}_4$ scales with $\mathcal{E}$ as $\overline{Z}_2 = \mathcal{E}\widetilde{Z}_2$ and $\overline{Z}_4 = \mathcal{E}\widetilde{Z}_4$, where

$$\widetilde{Z}_2 = \left[I_n\ 0_{n,n}\right] \overline{Z}^{-T} \begin{bmatrix} -I_{m_+} \\ 0_{(m-m_+),m_+} \end{bmatrix} \overline{\Delta}_+^{-1}, \quad \widetilde{Z}_4 = \left[I_n\ 0_{n,n}\right] \overline{Z}^{-T} \begin{bmatrix} 0_{(m-m_-),m_-} \\ -I_{m_-} \end{bmatrix} \overline{\Delta}_-^{-1}.$$

Secondly, using (10), (32) and (36) leads to $\mathcal{G}_L = \overline{P}_L(\overline{Z}_2^T + \overline{R}_L \overline{Z}_4^T)$ and we can rewrite $\widehat{\Xi}_L$ as

$$\widehat{\Xi}_L = \overline{P}_L \left(I_{m_+} + \widehat{q}\left(\widetilde{Z}_2^T + \overline{R}_L \widetilde{Z}_4^T\right) \mathcal{E} \widetilde{Z}_2 \overline{\Delta}_+\right),$$

where we have used that $\mathcal{G}_L = \mathcal{K}_L \mathcal{E}$. Now Lemma 1 yields $\det(\widehat{\Xi}_L) = \det(\overline{P}_L) \det(\Upsilon)$, where $\Upsilon \equiv I_n + \widehat{q}\,\mathcal{E}^{1/2} \widetilde{Z}_2 \overline{\Delta}_+ \left(\widetilde{Z}_2^T + \overline{R}_L \widetilde{Z}_4^T\right) \mathcal{E}^{1/2}$ and where we by $\mathcal{E}^{1/2}$ refer to the principal

square root of $\mathcal{E}$. The permutation matrix $\overline{P}_L$ is invertible but Υ must be checked. Thus we compute

$$\begin{aligned}\Upsilon\Upsilon^T = I_n &+ \widehat{q}^2\mathcal{E}^{1/2}\widetilde{Z}_2\overline{\Delta}_+\left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)\mathcal{E}\left(\widetilde{Z}_2 + \widetilde{Z}_4\overline{R}_L^T\right)\overline{\Delta}_+\widetilde{Z}_2^T\mathcal{E}^{1/2}\\ &+ \widehat{q}\,\mathcal{E}^{1/2}\left(\left(\widetilde{Z}_2 + \widetilde{Z}_4\overline{R}_L^T\right)\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_2\overline{\Delta}_+\left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)\right)\mathcal{E}^{1/2}.\end{aligned}$$

Next, thanks to the condition for well-posedness, $\overline{C}_L = \overline{\Delta}_- + \overline{R}_L^T\overline{\Delta}_+\overline{R}_L \le 0$, we obtain

$$\begin{aligned}0 &\le \left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)^T\overline{\Delta}_+\left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)\\ &= \widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{R}_L^T\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_2\overline{\Delta}_+\overline{R}_L\widetilde{Z}_4^T + \widetilde{Z}_4\overline{R}_L^T\overline{\Delta}_+\overline{R}_L\widetilde{Z}_4^T\\ &\le \widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{R}_L^T\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_2\overline{\Delta}_+\overline{R}_L\widetilde{Z}_4^T - \widetilde{Z}_4\overline{\Delta}_-\widetilde{Z}_4^T.\end{aligned}$$

Inserting this into $\Upsilon\Upsilon^T$ above, gives (recall that $\widehat{q}$ is positive)

$$\begin{aligned}\Upsilon\Upsilon^T \ge I_n &+ \widehat{q}^2\mathcal{E}^{1/2}\widetilde{Z}_2\overline{\Delta}_+\left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)\mathcal{E}\left(\widetilde{Z}_2 + \widetilde{Z}_4\overline{R}_L^T\right)\overline{\Delta}_+\widetilde{Z}_2^T\mathcal{E}^{1/2}\\ &+ \widehat{q}\,\mathcal{E}^{1/2}\left(\widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{\Delta}_-\widetilde{Z}_4^T\right)\mathcal{E}^{1/2}.\end{aligned}$$

We now let $\mathcal{E} = X\Lambda X^T$ be the eigendecomposition of $\mathcal{E}$, with—for simplicity—the eigenvalues sorted as $\Lambda = \mathrm{diag}(\Lambda_+, \Lambda_0)$, where $\Lambda_+ > 0$ and $\Lambda_0 = 0$. Furthermore, we denote

$$X^T\left(\widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{\Delta}_-\widetilde{Z}_4^T\right)X = \begin{bmatrix}\Theta_1 & \Theta_3\\ \Theta_2 & \Theta_4\end{bmatrix},$$

where Θ_1 and Θ_4 have the same sizes as Λ_+ and Λ_0, respectively. Using (32), (36) and (7) leads to $\overline{Z}_2\overline{\Delta}_+\overline{Z}_2^T + \overline{Z}_4\overline{\Delta}_-\overline{Z}_4^T = 0$, that is

$$0 = \overline{Z}_2\overline{\Delta}_+\overline{Z}_2^T + \overline{Z}_4\overline{\Delta}_-\overline{Z}_4^T = \mathcal{E}\left(\widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{\Delta}_-\widetilde{Z}_4^T\right)\mathcal{E} = X\begin{bmatrix}\Lambda_+\Theta_1\Lambda_+ & 0\\ 0 & 0\end{bmatrix}X^T,$$

which means that $\Theta_1 = 0$ must hold. This in turn leads to

$$\mathcal{E}^{1/2}\left(\widetilde{Z}_2\overline{\Delta}_+\widetilde{Z}_2^T + \widetilde{Z}_4\overline{\Delta}_-\widetilde{Z}_4^T\right)\mathcal{E}^{1/2} = X\begin{bmatrix}\Lambda_+^{1/2}\Theta_1\Lambda_+^{1/2} & 0\\ 0 & 0\end{bmatrix}X^T = 0,$$

where we have used that $\mathcal{E}^{1/2} = X\Lambda^{1/2}X^T$. Thus $\Upsilon\Upsilon^T$ is non-singular, since

$$\Upsilon\Upsilon^T \ge I_n + \widehat{q}^2\mathcal{E}^{1/2}\widetilde{Z}_2\overline{\Delta}_+\left(\widetilde{Z}_2^T + \overline{R}_L\widetilde{Z}_4^T\right)\mathcal{E}\left(\widetilde{Z}_2 + \widetilde{Z}_4\overline{R}_L^T\right)\overline{\Delta}_+\widetilde{Z}_2^T\mathcal{E}^{1/2} \ge I_n > 0.$$

It follows that Υ is non-singular, since $\mathrm{rank}(\Upsilon\Upsilon^T) = \mathrm{rank}(\Upsilon)$, and consequently so is $\widehat{\Xi}_L$, since $\det(\widehat{\Xi}_L) = \det(\overline{P}_L)\det(\Upsilon)$. The same derivations can be repeated for the right boundary.

Appendix C: Proof of Proposition 1 and Examples of q

In Proposition 1 we claim that the inverse of $\widetilde{A}_S = A_S + \delta E_0$ has the structure $\widetilde{A}_S^{-1} = J/\delta + K_0$ and that the corners of $\widetilde{M}^{-1} = S\widetilde{A}_S^{-1}S^T$ are independent of δ. We prove this below.

Proof First we make sure that $\widetilde{A}_S$ is non-singular. By numerically investigating the eigenvalue of $\widetilde{A}_S$ which is closest to zero, we see that (for all operators in this paper) it scales almost as $c(\delta)/N$, where $c(\delta)$ is nearly independent of the operator and where $c(\delta) = 0$ only if $\delta = 0$. We now show that the inverse of $\widetilde{A}_S$ is $J/\delta + K_0$. We denote the parts of A_S, $\widetilde{A}_S$, K_0 and J

$$A_S = \begin{bmatrix} a & \mathbf{a}^T \\ \mathbf{a} & \bar{A} \end{bmatrix}, \qquad \widetilde{A}_S = \begin{bmatrix} a+\delta & \mathbf{a}^T \\ \mathbf{a} & \bar{A} \end{bmatrix}, \qquad K_0 = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \bar{A}^{-1} \end{bmatrix}, \qquad J = \begin{bmatrix} 1 & \mathbf{1}^T \\ \mathbf{1} & \bar{J} \end{bmatrix},$$

where $\mathbf{a}$ and $\mathbf{1} = [1, 1, \ldots, 1]^T$ are vectors and $\bar{A}$ and $\bar{J}$ are matrices of size $N \times N$. Since A_S consists of consistent difference operators, it yields zero when operating on constants. Therefore, $A_S J = 0$ (because J is an all-ones matrix) and $\mathbf{a} + \bar{A}\mathbf{1} = \mathbf{0}$. Note that the relation $\mathbf{a} + \bar{A}\mathbf{1} = \mathbf{0}$ leads to $\begin{bmatrix} \mathbf{a} & \bar{A} \end{bmatrix} = \bar{A}B$, where $B = \begin{bmatrix} -\mathbf{1} & \bar{I} \end{bmatrix}$ is an $N \times (N+1)$ matrix of rank N, in which case it holds that $\operatorname{rank}(\bar{A}B) = \operatorname{rank}(\bar{A})$. Moreover, since $\widetilde{A}_S$ is non-singular, $\begin{bmatrix} \mathbf{a} & \bar{A} \end{bmatrix}$ must have full rank N. Hence $\operatorname{rank}(\bar{A}) = \operatorname{rank}(\bar{A}B) = N$, i.e., $\bar{A}$ has full rank N and is invertible. Next, due to the structure of K_0, we know that $E_0 K_0 = 0$. Thus we have

$$(A_S + \delta E_0)(J/\delta + K_0) = A_S K_0 + E_0 J = \begin{bmatrix} 1 & \mathbf{a}^T \bar{A}^{-1} + \mathbf{1}^T \\ \mathbf{0} & \bar{I} \end{bmatrix} = I.$$

In the last step we have used that $\mathbf{a}^T + \mathbf{1}^T \bar{A}^T = \mathbf{0}^T$ and that $\bar{A}$ is symmetric.

The first and the last row of the matrix S are consistent difference stencils. We can thus write $S = [\mathbf{s}_0, \times, \mathbf{s}_N]^T$, where the vectors $\mathbf{s}_{0,N}$ have the property $\mathbf{s}_{0,N}^T J = 0$. The interior rows of S are marked by a $\times$ because they are not uniquely defined. We compute

$$\widetilde{M}^{-1} = S\widetilde{A}_S^{-1} S^T = S\,(J/\delta + K_0)\,S^T = \begin{bmatrix} \mathbf{s}_0^T K_0 \mathbf{s}_0 & \times & \mathbf{s}_0^T K_0 \mathbf{s}_N \\ \times & \times & \times \\ \mathbf{s}_N^T K_0 \mathbf{s}_0 & \times & \mathbf{s}_N^T K_0 \mathbf{s}_N \end{bmatrix}.$$

We see that the corner elements of $\widetilde{M}^{-1}$ are independent of δ. We conclude that if S is defined such that it is non-singular, the constants in (42) can be computed using (60). □

As an example, consider the narrow (2, 0) order operator in Table 1, specified by D_2 below and associated with the following matrices H and S

$$D_2 = \frac{1}{h^2}\begin{bmatrix} 0 & 0 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2 & 1 & \\ & & & 0 & 0 & \end{bmatrix}, \quad H = h\begin{bmatrix} 1/2 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1/2 \end{bmatrix},$$
$$S = \frac{1}{h}\begin{bmatrix} -1 & 1 & & & & \\ \times & \times & \times & \times & \times & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ \times & \times & \times & \times & \times & \\ & & & -1 & 1 & \end{bmatrix}. \tag{73}$$

Using (73) and (39) we obtain A_S such that we can compute K_0 and $\widetilde{M}^{-1}$ as

$$K_0 = h\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \dots & N-1 & N-1 \\ 0 & 1 & \dots & N-1 & N \end{bmatrix}, \qquad \widetilde{M}^{-1} = \frac{1}{h}\begin{bmatrix} 1 & \times & \dots & \times & 0 \\ \times & \times & \dots & \times & \times \\ \vdots & \vdots & & \vdots & \vdots \\ \times & \times & \dots & \times & \times \\ 0 & \times & \dots & \times & 1 \end{bmatrix}$$

In this case we get $q_0 = q_N = 1/h$ and $q_c = 0$, such that $q = 1/h$.

In addition to the operator discussed above, we use the diagonal-norm operators in [13]. For the higher order accurate operators found in [13], q varies with N. For example, for the narrow (4, 2) order accurate operator, we have

N	$q_0 h$	$q_c h$	qh
8	3.986350339808304	0.000041141179445	3.986391480987749
9	3.986350339313381	0.000002953803786	3.986353293117168
10	3.986350339310830	0.000000212073570	3.986350551384400
11	3.986350339310817	0.000000015226197	3.986350354537014
12	3.986350339310817	0.000000001093192	3.986350340404008

Since the values do not differ so much, it is practical to use the largest value, the one for $N = 8$, regardless of the number of grid points.

Appendix D: Time-Dependent Numerical Examples

For simplicity we mainly consider stationary numerical examples. Below we give a couple of examples confirming the superconvergence also for time-dependent problems.

The Heat Equation with Dirichlet Boundary Conditions

We consider the heat equation. We solve $\mathcal{U}_t = \varepsilon \mathcal{U}_{xx} + \mathcal{F}(x,t)$ with $\varepsilon = 0.01$ and the exact solution $\mathcal{U}(x,t) = \cos(30x) + \sin(20x)\cos(10t) + \sin(35t)$. For the time propagation the classical 4th order accurate Runge–Kutta scheme is used, with sufficiently small time steps, $\Delta t = 10^{-4}$, such that the spatial errors dominate. In Fig. 13 the errors obtained using the narrow (6, 3) order scheme are shown as a function of time.

The corresponding spatial order of convergence (at time $t = 1$) is shown in Table 2. The simulations confirm the steady results, namely that both $\omega = 2\varepsilon$ and $\omega = q\varepsilon$ give superconvergent functionals but that choosing the factorization parameter as $\omega \sim \varepsilon/h$ improves the solution significantly compared to when using the eigendecomposition.

The Heat Equation with Neumann Boundary Conditions

We solve $\mathcal{U}_t = \varepsilon \mathcal{U}_{xx} + \mathcal{F}(x,t)$ again, but this time with Neumann boundary conditions, and the penalty parameters are now given by (65) with $a = 0$, $\varepsilon = 0.01$, $\alpha_{L,R} = 0$ and $\beta_{L,R} = 1$. In contrast to when having Dirichlet boundary conditions, the spectral radius ρ does not depend so strongly on ω and therefore we can let $\omega \to \infty$ (we can use $\omega = q\varepsilon$ here too, it gives the same convergence rates as $\omega = \infty$). In Table 3 we show the errors and convergence orders (at time $t = 1$) for the same setup as in the previous section, that is when solving using the 4th order Runge–Kutta scheme with $\Delta t = 10^{-4}$ and having the

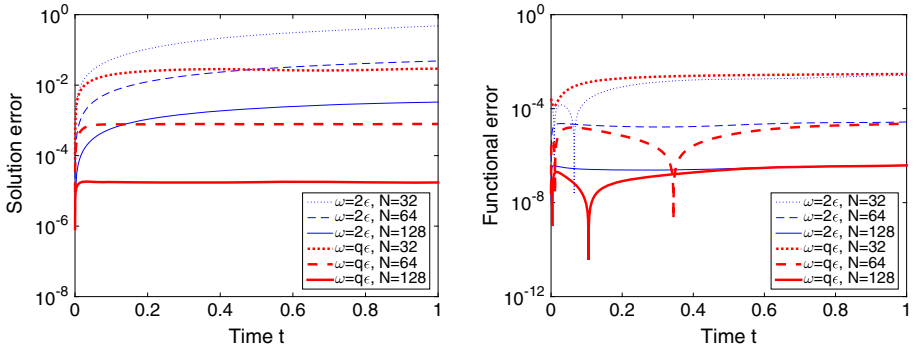

Fig. 13 Errors when solving the heat equation with Dirichlet boundary conditions, using the narrow (6, 3) order scheme. **a** Solution error $\|e\|_H$. **b** Functional error $|E|$ with $\mathcal{G}(x) = 1$

Table 2 The errors and convergence rates at $t = 1$ for the narrow (6,3) order scheme, when using Dirichlet boundary conditions

N	$\omega = 2\varepsilon$				$\omega = q\varepsilon$			
	$\|e\|_H$	Order	$\|E\|$	Order	$\|e\|_H$	Order	$\|E\|$	Order
32	0.480872	–	0.00258741	–	0.029297	–	0.00297573	–
64	0.048501	3.3096	0.00002704	6.5804	0.000790	5.2121	0.00002315	7.0064
128	0.003307	3.8743	0.00000038	6.1559	0.000017	5.5131	0.00000039	5.9055

Table 3 The errors and convergence rates at $t = 1$ for the narrow (6,3) order scheme, when using Neumann boundary conditions

N	$\omega = 2\varepsilon$				$\omega = \infty$			
	$\|e\|_H$	Order	$\|E\|$	Order	$\|e\|_H$	Order	$\|E\|$	Order
32	0.480840	–	0.00286818	–	0.013312	–	0.00286818	–
64	0.048501	3.3095	0.00000621	8.8506	0.000282	5.5629	0.00000621	8.8506
128	0.003307	3.8743	0.00000005	7.0638	0.000006	5.6406	0.00000005	7.0638

exact solution $\mathcal{U}(x, t) = \cos(30x) + \sin(20x)\cos(10t) + \sin(35t)$ and the weight function $\mathcal{G} = 1$. We note that the convergence rates behaves similarly to the Dirichlet case.

References

1. Berg, J., Nordström, J.: Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form. J. Comput. Phys. **231**(20), 6846–6860 (2012)
2. Berg, J., Nordström, J.: On the impact of boundary conditions on dual consistent finite difference discretizations. J. Comput. Phys. **236**, 41–55 (2013)
3. Berg, J., Nordström, J.: Duality based boundary conditions and dual consistent finite difference discretizations of the Navier–Stokes and Euler equations. J. Comput. Phys. **259**, 135–153 (2014)
4. Carpenter, M.H., Nordström, J., Gottlieb, D.: A stable and conservative interface treatment of arbitrary spatial accuracy. J. Comput. Phys. **148**(2), 341–365 (1999)

5. Eriksson, S., Nordström, J.: Analysis of the order of accuracy for node-centered finite volume schemes. Appl. Numer. Math. **59**(10), 2659–2676 (2009)
6. Fernández, D.C.D.R., Hicken, J.E., Zingg, D.W.: Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations. Comput. Fluids **95**, 171–196 (2014)
7. Gustafsson, B., Kreiss, H.O., Oliger, J.: Time-Dependent Problems and Difference Methods. Wiley, New York (2013)
8. Hicken, J.E.: Output error estimation for summation-by-parts finite-difference schemes. J. Comput. Phys. **231**(9), 3828–3848 (2012)
9. Hicken, J.E., Zingg, D.W.: Superconvergent functional estimates from summation-by-parts finite-difference discretizations. SIAM J. Sci. Comput. **33**(2), 893–922 (2011)
10. Hicken, J.E., Zingg, D.W.: Summation-by-parts operators and high-order quadrature. J. Comput. Appl. Math. **237**(1), 111–125 (2013)
11. Kreiss, H.O., Lorenz, J.: Initial-Boundary Value Problems and the Navier–Stokes Equations. Academic Press, New York (1989)
12. Mattsson, K.: Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. J. Sci. Comput. **51**(3), 650–682 (2012)
13. Mattsson, K., Nordström, J.: Summation by parts operators for finite difference approximations of second derivatives. J. Comput. Phys. **199**(2), 503–540 (2004)
14. Nordström, J., Eriksson, S., Eliasson, P.: Weak and strong wall boundary procedures and convergence to steady-state of the Navier–Stokes equations. J. Comput. Phys. **231**(14), 4867–4884 (2012)
15. Nordström, J., Svärd, M.: Well-posed boundary conditions for the Navier–Stokes equations. SIAM J. Numer. Anal. **43**(3), 1231–1255 (2005)
16. Quarteroni, A., Sacco, R., Saleri, F.: Numerical Mathematics. Springer, Berlin (2000)
17. Strand, B.: Summation by parts for finite difference approximation for d/dx. J. Comput. Phys. **110**(1), 47–67 (1994)
18. Svärd, M., Nordström, J.: On the order of accuracy for difference approximations of initial-boundary value problems. J. Comput. Phys. **218**(1), 333–352 (2006)
19. Svärd, M., Nordström, J.: Review of summation-by-parts schemes for initial-boundary-value problems. J. Comput. Phys. **268**, 17–38 (2014)