Journal of Scientific Computing (2021) 89:30
https://doi.org/10.1007/s10915-021-01606-9

# Inverses of SBP-SAT Finite Difference Operators Approximating the First and Second Derivative

Sofia Eriksson[1]

Received: 19 April 2020 / Revised: 21 June 2021 / Accepted: 26 July 2021 /
Published online: 21 September 2021

**Abstract**
The scalar, one-dimensional advection equation and heat equation are considered. These equations are discretized in space, using a finite difference method satisfying summation-by-parts (SBP) properties. To impose the boundary conditions, we use a penalty method called simultaneous approximation term (SAT). Together, this gives rise to two semi-discrete schemes where the discretization matrices approximate the first and the second derivative operators, respectively. The discretization matrices depend on free parameters from the SAT treatment. We derive the inverses of the discretization matrices, interpreting them as discrete Green's functions. In this direct way, we also find out precisely which choices of SAT parameters that make the discretization matrices singular. In the second derivative case, it is shown that if the penalty parameters are chosen such that the semi-discrete scheme is dual consistent, the discretization matrix can become singular even when the scheme is energy stable. The inverse formulas hold for SBP-SAT operators of arbitrary order of accuracy. For second and fourth order accurate operators, the inverses are provided explicitly.

**Keywords** Finite differences · Summation by parts · Simultaneous approximation term · Discretization matrix inverses · Discrete fundamental solutions · Discrete Green's functions

## 1 Introduction

Consider the time-dependent partial differential equation (1a) below, where $\mathcal{L}$ represents a linear differential operator and $f(x)$ is a forcing function. We assume that some suitable initial condition and—for the moment homogeneous—boundary conditions are given such that we have a well-posed problem. Applying the method of lines, that is discretizing first in space while keeping time continuous, yields a system of ordinary differential equations (1b), where we refer to $L$ as the *discretizarion matrix*.

✉ Sofia Eriksson
sofia.eriksson@lnu.se

[1] Department of Mathematics, Linnaeus University, Växjö, Sweden

$$u_t + \mathcal{L}u = f, \qquad t \geq 0, \quad x \in [0, \ell], \tag{1a}$$

$$\mathbf{v}_t + L\mathbf{v} = \mathbf{f}, \qquad t \geq 0. \tag{1b}$$

We first look at the scalar advection equation and thereafter at the heat equation, both in one spatial dimension. Thus $L$ approximates either the first or the second derivative operator, including boundary treatments.

In this paper, $L$ is obtained using the SBP-SAT finite difference method. This class of finite difference method is based on difference operators fulfilling summation-by-parts (SBP) properties, and is modified by the penalty technique simultaneous approximation term (SAT) for treating the boundary conditions. The SBP operators were first developed for first derivatives [21,29] and then later for second derivatives [7,25] and are designed to facilitate the derivation of energy estimates. A means to impose boundary conditions without destroying these properties is to use SAT [6]. The SATs included in $L$ contain free parameters. We follow the common practice of determining these parameters using the energy method, such that (1b) is guaranteed to be time-stable. Thereafter, any remaining degrees of freedom in the SATs can be used to make the scheme *dual consistent*. Dual consistency is advantageous when computing functionals of the solution, since the order of accuracy of functionals from dual consistent schemes can be higher compared to those from non-dual consistent schemes [18]. For more details about SBP-SAT, see [12,31].

Thanks to the SBP-SAT properties, the discretization matrix can be factorized as $L = H^{-1}K$, where $H$ is a symmetric, positive definite matrix that has the role of a quadrature rule, see [19]. Now consider the steady version of (1a), $\mathcal{L}u = f$. Its solution $u(x)$ may be represented as in (2a) below, where $\mathcal{G}$ is the Green's function. The steady version of (1b) is $L\mathbf{v} = \mathbf{f}$. Solving for $\mathbf{v}$, yields (2b).

$$u(x) = \int_0^{\ell} \mathcal{G}(x, y) f(y)\,\mathrm{d}y, \tag{2a}$$

$$\mathbf{v} = K^{-1}H\mathbf{f}. \tag{2b}$$

With $H$'s role as a quadrature rule in mind, we can see a clear similarity between (2a) and (2b): Since $\mathbf{f}$ approximates $f$ and the multiplication by $H$ approximates the integration, we realize that $K^{-1}$ resembles the Green's function $\mathcal{G}$. It makes sense to refer to $K^{-1}$ as a *discrete Green's function*.

A finite difference analogue of the Green's function was introduced already in the fundamental article [9]. Thereafter, discrete Green's functions appear sporadically in the literature, see for example [8,10] and references therein. E.g. in [4] (and correspondingly in [9] for two-dimensional problems) the finite formula approximating (2a) is scaled with the spatial mesh size $h$, which then corresponds closely to (2b). However, since traditional finite difference stencils usually do not have an assigned quadrature rule in the same sense as the SBP operators, the term "discrete Green's functions" often refers to $L^{-1}$ rather than to $K^{-1}$, for example in [5,8,28].

In the above-mentioned articles, the standard way of enforcing boundary conditions, *injection*, has been used instead of SAT (for descriptions of these two boundary methods, see for example [31]). In [14], the first and second derivatives were approximated using an SBP-SAT finite volume method, the inverses analogous to $K^{-1}$ were derived and used for analysing errors. Here, we derive formulas for $K^{-1}$ corresponding to the first and second derivatives as well, however, as an extension to the results in [14], our formulas hold for arbitrary orders of accuracy and in the second derivative case we consider general Robin boundary conditions instead of only Dirichlet boundary conditions.

The inverses are full matrices and are therefore probably not competitive for solving systems $L\mathbf{v} = \mathbf{f}$ directly, compared to fast solvers for banded matrices. It is however often advisable to use pre-conditioning to improve the convergence of iterative methods [16]. A preconditioning matrix $P$ should ideally approximate the inverse of $L$ in some sense, and knowledge about the structure of the inverses could—speculatively—be used when designing preconditioning matrices. If $P$ is a sparse approximate inverse, the computations are cheap, but preconditioners $P$ may also be essentially dense matrices, as for example the fundamental solution preconditioners considered in [5].

The paper is organized as follows: in Sect. 2, we look at the semi-discrete scheme approximating the advection equation. The matrix $K$ associated with $\frac{\partial}{\partial x}$ is denoted $\widetilde{Q}$, and its inverse is presented in Theorem 2.1. In Sect. 3, we consider the heat equation, thus approximating $\frac{\partial^2}{\partial x^2}$. The related matrix $K$, denoted $\widetilde{A}$, is inverted in Theorem 3.1. The SAT parameters are chosen to give stability and dual consistency, and additionally it is of interest to know if some choices of SAT parameters result in a singular discretization matrix $L$. In the second derivative case, it turns out that an energy stable scheme can actually have a singular $L$ if the scheme is also dual consistent. Some relations between stability, dual consistency and a singular discretization matrix are discussed in Sect. 3.3. We also discuss the relations between two different ways of showing energy stability, in Sect. 3.4. The paper is summarized in Sect. 4.

## 2 The First Derivative

Consider the scalar advection equation with a Dirichlet boundary condition at the inflow boundary, that is

$$\begin{aligned} u_t + u_x &= f, \quad x \in [0, \ell], \\ u &= g_{\mathrm{L}}, \; x = 0, \end{aligned} \tag{3}$$

valid for $t \geq 0$, with initial condition $u(x, 0) = u_0(x)$. The forcing function $f(x, t)$, the initial data $u_0(x)$ and the boundary data $g_{\mathrm{L}}(t)$ are known functions.

We call (3) well-posed if it has a unique solution and is stable (can be bounded by data). Techniques for showing existence and uniqueness can be found in for example [17,20]. We focus on showing stability, since we will derive a corresponding stable discrete problem later. We use the energy method, and multiply the partial differential equation in (3) by $u$, and integrate over the spatial domain. Thereafter, we use integration by parts and apply the boundary condition. For simplicity, we consider the homogeneous case, that is with the data $f = 0$ and $g_{\mathrm{L}} = 0$. This yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 = -u(\ell, t)^2$$

where $\|u\|^2 = \int_0^\ell u^2 \, \mathrm{d}x$ and where we have used that $(u^2)_t = 2uu_t$. In the homogeneous case, the growth rate thus becomes $\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 \leq 0$. Integrating this in time yields the energy estimate $\|u\|^2 \leq \|u_0\|^2$ and the solution is thus bounded. Since (3) is an one-dimensional hyperbolic problem it is also possible to show strong well-posedness, i.e., that $\|u\|$ is bounded by the data $f$, $g_{\mathrm{L}}$ and $u_0$. See [17,20] for different definitions of well-posedness.

## 2.1 The Semi-discrete Scheme

We first discretize in space, on the interval $x \in [0, \ell]$, using $n+1$ equidistant grid points $x_i = ih$, where $h = \ell/n$ and $i = 0, 1, \dots, n$. Using the SBP-SAT finite difference method, we obtain a semi-discrete scheme approximating (3) as

$$\mathbf{v}_t + D_1\mathbf{v} = \mathbf{f} + H^{-1}\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}\left(\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\mathbf{v} - g_{\mathrm{L}}\right), \tag{4}$$

where $\mathbf{v}(t) = [v_0, v_1, \dots, v_n]^{\mathsf{T}}$ is the approximation of the continuous solution $u(x,t)$, and where $\mathbf{f} = [f(x_0,t), f(x_1,t), \dots, f(x_n,t)]^{\mathsf{T}}$ is the restriction of $f(x,t)$ to the grid. In the same way, we let the initial data be $\mathbf{v}(0) = [u_0(x_0), u_0(x_1), \dots, u_0(x_n)]^{\mathsf{T}}$. The matrix $D_1$ approximates the first derivative operator $\partial/\partial x$, and fulfills the SBP-properties [21,29]

$$D_1 = H^{-1}Q, \qquad H = H^{\mathsf{T}} > 0, \qquad Q + Q^{\mathsf{T}} = \mathbf{e}_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} - \mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} \tag{5}$$

where $\mathbf{e}_{\mathrm{L}} = [1, 0, \dots, 0]^{\mathsf{T}}$ and $\mathbf{e}_{\mathrm{R}} = [0, \dots, 0, 1]^{\mathsf{T}}$. By the notation $>$, we mean that the matrix $H$ is positive definite. As mentioned in the introduction, $H$ has the role of a quadrature rule and $\|\mathbf{v}\|_H^2 \equiv \mathbf{v}^{\mathsf{T}}H\mathbf{v}$ approximates the $L^2$-norm of $u(x,t)$, see [19]. The scalar $\sigma_{\mathrm{L}}$ determines the strength of the SAT, and will be chosen below such that the scheme (4) is energy stable and dual consistent.

### 2.1.1 Stability and Dual Consistency

To show energy stability, we multiply (4) by $\mathbf{v}^{\mathsf{T}}H$ from the left and use the relations (5). We thereafter add the transpose, and we consider $\mathbf{f} = \mathbf{0}$ and $g_{\mathrm{L}} = 0$, just as in the continuous case. This yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{v}\|_H^2 = -v_n^2 + (1 + 2\sigma_{\mathrm{L}})v_0^2,$$

where $v_0 = \mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\mathbf{v}$ and $v_n = \mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\mathbf{v}$. We need $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{v}\|_H^2 \leq 0$, which is guaranteed if $\sigma_{\mathrm{L}} \leq -1/2$. For a dual consistent scheme, we need $\sigma_{\mathrm{L}} = -1$, see [3,18].

## 2.2 The Inverse of the Discretization Matrix

We first rewrite (4) as

$$\mathbf{v}_t + H^{-1}\widetilde{Q}\mathbf{v} = \widetilde{\mathbf{f}}, \tag{6}$$

where

$$\widetilde{Q} = Q - \sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}, \qquad \widetilde{\mathbf{f}} = \mathbf{f} - H^{-1}\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}g_{\mathrm{L}}. \tag{7}$$

We identify $\widetilde{Q}$ as the first derivative version of $K$ discussed in the introduction. The second order accurate version of $\widetilde{Q}$ was inverted in [14] and inspired by those results, we make a similar ansatz and derive $\widetilde{Q}^{-1}$ of arbitrary order of accuracy. The result is given in Theorem 2.1.

**Theorem 2.1** *Consider the $(n+1) \times (n+1)$-matrices $Q$ from (5) and $\widetilde{Q}$ found in (7). The structures of $Q$ and $\widetilde{Q}$ are*

$$Q = \begin{bmatrix} -1/2 & \vec{q}^{\mathsf{T}} \\ -\vec{q} & \underline{Q} \end{bmatrix}, \quad \widetilde{Q} = \begin{bmatrix} -1/2 - \sigma_{\mathrm{L}} & \vec{q}^{\mathsf{T}} \\ -\vec{q} & \underline{Q} \end{bmatrix}, \tag{8}$$

*where $\vec{q}$ is an $n \times 1$-vector and $\overline{Q}$ is an $n \times n$-matrix. It is assumed that $\overline{Q}$ is invertible. The inverse of $\widetilde{Q}$ is*

$$\widetilde{Q}^{-1} = G_1 - \frac{1}{\sigma_\mathrm{L}} \mathbf{1}\mathbf{b}^\mathsf{T}, \tag{9}$$

*where*

$$G_1 = \begin{bmatrix} 0 & \vec{0}^\mathsf{T} \\ \vec{0} & \overline{Q}^{-1} \end{bmatrix}, \qquad \mathbf{1} = [1, 1, \dots, 1]^\mathsf{T}, \qquad \mathbf{b}^\mathsf{T} = \begin{bmatrix} 1 & -\vec{q}^\mathsf{T}\overline{Q}^{-1} \end{bmatrix}. \tag{10}$$

***Proof of Theorem 2.1*** We aim to show that $\widetilde{Q}\widetilde{Q}^{-1} = I$, where $I$ is the $(n+1) \times (n+1)$ identity matrix. Using $\widetilde{Q}$ from (7) and $\widetilde{Q}^{-1}$ from (9), we compute

$$\begin{aligned} \widetilde{Q}\widetilde{Q}^{-1} &= \left(Q - \sigma_\mathrm{L}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}\right)\left(G_1 - \frac{1}{\sigma_\mathrm{L}}\mathbf{1}\mathbf{b}^\mathsf{T}\right) \\ &= QG_1 - \frac{1}{\sigma_\mathrm{L}} Q\mathbf{1}\mathbf{b}^\mathsf{T} - \sigma_\mathrm{L}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}G_1 + \mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{1}\mathbf{b}^\mathsf{T}. \end{aligned}$$

Note that $D_1\mathbf{1} = 0$, since $D_1$ in (5) is a consistent difference operator. Hence, $Q\mathbf{1} = \mathbf{0}$. Furthermore, $\mathbf{e}_\mathrm{L}^\mathsf{T}G_1 = \mathbf{0}^\mathsf{T}$ since the first row of $G_1$ consists of zeros. These relations, the fact that $\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{1} = 1$ and the structures of the components in (8) and (10) yields

$$\widetilde{Q}\widetilde{Q}^{-1} = QG_1 + \mathbf{e}_\mathrm{L}\mathbf{b}^\mathsf{T} = \begin{bmatrix} 0 & \vec{q}^\mathsf{T}\overline{Q}^{-1} \\ \vec{0} & \overline{I} \end{bmatrix} + \begin{bmatrix} 1 & -\vec{q}^\mathsf{T}\overline{Q}^{-1} \\ \vec{0} & 0 \end{bmatrix} = I$$

where $\overline{I}$ is the $n \times n$ identity matrix. □

**Corollary 2.2** *The structure of $\widetilde{Q}^{-1}$ in (9) implies that $\widetilde{Q}$ is singular only if $\sigma_\mathrm{L} = 0$.*

The existence of $G_1$ and $\mathbf{b}$ in (10), and consequently the validity of Theorem 2.1 and Corollary 2.2, rely on the assumption that $\overline{Q}$ is invertible. In the (2,1) order accurate case—where we by the notation "(2,1) order accurate", refer to a matrix $D_1$ which has second order of accuracy in the interior finite difference stencil and first order of accuracy at the boundaries—the inverse of $\overline{Q}$ is derived and presented in "Appendix A.1", which directly proves its existence. The same is done for the inverse of the "Section A.2" of Appendix order accurate operator, which is presented in "Section A.2" of Appendix. Higher order operators, on the other hand, have free parameters. For example, for the diagonal norm (6,3) order accurate version of $D_1$ described in [29], $x_1$ is a free parameter. In this case, $\widetilde{Q}$ is invertible for commonly used parameter values $x_1$, see [27]. The invertibility of $\widetilde{Q}$ is also addressed for general SBP operators in [22], where it is shown that $\widetilde{Q}$ (with $\sigma_\mathrm{L} = -1$) is invertible if and only if $\mathbf{1}$ spans the nullspace of $D_1$.

The discussion above is focused on "classical FD-SBP operators", constructed around centred finite difference approximations with diagonal matrices $H$. However, Theorem 2.1 only requires consistency (such that $Q\mathbf{1} = 0$) and that the SAT makes $\widetilde{Q} = Q - \sigma_\mathrm{L}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}$. Thus it holds for a more general class of SBP operators where the boundary nodes are included in the operator, compare Definition 1 in [11]—as long as the corresponding $\overline{Q}$ is invertible. Moreover, in Theorem 2.1 it is implied that $Q + Q^\mathsf{T} = \mathbf{e}_\mathrm{R}\mathbf{e}_\mathrm{R}^\mathsf{T} - \mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}$, but this is not crucial for the proof and the result applies also for e.g. upwind operators.

***Remark 2.3*** For the steady version of (3), that is $u_x = f$ with $u(0) = g_\mathrm{L}$, we have

$$u(x) = g_\mathrm{L} + \int_0^\ell \mathcal{G}(x, y) f(y)\,\mathrm{d}y, \qquad \mathcal{G}(x, y) = \begin{cases} 1, & y < x, \\ 0, & x \le y, \end{cases}$$

where $\mathcal{G}$ is a Green's function. Starting from $\mathbf{v} = \widetilde{Q}^{-1}H\widetilde{\mathbf{f}}$, using (7) and (9) as well as the relations $\mathbf{b}^\mathsf{T}\mathbf{e}_\mathrm{L} = 1$ and $G_1\mathbf{e}_\mathrm{L} = \mathbf{0}$ deduced from (10), we obtain

$$\mathbf{v} = g_\mathrm{L}\mathbf{1} + \widetilde{Q}^{-1}H\mathbf{f}.$$

Recall from the introduction that $K^{-1} = \widetilde{Q}^{-1}$ resembles $\mathcal{G}$. E.g. the version of $\widetilde{Q}^{-1}$ found in (34) in "Section A.1" of Appendix (which corresponds to the second order accurate operator) is

$$(\widetilde{Q}^{-1})_{i,j} = \begin{cases} 1 - (1 + 1/\sigma_\mathrm{L})(-1)^j, & 0 \le j \le i \le n, \\ (-1)^{i+j} - (1 + 1/\sigma_\mathrm{L})(-1)^j, & 0 \le i \le j \le n. \end{cases}$$

The dual consistent choice $\sigma_\mathrm{L} = -1$ is optimal in the sense that it cancels the oscillations such that $(\widetilde{Q}^{-1})_{i,j} = 1$ for $j \le i$, however $(\widetilde{Q}^{-1})_{i,j} = (-1)^{i+j} \neq 0$ for $i \le j$. If we instead let $\sigma_\mathrm{L} \to -\infty$, interpreted as mimicking the injection treatment, results in $\widetilde{Q}^{-1} = G_1$. By writing the numerical solution as $\mathbf{v} = \mathbf{1}\left(g_\mathrm{L} - \frac{1}{\sigma_\mathrm{L}}\mathbf{b}^\mathsf{T}H\mathbf{f}\right) + G_1H\mathbf{f}$, we see that the constant level of the solution varies when $\sigma_\mathrm{L}$ is tuned. In particular, $\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{v} \to g_\mathrm{L}$ as $\sigma_\mathrm{L} \to -\infty$.

### 2.2.1 Interface SATs

The SBP-SAT methodology is well suited for dividing the computational domain into subdomains, coupled by interfaces [7]. As an example, we discretize (3) again, using two subdomains with the unknowns $\mathbf{v}_\mathrm{A,B}$, coupled such that $\mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{v}_\mathrm{A} \approx \mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{v}_\mathrm{B}$ at the interface. Modifying (4) to this two-subdomain system yields

$$\frac{\mathrm{d}}{\mathrm{d}t}V + \mathbb{H}^{-1}\widetilde{\mathbb{Q}}V = \widetilde{F},$$

with

$$V = \begin{bmatrix} \mathbf{v}_\mathrm{A} \\ \mathbf{v}_\mathrm{B} \end{bmatrix}, \quad \mathbb{H} = \begin{bmatrix} H_\mathrm{A} & 0 \\ 0 & H_\mathrm{B} \end{bmatrix}, \quad \widetilde{\mathbb{Q}} = \begin{bmatrix} \widetilde{Q}_\mathrm{A} - \mu_\mathrm{A}\mathbf{e}_\mathrm{R}\mathbf{e}_\mathrm{R}^\mathsf{T} & \mu_\mathrm{A}\mathbf{e}_\mathrm{R}\mathbf{e}_\mathrm{L}^\mathsf{T} \\ \mu_\mathrm{B}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{R}^\mathsf{T} & Q_\mathrm{B} - \mu_\mathrm{B}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T} \end{bmatrix}, \quad \widetilde{F} = \begin{bmatrix} \widetilde{\mathbf{f}}_\mathrm{A} \\ \mathbf{f}_\mathrm{B} \end{bmatrix},$$

where all quantities with subindex A belongs to the left subdomain and the ones marked with B to the right subdomain. The same vectors $\mathbf{e}_\mathrm{L,R}$ are used in both domains implying that they have the same number of grid points, but that is merely for ease of presentation. In particular, $\widetilde{Q}_\mathrm{A} = Q_\mathrm{A} - \sigma_\mathrm{L}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}$ and $\widetilde{\mathbf{f}}_\mathrm{A} = \mathbf{f}_\mathrm{A} - H_\mathrm{A}^{-1}\sigma_\mathrm{L}\mathbf{e}_\mathrm{L}g_\mathrm{L}$ are modified to impose the boundary condition, and $\mu_\mathrm{A,B}$ are the penalty parameters at the interface. For $\mu_\mathrm{A} - \mu_\mathrm{B} = 1$ with $\mu_\mathrm{A} + \mu_\mathrm{B} \le 0$, the scheme is conservative, dual consistent and stable.

Assume $Q_\mathrm{A,B}\mathbf{1} = \mathbf{0}$, and let $\widetilde{Q}_\mathrm{A}^{-1} = G_\mathrm{A} - \frac{1}{\sigma_\mathrm{L}}\mathbf{1}\mathbf{b}_\mathrm{A}^\mathsf{T}$, and $(Q_\mathrm{B} - \mu_\mathrm{B}\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T})^{-1} = G_\mathrm{B} - \frac{1}{\mu_\mathrm{B}}\mathbf{1}\mathbf{b}_\mathrm{B}^\mathsf{T}$. Then $\mathbb{Q}\mathbb{1} = 0$, where $\mathbb{Q} = \widetilde{\mathbb{Q}}(\sigma_\mathrm{L} = 0)$ and where $\mathbb{1}$ is given below. In this case Theorem 2.1 applies and the inverse of $\widetilde{\mathbb{Q}}$ has the form $\widetilde{\mathbb{Q}}^{-1} = \mathbb{G} - \frac{1}{\sigma_\mathrm{L}}\mathbb{1}\mathbb{b}^\mathsf{T}$, where

$$\mathbb{G} = \begin{bmatrix} G_\mathrm{A} & \frac{\mu_\mathrm{A}}{\mu_\mathrm{B}}G_\mathrm{A}\mathbf{e}_\mathrm{R}\mathbf{b}_\mathrm{B}^\mathsf{T} \\ \mathbf{1}\mathbf{e}_\mathrm{R}^\mathsf{T}G_\mathrm{A} & G_\mathrm{B} - \frac{1}{\mu_\mathrm{B}}\mathbf{1}\mathbf{b}_\mathrm{B}^\mathsf{T} + \frac{\mu_\mathrm{A}}{\mu_\mathrm{B}}\mathbf{1}\mathbf{e}_\mathrm{R}^\mathsf{T}G_\mathrm{A}\mathbf{e}_\mathrm{R}\mathbf{b}_\mathrm{B}^\mathsf{T} \end{bmatrix}, \quad \mathbb{1} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix}, \quad \mathbb{b}^\mathsf{T} = \begin{bmatrix} \mathbf{b}_\mathrm{A}^\mathsf{T} & \frac{\mu_\mathrm{A}\mathbf{b}_\mathrm{A}^\mathsf{T}\mathbf{e}_\mathrm{R}}{\mu_\mathrm{B}}\mathbf{b}_\mathrm{B}^\mathsf{T} \end{bmatrix},$$

are obtained using the formula for inverse of block-matrices together with the relations $\mathbf{e}_\mathrm{L}^\mathsf{T}G_\mathrm{B} = \mathbf{0}^\mathsf{T}$, $G_\mathrm{B}\mathbf{e}_\mathrm{L} = \mathbf{0}$, $\mathbf{b}_\mathrm{B}^\mathsf{T}\mathbf{e}_\mathrm{L} = 1$ and $\mathbf{e}_\mathrm{L,R}^\mathsf{T}\mathbf{1} = 1$.

As in the single domain case, $\widetilde{\mathbb{Q}}^{-1}$ can be interpreted as a discrete Green's function. In particular, we note an interesting behaviour when $\mu_\mathrm{A} = 0$ and $\mu_\mathrm{B} = -1$, i.e. a fully up-wind coupling. Then $\widetilde{\mathbb{Q}}$ is block-triangular, which leads to

$$\widetilde{\mathbb{Q}} = \begin{bmatrix} \widetilde{Q}_{\mathrm{A}} & 0 \\ -\mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} & Q_{\mathrm{B}} + \mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} \end{bmatrix}, \qquad \widetilde{\mathbb{Q}}^{-1} = \begin{bmatrix} \widetilde{Q}_{\mathrm{A}}^{-1} & 0 \\ \mathbf{1}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\widetilde{Q}_{\mathrm{A}}^{-1} & G_{\mathrm{B}} + \mathbf{1}\mathbf{b}_{\mathrm{B}}^{\mathsf{T}} \end{bmatrix}.$$

We see that with up-wind the continuous feature of having $\mathcal{G}(x, y) = 0$ for $x \leq y$ from Remark 2.3 is at least mimicked on block-matrix level.

# 3 The Second Derivative

Now consider the scalar heat equation with Robin boundary conditions, that is

$$\begin{aligned} u_t - u_{xx} &= f, \quad x \in [0, \ell], \\ \alpha_{\mathrm{L}} u - \beta_{\mathrm{L}} u_x &= g_{\mathrm{L}}, \ x = 0, \\ \alpha_{\mathrm{R}} u + \beta_{\mathrm{R}} u_x &= g_{\mathrm{R}}, \ x = \ell, \end{aligned} \tag{11}$$

valid for $t \geq 0$, with initial condition $u(x, 0) = u_0(x)$. The forcing function $f(x, t)$, the initial data $u_0(x)$ and the boundary data $g_{\mathrm{L,R}}(t)$ are known functions.

We multiply the partial differential equation in (11) by $u$ and integrate the result over the spatial domain, with the data put to $f = 0$ and $g_{\mathrm{L,R}} = 0$. Thereafter using integration by parts and the boundary conditions, yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 + 2\|u_x\|^2 = -2\frac{\beta_{\mathrm{R}}}{\alpha_{\mathrm{R}}} u_x(\ell, t)^2 - 2\frac{\beta_{\mathrm{L}}}{\alpha_{\mathrm{L}}} u_x(0, t)^2.$$

For a decaying growth rate, we need $\alpha_{\mathrm{L,R}}\beta_{\mathrm{L,R}} \geq 0$.

## 3.1 The Semi-discrete Scheme

Using the SBP-SAT finite difference method, we obtain a scheme approximating (11) as

$$\begin{aligned} \mathbf{v}_t - D_2\mathbf{v} = \mathbf{f} &+ H^{-1}(\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}} - \tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}})\left(\alpha_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\mathbf{v} - \beta_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\mathbf{v} - g_{\mathrm{L}}\right) \\ &+ H^{-1}(\sigma_{\mathrm{R}}\mathbf{e}_{\mathrm{R}} + \tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}})\left(\alpha_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\mathbf{v} + \beta_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\mathbf{v} - g_{\mathrm{R}}\right), \end{aligned} \tag{12}$$

where $\mathbf{v}$, $\mathbf{f}$, $H$ and $\mathbf{e}_{\mathrm{L,R}}$ are described as in Sect. 2.1. The matrix $D_2$ approximates the second derivative operator, and fulfills the SBP-properties

$$D_2 = H^{-1}(-A + \mathbf{e}_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} - \mathbf{e}_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}), \qquad A = A^{\mathsf{T}} \geq 0. \tag{13}$$

The vectors $\mathbf{d}_{\mathrm{L}}$ and $\mathbf{d}_{\mathrm{R}}$ are consistent finite difference stencils approximating the first derivative, see [7]. Two common categories of $D_2$ operators are *wide-stencil* and *narrow-stencil* operators. Wide-stencil operators can be factorized as $D_2 = D_1^2$, and the term "narrow" describes finite difference schemes with a minimal stencil width [26].

The penalty parameters $\sigma_{\mathrm{L,R}}$ and $\tau_{\mathrm{L,R}}$ in (12) are scalars that will be further specified and discussed in the next sections. Now, we use (13) to rewrite (12) as

$$\mathbf{v}_t + H^{-1}\widetilde{A}\mathbf{v} = \widetilde{\mathbf{f}}, \tag{14}$$

where

$$\widetilde{A} = A - \begin{bmatrix} \mathbf{e}_{\mathrm{L}}^{\mathsf{T}} \\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \sigma_{\mathrm{L}}\alpha_{\mathrm{L}} & 1 + \sigma_{\mathrm{L}}\beta_{\mathrm{L}} \\ \tau_{\mathrm{L}}\alpha_{\mathrm{L}} & \tau_{\mathrm{L}}\beta_{\mathrm{L}} \end{bmatrix} \begin{bmatrix} \mathbf{e}_{\mathrm{L}}^{\mathsf{T}} \\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \end{bmatrix} - \begin{bmatrix} \mathbf{e}_{\mathrm{R}}^{\mathsf{T}} \\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \sigma_{\mathrm{R}}\alpha_{\mathrm{R}} & 1 + \sigma_{\mathrm{R}}\beta_{\mathrm{R}} \\ \tau_{\mathrm{R}}\alpha_{\mathrm{R}} & \tau_{\mathrm{R}}\beta_{\mathrm{R}} \end{bmatrix} \begin{bmatrix} \mathbf{e}_{\mathrm{R}}^{\mathsf{T}} \\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \end{bmatrix} \tag{15}$$

and where $\widetilde{\mathbf{f}} = \mathbf{f} - H^{-1}(\sigma_\mathrm{L}\mathbf{e}_\mathrm{L} - \tau_\mathrm{L}\mathbf{d}_\mathrm{L})g_\mathrm{L} - H^{-1}(\sigma_\mathrm{R}\mathbf{e}_\mathrm{R} + \tau_\mathrm{R}\mathbf{d}_\mathrm{R})g_\mathrm{R}$. We identify $\widetilde{A}$ as the second derivative version of the matrix $K$ from the introduction.

### 3.1.1 Stability

To show energy stability, we multiply (12) by $\mathbf{v}^\mathsf{T} H$ from the left and use the relations (13). We thereafter add the transpose, and let $\mathbf{f} = \mathbf{0}$ and $g_{\mathrm{L,R}} = 0$. This yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{v}\|_H^2 + 2\mathbf{v}^\mathsf{T} A\mathbf{v} = 2\mathbf{v}^\mathsf{T}(\mathbf{e}_\mathrm{R}\mathbf{d}_\mathrm{R}^\mathsf{T} - \mathbf{e}_\mathrm{L}\mathbf{d}_\mathrm{L}^\mathsf{T})\mathbf{v} \\ + 2\mathbf{v}^\mathsf{T}(\sigma_\mathrm{L}\mathbf{e}_\mathrm{L} - \tau_\mathrm{L}\mathbf{d}_\mathrm{L})\left(\alpha_\mathrm{L}\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{v} - \beta_\mathrm{L}\mathbf{d}_\mathrm{L}^\mathsf{T}\mathbf{v}\right) \\ + 2\mathbf{v}^\mathsf{T}(\sigma_\mathrm{R}\mathbf{e}_\mathrm{R} + \tau_\mathrm{R}\mathbf{d}_\mathrm{R})\left(\alpha_\mathrm{R}\mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{v} + \beta_\mathrm{R}\mathbf{d}_\mathrm{R}^\mathsf{T}\mathbf{v}\right),\end{aligned} \tag{16}$$

where we need to show that $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{v}\|_H^2 \leq 0$. We will determine the stability limits of $\sigma_{\mathrm{L,R}}$ and $\tau_{\mathrm{L,R}}$ using a procedure sometimes called the *borrowing technique* [1,2,7,15,24,30,32]. The idea is to "borrow" a maximum amount $\gamma$ of "positivity" from $A$, more precisely as

$$A = \tilde{A}_\gamma + h\gamma(\mathbf{d}_\mathrm{L}\mathbf{d}_\mathrm{L}^\mathsf{T} + \mathbf{d}_\mathrm{R}\mathbf{d}_\mathrm{R}^\mathsf{T}), \qquad \tilde{A}_\gamma \geq 0, \quad \gamma > 0. \tag{17}$$

Inserting the relation in (17) into (16), we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{v}\|_H^2 + 2\mathbf{v}^\mathsf{T}\tilde{A}_\gamma\mathbf{v} = \begin{bmatrix}\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{v} \\ -\mathbf{d}_\mathrm{L}^\mathsf{T}\mathbf{v}\end{bmatrix}^\mathsf{T}\begin{bmatrix}2\sigma_\mathrm{L}\alpha_\mathrm{L} & 1+\sigma_\mathrm{L}\beta_\mathrm{L}+\tau_\mathrm{L}\alpha_\mathrm{L} \\ 1+\sigma_\mathrm{L}\beta_\mathrm{L}+\tau_\mathrm{L}\alpha_\mathrm{L} & 2\tau_\mathrm{L}\beta_\mathrm{L}-2h\gamma\end{bmatrix}\begin{bmatrix}\mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{v} \\ -\mathbf{d}_\mathrm{L}^\mathsf{T}\mathbf{v}\end{bmatrix} \\ + \begin{bmatrix}\mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{v} \\ \mathbf{d}_\mathrm{R}^\mathsf{T}\mathbf{v}\end{bmatrix}^\mathsf{T}\begin{bmatrix}2\sigma_\mathrm{R}\alpha_\mathrm{R} & 1+\sigma_\mathrm{R}\beta_\mathrm{R}+\tau_\mathrm{R}\alpha_\mathrm{R} \\ 1+\sigma_\mathrm{R}\beta_\mathrm{R}+\tau_\mathrm{R}\alpha_\mathrm{R} & 2\tau_\mathrm{R}\beta_\mathrm{R}-2h\gamma\end{bmatrix}\begin{bmatrix}\mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{v} \\ \mathbf{d}_\mathrm{R}^\mathsf{T}\mathbf{v}\end{bmatrix}.\end{aligned}$$

For stability, we need both the matrices in the two quadratic forms above to be negative semi-definite. This is fulfilled if

$$\begin{aligned}2\sigma_{\mathrm{L,R}}\alpha_{\mathrm{L,R}} &\leq 0 \\ 2(\tau_{\mathrm{L,R}}\beta_{\mathrm{L,R}} - h\gamma) &\leq 0 \\ (1+\tau_{\mathrm{L,R}}\alpha_{\mathrm{L,R}}+\sigma_{\mathrm{L,R}}\beta_{\mathrm{L,R}})^2 &\leq 4\sigma_{\mathrm{L,R}}\alpha_{\mathrm{L,R}}(\tau_{\mathrm{L,R}}\beta_{\mathrm{L,R}} - h\gamma).\end{aligned} \tag{18}$$

### 3.1.2 Dual Consistency

To make the scheme (12) dual consistent we first note that the operator $\partial^2/\partial x^2$ (including boundary conditions) is a symmetric operator and that the matrix $\widetilde{A}$ must be symmetric to mimic this. From (15) it is clear that $\widetilde{A}$ is symmetric if $1+\sigma_{\mathrm{L,R}}\beta_{\mathrm{L,R}} = \tau_{\mathrm{L,R}}\alpha_{\mathrm{L,R}}$. Let

$$\delta_\mathrm{L} \equiv 1+\sigma_\mathrm{L}\beta_\mathrm{L} - \tau_\mathrm{L}\alpha_\mathrm{L} \qquad \delta_\mathrm{R} \equiv 1+\sigma_\mathrm{R}\beta_\mathrm{R} - \tau_\mathrm{R}\alpha_\mathrm{R}, \tag{19}$$

where $\delta_{\mathrm{L,R}} = 0$ for dual consistent choices of penalty parameters. The relations in (19), with $\delta_{\mathrm{L,R}} = 0$, can also be derived from the penalty parameters of the scalar problem in [13]. For a background and more thorough descriptions of dual consistency, see [18].

Note that now, using the dual consistency parameters $\delta_{\mathrm{L,R}}$ defined in (19), the three stability requirements in (18) can be reformulated as

$$\sigma_{\mathrm{L,R}}\alpha_{\mathrm{L,R}} \leq 0, \qquad \tau_{\mathrm{L,R}}\beta_{\mathrm{L,R}} \leq h\gamma, \qquad \delta_{\mathrm{L,R}}^2 \leq -4\alpha_{\mathrm{L,R}}(\sigma_{\mathrm{L,R}}h\gamma + \tau_{\mathrm{L,R}}). \tag{20}$$

### 3.2 The Inverse of the Discretization Matrix

We consider the steady version of (14), that is $H^{-1}\widetilde{A}\mathbf{v} = \widetilde{\mathbf{f}}$, which has a unique solution $\mathbf{v} = \widetilde{A}^{-1}H\widetilde{\mathbf{f}}$, if $\widetilde{A}^{-1}$ exists. We derive this inverse and present the result in Theorem 3.1.

**Theorem 3.1** *Consider $\widetilde{A}$ in (15), which depends on $A$ and $\mathbf{d}_{\mathrm{L,R}}$ in (13) and on the boundary related scalars $\sigma_{\mathrm{L,R}}$, $\tau_{\mathrm{L,R}}$, $\alpha_{\mathrm{L,R}}$ and $\beta_{\mathrm{L,R}}$. Let the parts of $A$ be denoted as follows,*

$$A = \begin{bmatrix} a_{\mathrm{L}} & \vec{a}_{\mathrm{L}}^{\mathsf{T}} & a_{\mathrm{C}} \\ \vec{a}_{\mathrm{L}} & \bar{A} & \vec{a}_{\mathrm{R}} \\ a_{\mathrm{C}} & \vec{a}_{\mathrm{R}}^{\mathsf{T}} & a_{\mathrm{R}} \end{bmatrix}, \tag{21}$$

*where $a_{\mathrm{L}}$, $a_{\mathrm{R}}$ and $a_{\mathrm{C}}$ are scalars, $\vec{a}_{\mathrm{L,R}}$ are $(n-1)\times 1$-vectors and $\bar{A}$ is an $(n-1)\times(n-1)$-matrix. The inverse of $\widetilde{A}$ is*

$$\widetilde{A}^{-1} = G_2 + \left[-\tau_{\mathrm{L}}\mathbf{b}_{\mathrm{L}} \;\; -\tau_{\mathrm{R}}\mathbf{b}_{\mathrm{R}} \;\; \mathbf{1}-\mathbf{x}/\ell \;\; \mathbf{x}/\ell\right]\Sigma^{-1}\begin{bmatrix} \mathbf{b}_{\mathrm{L}}^{\mathsf{T}} \\ \mathbf{b}_{\mathrm{R}}^{\mathsf{T}} \\ \beta_{\mathrm{L}}(\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}} \\ \beta_{\mathrm{R}}\mathbf{x}^{\mathsf{T}}/\ell \end{bmatrix} \tag{22}$$

*where $\mathbf{1} = [1\;1\;1\;\ldots\;1]^{\mathsf{T}}$ and $\mathbf{x} = h[0\;1\;2\;\ldots\;n]^{\mathsf{T}}$, and where*

$$G_2 = \begin{bmatrix} 0 & \vec{0}^{\mathsf{T}} & 0 \\ \vec{0} & \bar{A}^{-1} & \vec{0} \\ 0 & \vec{0}^{\mathsf{T}} & 0 \end{bmatrix}, \qquad \mathbf{b}_{\mathrm{L}} \equiv \mathbf{1}-\mathbf{x}/\ell - G_2\mathbf{d}_{\mathrm{L}}, \qquad \mathbf{b}_{\mathrm{R}} \equiv \mathbf{x}/\ell + G_2\mathbf{d}_{\mathrm{R}}. \tag{23}$$

*Furthermore, $\Sigma$ in (22) is a $4\times 4$-matrix*

$$\Sigma = \begin{bmatrix} \sigma_{\mathrm{L}} + \tau_{\mathrm{L}}\xi_{\mathrm{L}} & -\tau_{\mathrm{R}}\xi_{\mathrm{C}} & 0 & 0 \\ -\tau_{\mathrm{L}}\xi_{\mathrm{C}} & \sigma_{\mathrm{R}} + \tau_{\mathrm{R}}\xi_{\mathrm{R}} & 0 & 0 \\ \delta_{\mathrm{L}} & 0 & \alpha_{\mathrm{L}} + \beta_{\mathrm{L}}/\ell & -\beta_{\mathrm{L}}/\ell \\ 0 & \delta_{\mathrm{R}} & -\beta_{\mathrm{R}}/\ell & \alpha_{\mathrm{R}} + \beta_{\mathrm{R}}/\ell \end{bmatrix} \tag{24}$$

*that depends on $\alpha_{\mathrm{L,R}}$ and $\beta_{\mathrm{L,R}}$, that is on the choices of boundary conditions in (11), on the choices of penalty parameters $\sigma_{\mathrm{L,R}}$ and $\tau_{\mathrm{L,R}}$ in (12) and on the duality parameters $\delta_{\mathrm{L,R}}$ in (19), as well as on the scalars*

$$\xi_{\mathrm{L}} \equiv -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\mathbf{b}_{\mathrm{L}}, \qquad \xi_{\mathrm{R}} \equiv \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\mathbf{b}_{\mathrm{R}} \qquad \xi_{\mathrm{C}} \equiv \mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\mathbf{b}_{\mathrm{R}} = -\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\mathbf{b}_{\mathrm{L}}. \tag{25}$$

***Proof of Theorem 3.1*** The proof is given in "Appendix B". □

Note that the quantities in (23), and thus the validity of Theorem 3.1, rely on the existence of $\bar{A}^{-1}$. In "Appendix D", the explicit values of $\bar{A}^{-1}$, as well as of $G_2$, $\mathbf{b}_{\mathrm{L,R}}$, $\xi_{\mathrm{L,R}}$ and $\xi_{\mathrm{C}}$, are provided for the (2,0), (2,1) and (4,2) order accurate narrow-stencil operators and the (2,0) order accurate wide-stencil operator. This directly proves the existence of $\bar{A}^{-1}$ for these operators. Higher order accurate operators have free parameters, but empirically we can draw the conclusion that $\bar{A}^{-1}$ must exist at least for the parameter choices in [25], since the operators therein have been applied successfully for many years.

Given the existence of $\bar{A}^{-1}$, we note that $\widetilde{A}$ in (22) is singular if and only if $\Sigma$ in (24) is singular. The matrix $\Sigma$ is in turn singular if any of the two relations

$$(\alpha_{\mathrm{L}} + \beta_{\mathrm{L}}/\ell)(\alpha_{\mathrm{R}} + \beta_{\mathrm{R}}/\ell) - \beta_{\mathrm{L}}\beta_{\mathrm{R}}/\ell^2 = 0 \tag{26}$$

$$(\sigma_{\mathrm{L}} + \tau_{\mathrm{L}}\xi_{\mathrm{L}})(\sigma_{\mathrm{R}} + \tau_{\mathrm{R}}\xi_{\mathrm{R}}) - \tau_{\mathrm{L}}\tau_{\mathrm{R}}\xi_{\mathrm{C}}^2 = 0 \tag{27}$$

holds. The first condition is related to the continuous boundary conditions, and makes the matrix singular if Neumann boundary conditions are imposed on both boundaries, i.e. if $\alpha_{\mathrm{L}} = \alpha_{\mathrm{R}} = 0$. The second condition has to do with the choice of penalty parameters, and leads us to the following corollary of Theorem 3.1:

**Corollary 3.2** *The matrix $\widetilde{A}$, described in (15), is singular when the penalty parameters simultaneous fulfill $\sigma_{\mathrm{L}} = -\left(\xi_{\mathrm{L}} + \zeta\,|\xi_{\mathrm{C}}|\right)\tau_{\mathrm{L}}$ and $\sigma_{\mathrm{R}} = -\left(\xi_{\mathrm{R}} + |\xi_{\mathrm{C}}|/\zeta\right)\tau_{\mathrm{R}}$, where $\zeta \neq 0$. If $\xi_{\mathrm{C}}$, $\tau_{\mathrm{L}}$ or $\tau_{\mathrm{R}}$ is zero, the matrix $\widetilde{A}$ is singular if either $\sigma_{\mathrm{L}} = -\tau_{\mathrm{L}}\xi_{\mathrm{L}}$ or if $\sigma_{\mathrm{R}} = -\tau_{\mathrm{R}}\xi_{\mathrm{R}}$.*

***Proof of Corollary 3.2*** We make the ansatz $\sigma_{\mathrm{L,R}} = -\tau_{\mathrm{L,R}}\xi_{\mathrm{L,R}} - \varepsilon_{\mathrm{L,R}}$ with some unknown scalars $\varepsilon_{\mathrm{L,R}}$. Inserting this into (27) above gives $\varepsilon_{\mathrm{L}}\varepsilon_{\mathrm{R}} = \tau_{\mathrm{L}}\tau_{\mathrm{R}}\xi_{\mathrm{C}}^2$ which is fulfilled for all pairs $\varepsilon_{\mathrm{L}} = \tau_{\mathrm{L}}|\xi_{\mathrm{C}}|\zeta$ and $\varepsilon_{\mathrm{R}} = \tau_{\mathrm{R}}|\xi_{\mathrm{C}}|/\zeta$ with arbitrary choices of $\zeta \neq 0$. If $\xi_{\mathrm{C}}$, $\tau_{\mathrm{L}}$ or $\tau_{\mathrm{R}}$ is equal to zero, it is enough if either $\varepsilon_{\mathrm{L}} = 0$ or $\varepsilon_{\mathrm{R}} = 0$. □

The requirements on $A$ and $\mathbf{d}_{\mathrm{L,R}}$ in Theorem 3.1 are only that $A$ is symmetric, that $\bar{A}^{-1}$ exists (as discussed above) and that $D_2$ and $\mathbf{d}_{\mathrm{L,R}}$ in (13) are consistent such that the relations (43) and (44) in "Appendix B" holds. In addition we will assume that $D_2$ is constructed such the left and right boundary closures are equivalent. This implies that $A$ is a centrosymmetric matrix, that is $A_{i,j} = A_{n-i,n-j}$ for all $0 \leq i, j \leq n$, and that $(\mathbf{d}_{\mathrm{L}})_i = -(\mathbf{d}_{\mathrm{R}})_{n-i}$ for $0 \leq i \leq n$. This additional assumption leads to $\xi_{\mathrm{L}} = \xi_{\mathrm{R}}$ (this is easiest seen by expressing the quantities in (25) as $\xi_{\mathrm{L,R}} = 1/\ell + \mathbf{d}_{\mathrm{L,R}}^{\mathsf{T}} G_2 \mathbf{d}_{\mathrm{L,R}}$ and $\xi_{\mathrm{C}} = 1/\ell + \mathbf{d}_{\mathrm{L,R}}^{\mathsf{T}} G_2 \mathbf{d}_{\mathrm{R,L}}$ and thereafter using the fact that the inverse of a centrosymmetric matrix is also centrosymmetric). For later reference we define

$$\xi_{\mathrm{T}} \equiv \xi_{\mathrm{L,R}} + |\xi_{\mathrm{C}}|, \tag{28}$$

and assume that the penalty is chosen to be equally strong on both boundaries:

**Assumption 3.3** Choosing an equal penalty strength on both boundaries corresponds to having $\zeta = 1$ in Corollary 3.2. If in addition equivalent boundary closures are assumed, such that $\xi_{\mathrm{L}} = \xi_{\mathrm{R}}$, we can use $\xi_{\mathrm{T}} \equiv \xi_{\mathrm{L,R}} + |\xi_{\mathrm{C}}|$ from (28). This simplifies the condition of singularity in Corollary 3.2 to $\sigma_{\mathrm{L,R}} = -\xi_{\mathrm{T}}\tau_{\mathrm{L,R}}$.

***Remark 3.4*** The inverse of $\widetilde{A}$ mimics a fundamental solution. For example, the Green's function $\mathcal{G}$ of Poisson's equation, $-u_{xx} = f$ with $u(0) = u(\ell) = 0$, is

$$u(x) = \int_0^{\ell} \mathcal{G}(x, y) f(y)\,\mathrm{d}y, \qquad \mathcal{G}(x, y) = \begin{cases} y(1 - x/\ell), & y < x, \\ x(1 - y/\ell), & x \leq y. \end{cases}$$

Recalling that the matrix $H$ has the role of a quadrature rule, we see the clear similarity to the time-independent, homogeneous version of (14), $\mathbf{v} = \widetilde{A}^{-1} H\mathbf{f}$. The resemblance is more obvious if the penalty dependent part in (22) is ignored, since then $\mathbf{v} = G_2 H\mathbf{f}$. For the second order accurate approximation given in (64), $G_2$ is exact in the grid points, as

$$(G_2)_{i,j} = \begin{cases} x_j(1 - x_i/\ell), & 0 \leq j \leq i \leq n, \\ x_i(1 - x_j/\ell), & 0 \leq i \leq j \leq n. \end{cases}$$

This is identical with the result noted for the classical finite difference method using *injection* instead of SAT, compare [4,28]. With Robin boundary conditions we have

$$u(x) = \int_0^{\ell} \mathcal{G}(x, y) f(y)\,\mathrm{d}y + c_{\mathrm{L}}(1 - x/\ell) + c_{\mathrm{R}}x/\ell$$

where $c_{\mathrm{L,R}}$ depends on the type and data of the boundary conditions from (11), as

$$\begin{bmatrix} c_{\mathrm{L}} \\ c_{\mathrm{R}} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{L}} + \beta_{\mathrm{L}}/\ell & -\beta_{\mathrm{L}}/\ell \\ -\beta_{\mathrm{R}}/\ell & \alpha_{\mathrm{R}} + \beta_{\mathrm{R}}/\ell \end{bmatrix}^{-1} \begin{bmatrix} g_{\mathrm{L}} + \beta_{\mathrm{L}} \int_0^{\ell} (1 - y/\ell))\, f(y)\, \mathrm{d}y \\ g_{\mathrm{R}} + \beta_{\mathrm{R}} \int_0^{\ell} (y/\ell) f(y)\, \mathrm{d}y \end{bmatrix}.$$

The discrete counterpart is still $\mathbf{v} = \widetilde{A}^{-1} H \widetilde{\mathbf{f}}$, which, using relations in Theorem 3.1 and "Section B.1" of Appendix and with $\widetilde{\mathbf{f}} = \mathbf{f} - H^{-1}(\sigma_{\mathrm{L}} \mathbf{e}_{\mathrm{L}} - \tau_{\mathrm{L}} \mathbf{d}_{\mathrm{L}}) g_{\mathrm{L}} - H^{-1}(\sigma_{\mathrm{R}} \mathbf{e}_{\mathrm{R}} + \tau_{\mathrm{R}} \mathbf{d}_{\mathrm{R}}) g_{\mathrm{R}}$, can be written

$$\begin{aligned} \mathbf{v} = {} & G_2 H \mathbf{f} - \tau_{\mathrm{L}} \eta_{\mathrm{L}} \mathbf{b}_{\mathrm{L}} - \tau_{\mathrm{R}} \eta_{\mathrm{R}} \mathbf{b}_{\mathrm{R}} \\ & + \begin{bmatrix} \mathbf{1} - \mathbf{x}/\ell & \mathbf{x}/\ell \end{bmatrix} \begin{bmatrix} \alpha_{\mathrm{L}} + \beta_{\mathrm{L}}/\ell & -\beta_{\mathrm{L}}/\ell \\ -\beta_{\mathrm{R}}/\ell & \alpha_{\mathrm{R}} + \beta_{\mathrm{R}}/\ell \end{bmatrix}^{-1} \begin{bmatrix} g_{\mathrm{L}} + \beta_{\mathrm{L}} (\mathbf{1} - \mathbf{x}/\ell)^{\mathsf{T}} H \mathbf{f} - \delta_{\mathrm{L}} \eta_{\mathrm{L}} \\ g_{\mathrm{R}} + \beta_{\mathrm{R}} (\mathbf{x}/\ell)^{\mathsf{T}} H \mathbf{f} - \delta_{\mathrm{R}} \eta_{\mathrm{R}} \end{bmatrix}, \end{aligned}$$

where

$$\begin{bmatrix} \eta_{\mathrm{L}} \\ \eta_{\mathrm{R}} \end{bmatrix} = \begin{bmatrix} \sigma_{\mathrm{L}} + \tau_{\mathrm{L}} \xi_{\mathrm{L}} & -\tau_{\mathrm{R}} \xi_{\mathrm{C}} \\ -\tau_{\mathrm{L}} \xi_{\mathrm{C}} & \sigma_{\mathrm{R}} + \tau_{\mathrm{R}} \xi_{\mathrm{R}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{b}_{\mathrm{L}}^{\mathsf{T}} \\ \mathbf{b}_{\mathrm{R}}^{\mathsf{T}} \end{bmatrix} H \mathbf{f}.$$

Unless $\mathbf{f} = 0$, such that $\eta_{\mathrm{L,R}} = 0$, the numerical solution $\mathbf{v}$ differs depending on the choice of penalty parameters, where the vectors $\mathbf{1}$, $\mathbf{x}$ and $\mathbf{b}_{\mathrm{L,R}}$ span the possible perturbations. As long as choices resulting in $\sigma_{\mathrm{L,R}} + \xi_{\mathrm{T}} \tau_{\mathrm{L,R}} \approx 0$ are avoided, this perturbation is slight.

### 3.3 Relations Between Stability, Singularity and Dual Consistency

We take a look at the relation between the stability requirements on the scheme (12) and the conditions that make its discretization matrix singular. First, we note that:

**Theorem 3.5** *Consider $\gamma$ in (17) and $\xi_{\mathrm{T}}$ in (28). It holds that $h\gamma = 1/\xi_{\mathrm{T}}$.*

***Proof*** Theorem 3.5 is proven in "Section C.1" of Appendix. □

A consequence of Theorem 3.5 is that the stability demands in (20) can be written

$$\sigma_{\mathrm{L,R}} \alpha_{\mathrm{L,R}} \le 0, \qquad \tau_{\mathrm{L,R}} \beta_{\mathrm{L,R}} \le 1/\xi_{\mathrm{T}}, \qquad \delta_{\mathrm{L,R}}^2 \le -4 \alpha_{\mathrm{L,R}} (\sigma_{\mathrm{L,R}}/\xi_{\mathrm{T}} + \tau_{\mathrm{L,R}}), \tag{29}$$

with $\delta_{\mathrm{L,R}}$ from (19). We will see that the penalty can be chosen such that we have energy stability and a singular discretization matrix at the same time: from Assumption 3.3 we know that the matrix $\widetilde{A}$ is singular when $\sigma_{\mathrm{L,R}} = -\tau_{\mathrm{L,R}} \xi_{\mathrm{T}}$. Inserting this into (29), the third stability demand becomes $\delta_{\mathrm{L,R}}^2 \le 0$, which is only fulfilled if the penalty parameters are chosen in a dual consistent way. This means that if (12) is an energy stable scheme, it must also be dual consistent to risk having a singular discretization matrix. Note though that even if the scheme is dual consistent, a singular discretization matrix is avoided by choosing $\sigma_{\mathrm{L,R}} \neq -\tau_{\mathrm{L,R}} \xi_{\mathrm{T}}$. To be precise, simultaneous having $\sigma_{\mathrm{L,R}} = -\xi_{\mathrm{T}}/(\beta_{\mathrm{L,R}} \xi_{\mathrm{T}} + \alpha_{\mathrm{L,R}})$ and $\tau_{\mathrm{L,R}} = 1/(\beta_{\mathrm{L,R}} \xi_{\mathrm{T}} + \alpha_{\mathrm{L,R}})$ should be avoided, since this particular choice makes $\delta_{\mathrm{L,R}} = 0$, fulfills the stability demands but at the same time makes $\widetilde{A}$ singular.

In Assumption 3.3, one can argue that $\zeta = -1$ gives just as an equal penalty strength as $\zeta = 1$, simplifying Corollary 3.2 to $\sigma_{\mathrm{L,R}} = -\left(\xi_{\mathrm{L,R}} - |\xi_{\mathrm{C}}|\right) \tau_{\mathrm{L,R}}$. However, these choices do not give energy stability and are therefore not interesting for our further discussions. Besides, $|\xi_{\mathrm{C}}|$ tend to be very small so in practice it does not make much of a difference.

### 3.4 Relations to the Stability Demands in [13]

In Sect. 3.1.1 the "borrowing technique" is used for deriving the stability restrictions on the penalty parameters. In [13], a different approach (inspired by [3,18] where wide-stencil discretizations are rewritten as first order systems) is used for showing stability, and here we are going to comment on some connections between the two methods.

In [13], it is assumed that $A$ can be decomposed as in [7], that is as

$$A = A^{\mathsf{T}} = S^{\mathsf{T}}MS, \qquad \mathbf{d}_{\mathrm{L}} = S^{\mathsf{T}}\mathbf{e}_{\mathrm{L}}, \qquad \mathbf{d}_{\mathrm{R}} = S^{\mathsf{T}}\mathbf{e}_{\mathrm{R}}, \tag{30}$$

and the strategy for showing stability is to modify the approximation of $u_x$ from $S\mathbf{v}$ to the auxiliary variable $\mathbf{w} = S\mathbf{v} + M^{-1}\mathbf{e}_{\mathrm{L}}\rho_{\mathrm{L}} + M^{-1}\mathbf{e}_{\mathrm{R}}\rho_{\mathrm{R}}$. In [13], $\rho_{\mathrm{L,R}}$ are penalty-like terms proportional to the solution deviations from boundary data, but other options are possible. Computing $\mathbf{w}^{\mathsf{T}}M\mathbf{w}$ makes the terms

$$2\mathbf{v}^{\mathsf{T}}\mathbf{d}_{\mathrm{L}}\rho_{\mathrm{L}} + 2\mathbf{v}^{\mathsf{T}}\mathbf{d}_{\mathrm{R}}\rho_{\mathrm{R}} + q_{\mathrm{L}}\rho_{\mathrm{L}}^2 + 2q_{\mathrm{C}}\rho_{\mathrm{L}}\rho_{\mathrm{R}} + q_{\mathrm{R}}\rho_{\mathrm{R}}^2 \leq 2\mathbf{v}^{\mathsf{T}}(\mathbf{d}_{\mathrm{L}}\rho_{\mathrm{L}} + \mathbf{d}_{\mathrm{R}}\rho_{\mathrm{R}}) + q_{\mathrm{T}}(\rho_{\mathrm{L}}^2 + \rho_{\mathrm{R}}^2)$$

available to the boundary terms in (16), where $q_{\mathrm{L,R}}$, $q_{\mathrm{C}}$ and $q_{\mathrm{T}}$ are defined as

$$q_{\mathrm{L,R}} \equiv \mathbf{e}_{\mathrm{L,R}}^{\mathsf{T}}M^{-1}\mathbf{e}_{\mathrm{L,R}}, \qquad q_{\mathrm{C}} \equiv \mathbf{e}_{\mathrm{L}}^{\mathsf{T}}M^{-1}\mathbf{e}_{\mathrm{R}} = \mathbf{e}_{\mathrm{R}}^{\mathsf{T}}M^{-1}\mathbf{e}_{\mathrm{L}}, \qquad q_{\mathrm{T}} \equiv q_{\mathrm{L,R}} + |q_{\mathrm{C}}|. \tag{31}$$

The "borrowing technique" on the other hand, makes the terms $-h\gamma\mathbf{v}^{\mathsf{T}}(\mathbf{d}_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} + \mathbf{d}_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}})\mathbf{v}$ available for the boundary terms in (16).

Although these two approaches of showing stability are different, they are closely related. In Lemma 3.6 we formalize this relation and show that $q_{\mathrm{T}} = 1/(h\gamma)$.

**Lemma 3.6** *Assume that $A$ in (13) can be factorized as in (30) with $M > 0$, and define $q_{\mathrm{T}}$ as stated in (31). Next, consider (17), where the parameter $\gamma$ is defined as the maximum number such that $\tilde{A}_{\gamma} \geq 0$ still holds. Then it holds that $h\gamma = 1/q_{\mathrm{T}}$.*

***Proof*** Lemma 3.6 is proven in "Section C.2" of Appendix. □

For wide-stencil operators, $S = D_1$ and $M = H$ in (30), and the parameters $q_{\mathrm{L,R}}$ and $q_{\mathrm{C}}$ in (31) are easily obtained since $M$ is known. For narrow-stencil operators on the other hand, $M$ and the interior of $S$ are not uniquely defined. In [13], the strategy was (under the contrary assumption that $S$ is non-singular and $M$ is singular) to compute

$$\widetilde{q}_{\mathrm{L,R}} \equiv \mathbf{e}_{\mathrm{L,R}}^{\mathsf{T}}\widetilde{M}^{-1}\mathbf{e}_{\mathrm{L,R}}, \qquad \widetilde{q}_{\mathrm{C}} \equiv \mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\widetilde{M}^{-1}\mathbf{e}_{\mathrm{R}} = \mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\widetilde{M}^{-1}\mathbf{e}_{\mathrm{L}}, \qquad \widetilde{q}_{\mathrm{T}} \equiv \widetilde{q}_{\mathrm{L,R}} + |\widetilde{q}_{\mathrm{C}}| \tag{32}$$

instead, where $\widetilde{M} \equiv S^{-\mathsf{T}}(A + p\mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}})S^{-1}$ with $p \neq 0$ being a perturbation parameter. For wide-stencil operators though, it can easily be checked numerically that $q_{\mathrm{L,R}} \neq \widetilde{q}_{\mathrm{L,R}}$ and $q_{\mathrm{C}} \neq \widetilde{q}_{\mathrm{C}}$. This is somewhat alarming, but it can as easily be checked that it still holds that $q_{\mathrm{T}} = \widetilde{q}_{\mathrm{T}}$. We confirm this analytically in Theorem 3.8 below, and the use of $\widetilde{q}_{\mathrm{T}}$ in [13] is thus justified. First though, we note the following:

**Lemma 3.7** *The quantities $\widetilde{q}_{\mathrm{L,R}}$ and $\widetilde{q}_{\mathrm{C}}$ defined in (32) are identical to the quantities $\xi_{\mathrm{L,R}}$ and $\xi_{\mathrm{C}}$ in (25).*

***Proof*** Lemma 3.7 is proven in "Section C.3" of Appendix. □

Thus, in summary, we have that:

**Theorem 3.8** *Assume that $A$ in (13) can be factorized as in (30) with $M > 0$, and define $q_{\mathrm{T}}$ as stated in (31). Next, assume that $M$ is singular instead, with $M \geq 0$, and define $\widetilde{q}_{\mathrm{T}}$ as stated in (32). Then it holds that $q_{\mathrm{T}} = \widetilde{q}_{\mathrm{T}}$.*

**Table 1** The borrowing parameter $\gamma$ computed in [24,32], for narrow-stencil second derivative operators from [23,25]

| Order | $h\widetilde{q}_{\mathrm{T}}$ from [13] | $1/\gamma$ from [24,32] |
|---|---|---|
| (2,0) | 1 | – |
| (2,1) | $\underline{2.5}$ | $\underline{2.5}$ |
| (4,2) | $\underline{3.986}391480987749$ (for $n = 8$) | $\underline{3.986}350339$ |
| (6,3) | $\underline{5.322}804652661742$ (for $n = 12$) | $\underline{5.322}787044$ |
| (8,4) | $\underline{633.6}9326893357$ (for $n = 16$) | $\underline{633.6}2285$ |
| (10,5) | – | 28.4736205 |

In comparison the $\widetilde{q}_{\mathrm{T}}$-values (scaled with $h$) from [13]

***Proof*** From Lemma 3.6 we have that $q_{\mathrm{T}} = 1/(h\gamma)$ and from Theorem 3.5 we have that $1/(h\gamma) = \xi_{\mathrm{T}}$. Combining Lemma 3.7 with the definitions in (32) and (28) we deduce that $\xi_{\mathrm{T}} = \widetilde{q}_{\mathrm{T}}$. All in all, this gives $q_{\mathrm{T}} = 1/(h\gamma) = \xi_{\mathrm{T}} = \widetilde{q}_{\mathrm{T}}$ concluding the proof. □

For an example, see the derived values of $\widetilde{q}_{\mathrm{L,R,C}}$ and $q_{\mathrm{L,R,C}}$ for the wide-stencil (2,0) order operator in "Section D.4" of Appendix. As a numerical confirmation, in Table 1 we compare the values of $h\widetilde{q}_{\mathrm{T}}$ from [13] to the values of $\gamma$ computed in [24,32]. In Table 1 though, it appears that $h\widetilde{q}_{\mathrm{T}} \geq 1/\gamma$. This is because the listed $\gamma$ are computed for $n \to \infty$, and are as such slightly too large for very coarse meshes.

## 4 Conclusions

We discretize the scalar advection equation and the heat equation in one-dimensional space, using the SBP-SAT finite difference method. This gives rise to two semi-discrete schemes of the form $\mathbf{v}_t + L\mathbf{v} = \widetilde{\mathbf{f}}$, where the discretization matrix $L$ is approximating either the first derivative or the second derivative, including treatment of the boundary conditions. The matrix $L$ is, due to properties of the SBP-SAT method, associated with a positive definite matrix $H$ such that $L = H^{-1}K$, where the inverse of $K$ is interpreted as a discrete Green's function. We derive the general forms of these inverses, and provide explicit examples of $K^{-1}$ for some operators $L$ of second and fourth order accuracy.

The boundary treatment SAT induces free parameters in $L$. We first determine these parameters such that the semi-discrete schemes are energy stable. Any remaining degrees of freedom can be used to make the schemes dual consistent. Another important question is whether the discretization matrices $L$ are invertible. Conveniently, the formula for $K^{-1}$ reveals precisely which combinations of SAT parameters that make $L$ singular.

In the second derivative case, it turns out that for one very particular choice of SAT parameters, $L$ can become singular even when the scheme is energy stable. Here, we can avoid this and instead choose the parameters such that the scheme is energy stable, dual consistent and guaranteed to have an invertible discretization matrix (and consequently a unique solution). However, for more complex problems it might not be feasible to *prove* that the discretization matrix is invertible, not even for energy stable schemes.

Last, we take a look at two supposedly different approaches of proving energy stability. Curiously, they are closely related, leading to the same demands on the SAT parameters.

**Acknowledgements** I would like to thank Jan Nordström and Anna Nissen for inspiration and encouragement in early discussions about this work.

**Funding** Open access funding provided by Linnaeus University.

## A Explicit Inverses of the First Derivative Operator

### A.1 The (2,1) Order Accurate Operator

In the second order case, we have

$$D_1 = \frac{1}{h}\begin{bmatrix} -1 & 1 & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & & \\ & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & -1 & 1 \end{bmatrix}, \quad \widetilde{Q} = \begin{bmatrix} -\frac{1}{2} - \sigma_{\mathrm{L}} & \frac{1}{2} & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & & \\ & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{33}$$

with the associated norm-matrix $H = h\,\mathrm{diag}\left(\frac{1}{2},\quad 1,\quad 1,\quad \ldots,\quad 1,\quad 1,\quad \frac{1}{2}\right)$. In (33), we identify $\vec{q}^{\mathsf{T}} = \begin{bmatrix} \frac{1}{2} & 0 & \ldots & 0 & 0 \end{bmatrix}$ and $\overline{Q}$ (given below) according to (8). Using Gauss–Jordan elimination we find the inverse of $\overline{Q}$, as

$$\overline{Q} = \frac{1}{2}\begin{bmatrix} 0 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ & & & & -1 & 1 \end{bmatrix} \quad \Longrightarrow \quad \overline{Q}^{-1} = 2\begin{bmatrix} 1 & -1 & 1 & -1 & 1 & \cdots \\ 1 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 1 & -1 & 1 & \cdots \\ 1 & 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & 1 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

We compute $\vec{q}^{\mathsf{T}}\overline{Q}^{-1} = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & \cdots \end{bmatrix}$ as well. Inserting these results into (9) and (10) yields

$$\widetilde{Q}^{-1} = 2\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -1 & 1 & -1 & 1 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 1 & -1 & 1 & \cdots \\ 0 & 1 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 1 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} - \frac{1}{\sigma_{\mathrm{L}}}\begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{34}$$

The formula (34) holds both for even and odd number of grid points $n$. If $n$ is even (as assumed in the derivation in [14]), the bottom last element of $\widetilde{Q}^{-1}$ is $-1/\sigma_{\mathrm{L}}$, if $n$ is odd, the bottom last element of $\widetilde{Q}^{-1}$ is $2 + 1/\sigma_{\mathrm{L}}$.

### A.2 The (4,2) Order Accurate Operator

In [29], we find $D_1$ with fourth order interior accuracy and the associated $H$. Together with (5) and (7), this gives us

$$\widetilde{Q} = \begin{bmatrix} -\frac{1}{2}-\sigma_{\mathrm{L}} & \frac{59}{96} & -\frac{1}{12} & -\frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{59}{96} & 0 & \frac{59}{96} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{12} & -\frac{59}{96} & 0 & \frac{59}{96} & -\frac{1}{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{32} & 0 & -\frac{59}{96} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{59}{96} & 0 & -\frac{1}{32} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{59}{96} & 0 & \frac{59}{96} & -\frac{1}{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{59}{96} & 0 & \frac{59}{96} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{32} & \frac{1}{12} & -\frac{59}{96} & \frac{1}{2} \end{bmatrix}. \tag{35}$$

We identify $\overline{Q}$ and $\vec{q}$ as indicated in (8). We are now looking for a matrix $\overline{G}$ such that $\overline{QG} = \overline{I}$. Let $\overline{G}$ be composed as

$$\overline{G} = \begin{bmatrix} \vec{g}_1 & \vec{g}_2 & \dots & \vec{g}_n \end{bmatrix}, \vec{g}_j = \begin{bmatrix} g_{1,j} & g_{2,j} & \dots & g_{n,j} \end{bmatrix}^{\mathsf{T}}.$$

For $\overline{QG} = \overline{I}$ to hold, $\overline{Q}\vec{g}_j = \vec{e}_j$ must be fulfilled for all $j = 1, 2, \dots, n$, where the $n \times 1$ vector $\vec{e}_j = [0, \dots, 0, 1, 0, \dots, 0]^{\mathsf{T}}$ is non-zero only in its $j$th element. For $\overline{Q}\vec{g}_j = \vec{e}_j$ to be fulfilled, the interior rows lead to $g_{i-2,j} - 8g_{i-1,j} + 8g_{i+1,j} - g_{i+2,j} = 12\delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. Hence, the fourth order linear homogeneous recurrence relation $g_{i-2,j} - 8g_{i-1,j} + 8g_{i+1,j} - g_{i+2,j} = 0$ has to be fulfilled by most $g_{i,j}$. The general, explicit solution to this recursive relation has the form $g_{i,j} = c_1 + c_2(-1)^i + c_3\phi^i + c_4\phi^{-i}$, where $\phi = 4 + \sqrt{15} \approx 7.873$ and where $c_{1,2,3,4}$ are $j$-dependent constants.

The requirement $\overline{Q}\vec{g}_j = \vec{e}_j$ takes slightly different forms depending on $j$. For $j = 1$, we have $\overline{Q}\vec{g}_1 = \vec{e}_1$, which is expressed explicitly as

$$\frac{1}{96}\begin{bmatrix} 59g_{2,1} \\ -59g_{1,1} + 59g_{3,1} - 8g_{4,1} \\ -59g_{2,1} + 64g_{4,1} - 8g_{5,1} \\ 8\left(g_{2,1} - 8g_{3,1} + 8g_{5,1} - g_{6,1}\right) \\ \vdots \\ 8\left(g_{i-2,1} - 8g_{i-1,1} + 8g_{i+1,1} - g_{i+2,1}\right) \\ \vdots \\ 8\left(g_{n-6,1} - 8g_{n-5,1} + 8g_{n-3,1} - g_{n-2,1}\right) \\ 8g_{n-5,1} - 64g_{n-4,1} + 59g_{n-2,1} - 3g_{n,1} \\ 8g_{n-4,1} - 59g_{n-3,1} + 59g_{n-1,1} - 8g_{n,1} \\ -59g_{n-2,1} + 59g_{n,1} \\ 3g_{n-3,1} + 8g_{n-2,1} - 59g_{n-1,1} + 48g_{n,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{36}$$

The ansatz $g_{i,1} = c_1 + c_2(-1)^i + c_3\phi^i + c_4\phi^{-i}$ holds for $2 \leq i \leq n-2$ where $c_{1,2,3,4}$ are unknowns to be determined. In addition, we have the three unknowns $g_{1,1}$, $g_{n-1,1}$ and $g_{n,1}$. The three first and the four last rows in (36) gives us seven conditions. Inserting the above mentioned expressions for $g_{i,1}$ into (36), gives a linear system with seven unknowns and seven conditions, as

$$\left[\begin{array}{ccccccc|c}
0 & 59 & 59 & 59\phi^2 & 59\phi^{-2} & 0 & 0 & 96 \\
-59 & 51 & -67 & 59\phi^3 - 8\phi^4 & 59\phi^{-3} - 8\phi^{-4} & 0 & 0 & 0 \\
0 & -3 & 13 & -59\phi^2 + 8\phi^3 & -59\phi^{-2} + 8\phi^{-3} & 0 & 0 & 0 \\
0 & 3 & -13(-1)^n & \phi^n(-8\phi^{-3} + 59\phi^{-2}) & \phi^{-n}(-8\phi^3 + 59\phi^2) & 0 & -3 & 0 \\
0 & -51 & 67(-1)^n & \phi^n(8\phi^{-4} - 59\phi^{-3}) & \phi^{-n}(8\phi^4 - 59\phi^3) & 59 & -8 & 0 \\
0 & -59 & -59(-1)^n & -59\phi^{n-2} & -59\phi^{2-n} & 0 & 59 & 0 \\
0 & 11 & 5(-1)^n & \phi^n(3\phi^{-3} + 8\phi^{-2}) & \phi^{-n}(3\phi^3 + 8\phi^2) & -59 & 48 & 0
\end{array}\right]$$

with the unknowns sorted as $g_{1,1}$, $c_1$, $c_2$, $c_3$, $c_4$, $g_{n-1,1}$ and $g_{n,1}$, and where we have used the relation $\phi + \phi^{-1} = 8$ to simplify the expressions.

### A.2.1 The Inverse with an Even Number of Grid Points *n*

To make the expressions manageable, we simplify by assuming that $n$ is an even number. In this particular case, when solving the $7 \times 7$ system above, we obtain

$$g_{1,1} = \frac{1}{2}\left(\frac{12\mathcal{C}_n}{59\mathcal{D}_n}\right)^2, \qquad g_{n-1,1} = \frac{12}{59}\left(\frac{\mathcal{C}_n}{\mathcal{D}_n} - \frac{9}{59\mathcal{D}_n^2}\right), \qquad g_{n,1} = \frac{12\mathcal{C}_n}{59\mathcal{D}_n}.$$

where $\mathcal{C}_n$ and $\mathcal{D}_n$ are integers given in (37) below. Note that $\mathcal{D}_n \geq 1$ for even $n$, so there is no risk of division by zero. Moreover, we obtain

$$c_1 = \frac{12\mathcal{C}_n}{59\mathcal{D}_n}, \qquad c_2 = \frac{36}{590\mathcal{D}_n^2}, \qquad c_3 = \frac{6(\phi - 1)\phi^{1-n}}{590\mathcal{D}_n^2}, \qquad c_4 = \frac{6(\phi^{-1} - 1)\phi^{n-1}}{590\mathcal{D}_n^2},$$

which inserted into the ansatz $g_{i,1} = c_1 + c_2(-1)^i + c_3\phi^i + c_4\phi^{-i}$ leads to

$$g_{i,1} = \frac{12}{59}\left(\frac{\mathcal{C}_n}{\mathcal{D}_n} - \frac{3\mathcal{B}_{n-i}}{\mathcal{D}_n^2}\right), \qquad \text{for } 2 \leq i \leq n-2.$$

The quantities $\mathcal{B}_j$ are integers for integers $j$, and are specified below

$$\begin{aligned}
\mathcal{D}_n &= \frac{\nu_{n/2-1} + \nu_{n/2-2}}{10}, & \mathcal{C}_n &= \frac{9\nu_{n/2-1} + 4\nu_{n/2-2}}{10} \\
\mathcal{B}_j &= \frac{\nu_{j-1} - \nu_{j-2} - 6(-1)^j}{60}, & \mathcal{A}_j &= \frac{\frac{1-(-1)^j}{2}\nu_{n/2-1} + \frac{1+(-1)^j}{2}\nu_{n/2-2} - \nu_{n/2-j}}{60}.
\end{aligned} \tag{37}$$

where $\nu_j = \phi^j + \phi^{-j}$. For convenience, all the $g_{i,1}$ presented above will be restated in (38) and (39), wherein we will also make use of $\mathcal{A}_j$ defined above.

We use the same strategy for the other columns $j > 1$. For $2 \leq j \leq n-2$, we need two different versions of the constants $c_{1,2,3,4}$, depending on if we consider $g_{i,j}$ for $i \leq j$ or for $i \geq j$. We let $g_{i,j} = c_1^u + c_2^u(-1)^i + c_3^u\phi^i + c_4^u\phi^{-i}$ for $2 \leq i \leq j \leq n-2$ and $g_{i,j} = c_1^l + c_2^l(-1)^i + c_3^l\phi^i + c_4^l\phi^{-i}$ for $2 \leq j \leq i \leq n-2$. Thus for every $2 \leq j \leq n-2$, we have eight unknown constants, as well as the three remaining unknowns $g_{1,j}$, $g_{n-1,j}$ and $g_{n,j}$. The three first and the four last rows in the system above gives us seven conditions. From the rows $i = j-1, j, j+1$, we get three more conditions and in addition, we demand that the two versions of $g_{j,j}$ are identical. All in all, this gives a linear system with eleven unknowns $g_{1,j}$, $c_1^u$, $c_2^u$, $c_3^u$, $c_4^u$, $c_1^l$, $c_2^l$, $c_3^l$, $c_4^l$, $g_{n-1,j}$ and $g_{n,j}$ and eleven conditions.

We still consider even numbers of $n$. Solving for the unknowns and inserting $c^u_{1,2,3,4}$ and $c^l_{1,2,3,4}$ into their respective ansatz, we eventually end up with $g_{i,j}$ for the inner columns, presented below in (40) and (41). Furthermore, repeating the procedure for the last two columns, we obtain $g_{i,j}$ for $j = n-1$ and $j = n$, given in (38) and (39). To simplify the expressions in (39)–(41), we have used $\mathcal{A}_j$ in (37).

In summary, when $n$ is even, the inverse of $\overline{Q}$ is given by $\overline{Q}^{-1} = (g_{i,j})_{n\times n}$ with $g_{i,j}$ as described in (38)–(41) below. First, the corner elements are

$$\begin{aligned} g_{1,1} &= \frac{72\mathcal{C}_n^2}{59^2\mathcal{D}_n^2}, & g_{1,n-1} &= \frac{12}{59}\left(\frac{12\mathcal{C}_n^2+9}{59\mathcal{D}_n^2} - \frac{\mathcal{C}_n}{\mathcal{D}_n}\right), & g_{1,n} &= -\frac{12\mathcal{C}_n}{59\mathcal{D}_n}, \\ g_{n-1,1} &= \frac{12}{59}\left(\frac{\mathcal{C}_n}{\mathcal{D}_n} - \frac{9}{59\mathcal{D}_n^2}\right), & g_{n-1,n-1} &= \frac{72\mathcal{C}_n^2}{59^2\mathcal{D}_n^2}, & g_{n-1,n} &= -\frac{12\mathcal{C}_n}{59\mathcal{D}_n}, \\ g_{n,1} &= \frac{12\mathcal{C}_n}{59\mathcal{D}_n}, & g_{n,n-1} &= \frac{12\mathcal{C}_n}{59\mathcal{D}_n}, & g_{n,n} &= 0. \end{aligned} \tag{38}$$

For $2 \le i \le n-2$, we obtain

$$g_{i,1} = \frac{12}{59}\left(\frac{\mathcal{C}_n}{\mathcal{D}_n} - \frac{3\mathcal{B}_{n-i}}{\mathcal{D}_n^2}\right), \quad g_{i,n-1} = 36\frac{4\mathcal{C}_n\mathcal{A}_i + \mathcal{B}_i}{59\mathcal{D}_n^2} - \frac{12\mathcal{A}_i}{\mathcal{D}_n}, \quad g_{i,n} = \frac{-12\mathcal{A}_i}{\mathcal{D}_n}, \tag{39}$$

while we for $2 \le j \le n-2$ have

$$g_{1,j} = 36\frac{4\mathcal{C}_n\mathcal{A}_j + \mathcal{B}_{n-j}}{59\mathcal{D}_n^2} - \frac{12\mathcal{C}_n}{59\mathcal{D}_n}, \quad g_{n-1,j} = \frac{12\mathcal{A}_j}{\mathcal{D}_n} - \frac{36\mathcal{B}_j}{59\mathcal{D}_n^2}, \quad g_{n,j} = \frac{12\mathcal{A}_j}{\mathcal{D}_n}. \tag{40}$$

Finally, the interior elements are

$$\begin{aligned} g_{i,j} &= 12^2\frac{\mathcal{A}_i\mathcal{A}_j}{\mathcal{D}_n^2} - 12\left(\frac{\mathcal{A}_i}{\mathcal{D}_n} - \frac{\mathcal{B}_i\mathcal{B}_{n-j}}{\mathcal{D}_n^2}\right), & &\text{for } 2 \le i \le j \le n-2, \\ g_{i,j} &= 12\left(\frac{\mathcal{A}_j}{\mathcal{D}_n} - \frac{\mathcal{B}_j\mathcal{B}_{n-i}}{\mathcal{D}_n^2}\right), & &\text{for } 2 \le j \le i \le n-2. \end{aligned} \tag{41}$$

In the expressions above we have used $\mathcal{D}_n$, $\mathcal{C}_n$, $\mathcal{B}_j$ and $\mathcal{A}_j$ defined in (37). Next, we recall the structure in (8), and identify $\vec{q}$ in (35) as

$$\vec{q}^{\mathsf{T}} = \begin{bmatrix} \frac{59}{96} & -\frac{1}{12} & -\frac{1}{32} & 0 & 0 & \cdots & 0 \end{bmatrix},$$

and compute $\vec{q}^{\mathsf{T}}\overline{Q}^{-1}$ as $(\vec{q}^{\mathsf{T}}\overline{Q}^{-1})_j = \frac{59}{96}g_{1,j} - \frac{1}{12}g_{2,j} - \frac{1}{32}g_{3,j}$. This gives

$$(\vec{q}^{\mathsf{T}}\overline{Q}^{-1})_j = \begin{cases} \frac{12\mathcal{C}_n}{59\mathcal{D}_n} - 1 & \text{for } j = 1 \\ \frac{12\mathcal{A}_j}{\mathcal{D}_n} - 1 & \text{for } j = 2, \ldots, n-2 \\ \frac{12\mathcal{C}_n}{59\mathcal{D}_n} - 1 & \text{for } j = n-1 \\ -1 & \text{for } j = n, \end{cases} \tag{42}$$

where we have used the structures of $g_{i,j}$ in (38)–(41), together with the following relations:

$$\begin{aligned} \mathcal{B}_2 &= 0, & \mathcal{A}_2 &= 0, & \mathcal{B}_{n-2} &= \frac{\mathcal{C}_n\mathcal{D}_n - 8\mathcal{D}_n^2}{3} \\ \mathcal{B}_3 &= 1, & \mathcal{A}_3 &= \frac{\mathcal{C}_n - 8\mathcal{D}_n}{3}, & \mathcal{B}_{n-3} &= \frac{-2\mathcal{C}_n^2 + 33\mathcal{C}_n\mathcal{D}_n - 136\mathcal{D}_n^2}{3}. \end{aligned}$$

As an example, we write out $\overline{Q}^{-1}$ from (38)–(41) explicitly, for $n = 8$, as

$$\overline{Q}^{-1} = \frac{12}{55^2}\begin{bmatrix} \frac{1291776}{3481} & -\frac{24200}{59} & \frac{19192}{59} & -\frac{19931}{59} & \frac{19027}{59} & -\frac{25520}{59} & \frac{1077881}{3481} & -\frac{25520}{59} \\ \frac{24200}{59} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{25352}{59} & 0 & 384 & -337 & 329 & -440 & \frac{18587}{59} & -440 \\ \frac{25499}{59} & 0 & 433 & 6 & 48 & -55 & \frac{2344}{59} & -55 \\ \frac{25517}{59} & 0 & 439 & 48 & 384 & -440 & \frac{18752}{59} & -440 \\ \frac{25520}{59} & 0 & 440 & 55 & 440 & 0 & \frac{1320}{59} & 0 \\ \frac{1505671}{3481} & 0 & \frac{25957}{59} & \frac{3224}{59} & \frac{25792}{59} & -\frac{1320}{59} & \frac{1291776}{3481} & -\frac{25520}{59} \\ \frac{25520}{59} & 0 & 440 & 55 & 440 & 0 & \frac{25520}{59} & 0 \end{bmatrix},$$

where e.g. $\mathcal{D}_8 = 55$. Correspondingly, we have

$$\vec{q}^{\mathsf{T}}\overline{Q}^{-1} = \frac{1}{55}\left[\, \tfrac{2323}{59} \; -55 \; 41 \; -43 \; 41 \; -55 \; \tfrac{2323}{59} \; -55 \,\right].$$

Inserting $\overline{Q}^{-1}$ from (38)–(41), and (42) into (9) and (10) yields the inverse of $\widetilde{Q}$ in the (4,2) order accurate case (for $n$ even). In the example with $n = 8$, we have

$$\widetilde{Q}^{-1} \approx \begin{bmatrix} 1 & -0.72 & 1 & -0.75 & 0.78 & -0.75 & 1 & -0.72 & 1 \\ 1 & 0.76 & -0.63 & 0.54 & -0.56 & 0.53 & -0.72 & 0.51 & -0.72 \\ 1 & 0.91 & 1 & -0.75 & 0.78 & -0.75 & 1 & -0.72 & 1 \\ 1 & 0.99 & 1 & 0.78 & -0.56 & 0.56 & -0.75 & 0.53 & -0.75 \\ 1 & 1.00 & 1 & 0.97 & 0.81 & -0.56 & 0.78 & -0.56 & 0.78 \\ 1 & 1.00 & 1 & 1.00 & 0.97 & 0.78 & -0.75 & 0.54 & -0.75 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -0.63 & 1 \\ 1 & 1.00 & 1 & 1.00 & 1.00 & 0.99 & 0.91 & 0.76 & -0.72 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

for $\sigma_{\mathrm{L}} = -1$, which we see mimic the Green's function, as discussed in Remark 2.3.

Recall that we assumed that $n$ was even. Repeating the derivation for odd $n$, the resulting inverse $\widetilde{Q}^{-1}$ has a similar behaviour, but with other coefficients. For example, the denominators will instead be $\widetilde{\mathcal{D}}_n = (\phi^{(n-3)/2} - \phi^{(3-n)/2})/\sqrt{60}$, which are positive integers for odd $n \geq 5$.

## B Proof of Theorem 3.1

Theorem 3.1 states that the inverse of $\widetilde{A}$ from (15) is equal to the expression (22). This is shown in "Section B.2" of Appendix, however, first, we present some useful relations.

### B.1 Preliminaries

Note that $D_2\mathbf{1} = \mathbf{0}$ and $D_2\mathbf{x} = \mathbf{0}$, since $D_2$ approximates the second derivative operator (these two relations actually hold also for the inconsistent (2,0) order accurate operators in "Sections D.1" and "D.4" of Appendices). Furthermore, $\mathbf{d}_{\mathrm{L,R}}^{\mathsf{T}}$ consistently approximate the first derivative, so that $\mathbf{d}_{\mathrm{L,R}}^{\mathsf{T}}\mathbf{1} = 0$ and $\mathbf{d}_{\mathrm{L,R}}^{\mathsf{T}}\mathbf{x} = 1$. Hence

$$\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}(\ell\mathbf{1} - \mathbf{x}) = -1, \qquad \mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\mathbf{x} = 1, \qquad \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}(\ell\mathbf{1} - \mathbf{x}) = -1, \qquad \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\mathbf{x} = 1. \tag{43}$$

Combining the above relations with $A = -HD_2 + \mathbf{e}_\mathrm{R}\mathbf{d}_\mathrm{R}^\mathsf{T} - \mathbf{e}_\mathrm{L}\mathbf{d}_\mathrm{L}^\mathsf{T}$ from (13), gives

$$A(\ell\mathbf{1} - \mathbf{x}) = \mathbf{e}_\mathrm{L} - \mathbf{e}_\mathrm{R}, \qquad A\mathbf{x} = \mathbf{e}_\mathrm{R} - \mathbf{e}_\mathrm{L}. \tag{44}$$

Now, we define the additional $(n-1)\times 1$-vectors $\vec{1} = [1\ 1\ \ldots\ 1]^\mathsf{T}$ and $\vec{x} = h[1\ 2\ \ldots\ n-1]^\mathsf{T}$ (they are shorter versions of $\mathbf{1}$ and $\mathbf{x}$ in Theorem 3.1). With these new variables and with the notation from (21), the relations (44) can be expressed as

$$\begin{bmatrix} \ell a_\mathrm{L} + \vec{a}_\mathrm{L}^\mathsf{T}(\ell\vec{1} - \vec{x}) \\ \ell\vec{a}_\mathrm{L} + \bar{A}(\ell\vec{1} - \vec{x}) \\ \ell a_\mathrm{C} + \vec{a}_\mathrm{R}^\mathsf{T}(\ell\vec{1} - \vec{x}) \end{bmatrix} = \begin{bmatrix} 1 \\ \vec{0} \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} \vec{a}_\mathrm{L}^\mathsf{T}\vec{x} + \ell a_\mathrm{C} \\ \bar{A}\vec{x} + \ell\vec{a}_\mathrm{R} \\ \vec{a}_\mathrm{R}^\mathsf{T}\vec{x} + \ell a_\mathrm{R} \end{bmatrix} = \begin{bmatrix} -1 \\ \vec{0} \\ 1 \end{bmatrix}.$$

Given that $A$ is correctly constructed, such that $\bar{A}$ in invertible, this leads to the relations

$$\vec{1} - \vec{x}/\ell = -\bar{A}^{-1}\vec{a}_\mathrm{L}, \qquad \vec{x}/\ell = -\bar{A}^{-1}\vec{a}_\mathrm{R} \tag{45}$$

and

$$a_\mathrm{L} = \vec{a}_\mathrm{L}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{L} + \frac{1}{\ell}, \quad a_\mathrm{R} = \vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{R} + \frac{1}{\ell}, \quad a_\mathrm{C} = \vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{L} - \frac{1}{\ell} = \vec{a}_\mathrm{L}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{R} - \frac{1}{\ell}. \tag{46}$$

Now, multiplying $A$ from (21) by $G_2$ from (23) and using the relations (45), we get

$$AG_2 = \begin{bmatrix} 0 & \vec{a}_\mathrm{L}^\mathsf{T}\bar{A}^{-1} & 0 \\ \vec{0} & \bar{I} & \vec{0} \\ 0 & \vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1} & 0 \end{bmatrix} = I - \mathbf{e}_\mathrm{L}(\mathbf{1} - \mathbf{x}/\ell)^\mathsf{T} - \mathbf{e}_\mathrm{R}\mathbf{x}^\mathsf{T}/\ell \tag{47}$$

where $\bar{I}$ is the $(n-1)\times(n-1)$ identity matrix. From (23) we have $\mathbf{b}_\mathrm{L} = \mathbf{1} - \mathbf{x}/\ell - G_2\mathbf{d}_\mathrm{L}$ and $\mathbf{b}_\mathrm{R} = \mathbf{x}/\ell + G_2\mathbf{d}_\mathrm{R}$, and using the relations (44), (47) and (43), we arrive at

$$A\mathbf{b}_\mathrm{L} = -\mathbf{d}_\mathrm{L}, \qquad A\mathbf{b}_\mathrm{R} = \mathbf{d}_\mathrm{R}. \tag{48}$$

The vectors $\mathbf{e}_\mathrm{L,R}$ picks out the first and last elements in the vectors they are multiplied by, such that

$$\begin{aligned} \mathbf{e}_\mathrm{L}^\mathsf{T}(\mathbf{1} - \mathbf{x}/\ell) = 1, \quad \mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{x}/\ell = 0, \quad \mathbf{e}_\mathrm{R}^\mathsf{T}(\mathbf{1} - \mathbf{x}/\ell) = 0, \quad \mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{x}/\ell = 1, \\ \mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{b}_\mathrm{L} = 1, \quad \mathbf{e}_\mathrm{L}^\mathsf{T}\mathbf{b}_\mathrm{R} = 0, \quad \mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{b}_\mathrm{L} = 0, \quad \mathbf{e}_\mathrm{R}^\mathsf{T}\mathbf{b}_\mathrm{R} = 1. \end{aligned} \tag{49}$$

Finally, from (23) we have

$$\mathbf{e}_\mathrm{L}^\mathsf{T}G_2 = \mathbf{e}_\mathrm{R}^\mathsf{T}G_2 = \mathbf{0}^\mathsf{T}, \qquad \mathbf{d}_\mathrm{L}^\mathsf{T}G_2 = (\mathbf{1} - \mathbf{x}/\ell - \mathbf{b}_\mathrm{L})^\mathsf{T}, \qquad \mathbf{d}_\mathrm{R}^\mathsf{T}G_2 = (\mathbf{b}_\mathrm{R} - \mathbf{x}/\ell)^\mathsf{T}. \tag{50}$$

We are now ready to prove Theorem 3.1.

### B.2 Confirmation of Eq. (22) with (23)–(25)

We multiply $\widetilde{A}$ in (15) by the expression for $\widetilde{A}^{-1}$ in (22), with the aim of showing that $\widetilde{A}\widetilde{A}^{-1} = I$ indeed holds. In the first step, (22) yields

$$\widetilde{A}\widetilde{A}^{-1} = \widetilde{A}G_2 + \underbrace{\widetilde{A}\begin{bmatrix} -\tau_\mathrm{L}\mathbf{b}_\mathrm{L}^\mathsf{T} \\ -\tau_\mathrm{R}\mathbf{b}_\mathrm{R}^\mathsf{T} \\ (\mathbf{1} - \mathbf{x}/\ell)^\mathsf{T} \\ \mathbf{x}^\mathsf{T}/\ell \end{bmatrix}^\mathsf{T}}_{\Gamma} \Sigma^{-1} \begin{bmatrix} \mathbf{b}_\mathrm{L}^\mathsf{T} \\ \mathbf{b}_\mathrm{R}^\mathsf{T} \\ \beta_\mathrm{L}(\mathbf{1} - \mathbf{x}/\ell)^\mathsf{T} \\ \beta_\mathrm{R}\mathbf{x}^\mathsf{T}/\ell \end{bmatrix}. \tag{51}$$

We start by looking at the first term in (51). First using (15), followed by the relations in (47) and (50), and thereafter just rearranging the terms, we arrive at

$$\begin{aligned}
\widetilde{A}G_2 &= AG_2 \begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{L}}\alpha_{\mathrm{L}} & \sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1\\ \tau_{\mathrm{L}}\alpha_{\mathrm{L}} & \tau_{\mathrm{L}}\beta_{\mathrm{L}}\end{bmatrix}\begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix} G_2 \\
&\quad - \begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{R}}\alpha_{\mathrm{R}} & \sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1\\ \tau_{\mathrm{R}}\alpha_{\mathrm{R}} & \tau_{\mathrm{R}}\beta_{\mathrm{R}}\end{bmatrix}\begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix} G_2 \\
&= I - \mathbf{e}_{\mathrm{L}}(\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}} - \mathbf{e}_{\mathrm{R}}\mathbf{x}^{\mathsf{T}}/\ell - \begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1\\ \tau_{\mathrm{L}}\beta_{\mathrm{L}}\end{bmatrix}(\mathbf{b}_{\mathrm{L}}-\mathbf{1}+\mathbf{x}/\ell)^{\mathsf{T}} \\
&\quad - \begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1\\ \tau_{\mathrm{R}}\beta_{\mathrm{R}}\end{bmatrix}(\mathbf{b}_{\mathrm{R}}-\mathbf{x}/\ell)^{\mathsf{T}} \\
&= I - \begin{bmatrix}(\sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1)\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} - \tau_{\mathrm{L}}\beta_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\\ (\sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1)\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} + \tau_{\mathrm{R}}\beta_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\\ -\left(\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} - \tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\right)\\ -(\sigma_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} + \tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}})\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\mathbf{b}_{\mathrm{L}}^{\mathsf{T}}\\ \mathbf{b}_{\mathrm{R}}^{\mathsf{T}}\\ \beta_{\mathrm{L}}(\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}}\\ \beta_{\mathrm{R}}\mathbf{x}^{\mathsf{T}}/\ell\end{bmatrix}.
\end{aligned} \tag{52}$$

Next, we look at the part $\Gamma$ in (51). After rewriting $\widetilde{A}$ using (15), we use the relations in (48), (44), (49), (43) and (25). Thereafter, the resulting terms are rearranged. These steps are shown below in (53).

$$\begin{aligned}
\Gamma = A &\begin{bmatrix}-\tau_{\mathrm{L}}\mathbf{b}_{\mathrm{L}}^{\mathsf{T}}\\ -\tau_{\mathrm{R}}\mathbf{b}_{\mathrm{R}}^{\mathsf{T}}\\ (\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}}\\ \mathbf{x}^{\mathsf{T}}/\ell\end{bmatrix}^{\mathsf{T}} - \begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{L}}\alpha_{\mathrm{L}} & \sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1\\ \tau_{\mathrm{L}}\alpha_{\mathrm{L}} & \tau_{\mathrm{L}}\beta_{\mathrm{L}}\end{bmatrix}\begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}-\tau_{\mathrm{L}}\mathbf{b}_{\mathrm{L}}^{\mathsf{T}}\\ -\tau_{\mathrm{R}}\mathbf{b}_{\mathrm{R}}^{\mathsf{T}}\\ (\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}}\\ \mathbf{x}^{\mathsf{T}}/\ell\end{bmatrix}^{\mathsf{T}} \\
&- \begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{R}}\alpha_{\mathrm{R}} & \sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1\\ \tau_{\mathrm{R}}\alpha_{\mathrm{R}} & \tau_{\mathrm{R}}\beta_{\mathrm{R}}\end{bmatrix}\begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}-\tau_{\mathrm{L}}\mathbf{b}_{\mathrm{L}}^{\mathsf{T}}\\ -\tau_{\mathrm{R}}\mathbf{b}_{\mathrm{R}}^{\mathsf{T}}\\ (\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}}\\ \mathbf{x}^{\mathsf{T}}/\ell\end{bmatrix}^{\mathsf{T}} \\
= &\begin{bmatrix}\tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\\ -\tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\\ (\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}-\mathbf{e}_{\mathrm{R}}^{\mathsf{T}})/\ell\\ (\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}-\mathbf{e}_{\mathrm{L}}^{\mathsf{T}})/\ell\end{bmatrix}^{\mathsf{T}} - \begin{bmatrix}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}}\\ -\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{L}}\alpha_{\mathrm{L}} & \sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1\\ \tau_{\mathrm{L}}\alpha_{\mathrm{L}} & \tau_{\mathrm{L}}\beta_{\mathrm{L}}\end{bmatrix}\begin{bmatrix}-\tau_{\mathrm{L}} & 0 & 1 & 0\\ -\tau_{\mathrm{L}}\widetilde{q}_{\mathrm{L}} & \tau_{\mathrm{R}}\widetilde{q}_{\mathrm{C}} & 1/\ell & -1/\ell\end{bmatrix} \\
&- \begin{bmatrix}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}}\\ \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{R}}\alpha_{\mathrm{R}} & \sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1\\ \tau_{\mathrm{R}}\alpha_{\mathrm{R}} & \tau_{\mathrm{R}}\beta_{\mathrm{R}}\end{bmatrix}\begin{bmatrix}0 & -\tau_{\mathrm{R}} & 0 & 1\\ \tau_{\mathrm{L}}\widetilde{q}_{\mathrm{C}} & -\tau_{\mathrm{R}}\widetilde{q}_{\mathrm{R}} & -1/\ell & 1/\ell\end{bmatrix} \\
= &\begin{bmatrix}(\sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1)\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} - \tau_{\mathrm{L}}\beta_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\\ (\sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1)\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} + \tau_{\mathrm{R}}\beta_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\\ -\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} + \tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}}\\ -\sigma_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} - \tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\sigma_{\mathrm{L}}+\tau_{\mathrm{L}}\widetilde{q}_{\mathrm{L}} & -\tau_{\mathrm{R}}\widetilde{q}_{\mathrm{C}} & 0 & 0\\ -\tau_{\mathrm{L}}\widetilde{q}_{\mathrm{C}} & \sigma_{\mathrm{R}}+\tau_{\mathrm{R}}\widetilde{q}_{\mathrm{R}} & 0 & 0\\ \delta_{\mathrm{L}} & 0 & \alpha_{\mathrm{L}}+\frac{\beta_{\mathrm{L}}}{\ell} & -\frac{\beta_{\mathrm{L}}}{\ell}\\ 0 & \delta_{\mathrm{R}} & -\frac{\beta_{\mathrm{R}}}{\ell} & \alpha_{\mathrm{R}}+\frac{\beta_{\mathrm{R}}}{\ell}\end{bmatrix},
\end{aligned} \tag{53}$$

We note that the last 4 × 4-matrix is nothing but $\Sigma$ from (24). Inserting the results from (52) and (53) into (51) gives us

$$\widetilde{A}\widetilde{A}^{-1} = I - \begin{bmatrix} (\sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1)\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} - \tau_{\mathrm{L}}\beta_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \\ (\sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1)\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} + \tau_{\mathrm{R}}\beta_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \\ -\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} + \tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \\ -\sigma_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} - \tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{b}_{\mathrm{L}}^{\mathsf{T}} \\ \mathbf{b}_{\mathrm{R}}^{\mathsf{T}} \\ \beta_{\mathrm{L}}(\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}} \\ \beta_{\mathrm{R}}\mathbf{x}^{\mathsf{T}}/\ell \end{bmatrix}$$

$$+ \begin{bmatrix} (\sigma_{\mathrm{L}}\beta_{\mathrm{L}}+1)\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} - \tau_{\mathrm{L}}\beta_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \\ (\sigma_{\mathrm{R}}\beta_{\mathrm{R}}+1)\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} + \tau_{\mathrm{R}}\beta_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \\ -\sigma_{\mathrm{L}}\mathbf{e}_{\mathrm{L}}^{\mathsf{T}} + \tau_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} \\ -\sigma_{\mathrm{R}}\mathbf{e}_{\mathrm{R}}^{\mathsf{T}} - \tau_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \Sigma\Sigma^{-1} \begin{bmatrix} \mathbf{b}_{\mathrm{L}}^{\mathsf{T}} \\ \mathbf{b}_{\mathrm{R}}^{\mathsf{T}} \\ \beta_{\mathrm{L}}(\mathbf{1}-\mathbf{x}/\ell)^{\mathsf{T}} \\ \beta_{\mathrm{R}}\mathbf{x}^{\mathsf{T}}/\ell \end{bmatrix} = I,$$

concluding the proof.

## C Proofs of the Relations Between $\xi_{\mathrm{T}}$, $\gamma$, $q_{\mathrm{T}}$ and $\widetilde{q}_{\mathrm{T}}$

Below we present the proofs of Theorem 3.5 and the Lemmas 3.6 and 3.7.

### C.1 Proof of Theorem 3.5

We aim to relate $\gamma$ in (17) to $\xi_{\mathrm{T}}$ in (28). Note that the latter quantity relies on that $\xi_{\mathrm{L}} = \xi_{\mathrm{R}}$ in (25). To emphasize this, we introduce $\xi_{\mathrm{D}} = \xi_{\mathrm{L,R}}$.

We start by defining $\widetilde{\mathbf{v}} = \mathbf{v} - \mathbf{b}_{\mathrm{L}}\rho_{\mathrm{L}} + \mathbf{b}_{\mathrm{R}}\rho_{\mathrm{R}}$ with $\mathbf{b}_{\mathrm{L,R}}$ from (23), and compute

$$\widetilde{\mathbf{v}}^{\mathsf{T}}A\widetilde{\mathbf{v}} = \mathbf{v}^{\mathsf{T}}A\mathbf{v} + 2\rho_{\mathrm{L}}\mathbf{v}^{\mathsf{T}}\mathbf{d}_{\mathrm{L}} + 2\rho_{\mathrm{R}}\mathbf{v}^{\mathsf{T}}\mathbf{d}_{\mathrm{R}} + \rho_{\mathrm{L}}^2\xi_{\mathrm{L}} + 2\rho_{\mathrm{L}}\rho_{\mathrm{R}}\xi_{\mathrm{C}} + \rho_{\mathrm{R}}^2\xi_{\mathrm{R}} \tag{54}$$

using (48) and (25). The $(n+1)\times 1$-vector $\mathbf{v}$ is arbitrary and for the scalars $\rho_{\mathrm{L,R}}$ we make the ansatz $\rho_{\mathrm{L}} = (s_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} + t_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}})\mathbf{v}$ and $\rho_{\mathrm{R}} = (s_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} + t_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}})\mathbf{v}$ where $t_{\mathrm{L,R}}$ and $s_{\mathrm{L,R}}$ are scalars yet to be determined. Inserted into (54), this yields

$$\widetilde{\mathbf{v}}^{\mathsf{T}}A\widetilde{\mathbf{v}} = \mathbf{v}^{\mathsf{T}}A\mathbf{v} + \mathbf{v}^{\mathsf{T}}(z_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}\mathbf{d}_{\mathrm{L}}^{\mathsf{T}} + 2z_{\mathrm{C}}\mathbf{d}_{\mathrm{L}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}} + z_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}\mathbf{d}_{\mathrm{R}}^{\mathsf{T}})\mathbf{v} \tag{55}$$

where we have defined

$$\begin{aligned} z_{\mathrm{L}} &= 2s_{\mathrm{L}} + 2\xi_{\mathrm{C}}s_{\mathrm{L}}t_{\mathrm{L}} + \xi_{\mathrm{L}}s_{\mathrm{L}}^2 + \xi_{\mathrm{R}}t_{\mathrm{L}}^2 \\ z_{\mathrm{R}} &= 2s_{\mathrm{R}} + 2\xi_{\mathrm{C}}s_{\mathrm{R}}t_{\mathrm{R}} + \xi_{\mathrm{R}}s_{\mathrm{R}}^2 + \xi_{\mathrm{L}}t_{\mathrm{R}}^2 \\ z_{\mathrm{C}} &= t_{\mathrm{L}} + t_{\mathrm{R}} + \xi_{\mathrm{L}}s_{\mathrm{L}}t_{\mathrm{R}} + \xi_{\mathrm{R}}s_{\mathrm{R}}t_{\mathrm{L}} + \xi_{\mathrm{C}}s_{\mathrm{L}}s_{\mathrm{R}} + \xi_{\mathrm{C}}t_{\mathrm{L}}t_{\mathrm{R}}. \end{aligned} \tag{56}$$

Using the "borrowing technique", $\gamma$ is the maximum value such that $\tilde{A}_{\gamma} \geq 0$ still holds, referring to $\gamma$ and $\tilde{A}_{\gamma}$ from (17). For (55) to correspond to (17), we need $z_{\mathrm{L}} = z_{\mathrm{R}}$ and $z_{\mathrm{C}} = 0$, and under these constraints we must mimimize $z_{\mathrm{L,R}}$. To get there, we first define $x_{\mathrm{L}} = s_{\mathrm{L}} + t_{\mathrm{L}}$, $y_{\mathrm{L}} = s_{\mathrm{L}} - t_{\mathrm{L}}$, $x_{\mathrm{R}} = s_{\mathrm{R}} + t_{\mathrm{R}}$ and $y_{\mathrm{R}} = s_{\mathrm{R}} - t_{\mathrm{R}}$. Now

$$\begin{aligned} x_{\mathrm{L}} + y_{\mathrm{L}} &= 2s_{\mathrm{L}}, & x_{\mathrm{L}}^2 - y_{\mathrm{L}}^2 &= 4s_{\mathrm{L}}t_{\mathrm{L}}, & x_{\mathrm{L}}^2 + y_{\mathrm{L}}^2 &= 2(s_{\mathrm{L}}^2 + t_{\mathrm{L}}^2), \\ x_{\mathrm{R}} + y_{\mathrm{R}} &= 2s_{\mathrm{R}}, & x_{\mathrm{R}}^2 - y_{\mathrm{R}}^2 &= 4s_{\mathrm{R}}t_{\mathrm{R}}, & x_{\mathrm{R}}^2 + y_{\mathrm{R}}^2 &= 2(s_{\mathrm{R}}^2 + t_{\mathrm{R}}^2). \end{aligned}$$

Inserted into $z_{\mathrm{L}}$ and $z_{\mathrm{R}}$ in (56), these relations gives us

$$\begin{aligned} z_{\mathrm{L,R}} &= x_{\mathrm{L,R}} + y_{\mathrm{L,R}} + \xi_{\mathrm{C}} \frac{x_{\mathrm{L,R}}^2 - y_{\mathrm{L,R}}^2}{2} + \xi_{\mathrm{D}} \frac{x_{\mathrm{L,R}}^2 + y_{\mathrm{L,R}}^2}{2} \\ &= \frac{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}}{2} \left( x_{\mathrm{L,R}} + \frac{1}{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}} \right)^2 + \frac{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}}{2} \left( y_{\mathrm{L,R}} + \frac{1}{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}} \right)^2 - \frac{\xi_{\mathrm{D}}}{\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2} \end{aligned}$$

where we have used that $\xi_{\mathrm{D}} = \xi_{\mathrm{L}} = \xi_{\mathrm{R}}$. Note that for fixed values of $z_{\mathrm{L}}$ and $z_{\mathrm{R}}$, the pairs $(x_{\mathrm{L}}, y_{\mathrm{L}})$ and $(x_{\mathrm{R}}, y_{\mathrm{R}})$ describe ellipses. Reformulated in a parametric form, they are

$$\begin{aligned} x_{\mathrm{L}} &= \frac{-1}{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}} + \sqrt{\frac{2}{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}}}\, r_{\mathrm{L}} \cos(\theta_{\mathrm{L}}), \quad y_{\mathrm{L}} = \frac{-1}{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}} + \sqrt{\frac{2}{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}}}\, r_{\mathrm{L}} \sin(\theta_{\mathrm{L}}), \\ x_{\mathrm{R}} &= \frac{-1}{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}} + \sqrt{\frac{2}{\xi_{\mathrm{D}} + \xi_{\mathrm{C}}}}\, r_{\mathrm{R}} \cos(\theta_{\mathrm{R}}), \quad y_{\mathrm{R}} = \frac{-1}{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}} + \sqrt{\frac{2}{\xi_{\mathrm{D}} - \xi_{\mathrm{C}}}}\, r_{\mathrm{R}} \sin(\theta_{\mathrm{R}}), \end{aligned} \tag{57}$$

where $r_{\mathrm{L}}^2 = z_{\mathrm{L}} + \xi_{\mathrm{D}}/(\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2)$ and $r_{\mathrm{R}}^2 = z_{\mathrm{R}} + \xi_{\mathrm{D}}/(\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2)$. To enforce $z_{\mathrm{L}} = z_{\mathrm{R}}$, we simply let $r_{\mathrm{L}} = r_{\mathrm{R}} = r$. This gives us

$$z_{\mathrm{L,R}} = r^2 - \frac{\xi_{\mathrm{D}}}{\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2}. \tag{58}$$

Next, we need to fulfull the requirement $z_{\mathrm{C}} = 0$. Inserting the relations

$$t_{\mathrm{L,R}} = \frac{x_{\mathrm{L,R}} - y_{\mathrm{L,R}}}{2}, \quad s_{\mathrm{L}} t_{\mathrm{R}} + t_{\mathrm{L}} s_{\mathrm{R}} = \frac{x_{\mathrm{L}} x_{\mathrm{R}} - y_{\mathrm{L}} y_{\mathrm{R}}}{2}, \quad s_{\mathrm{L}} s_{\mathrm{R}} + t_{\mathrm{L}} t_{\mathrm{R}} = \frac{x_{\mathrm{L}} x_{\mathrm{R}} + y_{\mathrm{L}} y_{\mathrm{R}}}{2}$$

into $z_{\mathrm{C}}$ in (56), and thereafter using (57) with $r_{\mathrm{L,R}} = r$, leads to

$$\begin{aligned} 2 z_{\mathrm{C}} &= x_{\mathrm{L}} - y_{\mathrm{L}} + x_{\mathrm{R}} - y_{\mathrm{R}} + \xi_{\mathrm{D}} (x_{\mathrm{L}} x_{\mathrm{R}} - y_{\mathrm{L}} y_{\mathrm{R}}) + \xi_{\mathrm{C}} (x_{\mathrm{L}} x_{\mathrm{R}} + y_{\mathrm{L}} y_{\mathrm{R}}) \\ &= 2 \left( \frac{\xi_{\mathrm{C}}}{\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2} + r^2 \cos(\theta_{\mathrm{L}} + \theta_{\mathrm{R}}) \right). \end{aligned}$$

Now, we want $z_{\mathrm{C}} = 0$ while keeping $r^2$ to a minimum (in order to in turn minimize $z_{\mathrm{L,R}}$). We achieve this by putting

$$r^2 = \frac{|\xi_{\mathrm{C}}|}{\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2}, \qquad \cos(\theta_{\mathrm{L}} + \theta_{\mathrm{R}}) = -\mathrm{sgn}(\xi_{\mathrm{C}}).$$

It can be shown that $\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2 \geq 0$ (by inserting (48) into (25) and using that $A^{\mathsf{T}} = A \geq 0$), therefore the absolute value is only needed for $\xi_{\mathrm{C}}$. Inserting the above choice of $r^2$ into $z_{\mathrm{L,R}}$ in (58) and thereafter using (28) with $\xi_{\mathrm{L,R}} = \xi_{\mathrm{D}}$, we obtain

$$z_{\mathrm{L,R}} = \frac{|\xi_{\mathrm{C}}| - \xi_{\mathrm{D}}}{\xi_{\mathrm{D}}^2 - \xi_{\mathrm{C}}^2} = \frac{-1}{\xi_{\mathrm{D}} + |\xi_{\mathrm{C}}|} = -\frac{1}{\xi_{\mathrm{T}}}.$$

We have thereby shown that, with $z_{\mathrm{C}} = 0$ and $z_{\mathrm{L}} = z_{\mathrm{R}}$ in (55), $1/\xi_{\mathrm{T}}$ is the maximum amount of "positivity" in form of $(\mathbf{d}_{\mathrm{L}} \mathbf{d}_{\mathrm{L}}^{\mathsf{T}} + \mathbf{d}_{\mathrm{R}} \mathbf{d}_{\mathrm{R}}^{\mathsf{T}})$ we can extract from $A$. Inserting $z_{\mathrm{C}} = 0$ and $z_{\mathrm{L,R}} = -1/\xi_{\mathrm{T}}$ into (55) and noting that $\widetilde{\mathbf{v}}^{\mathsf{T}} A \widetilde{\mathbf{v}} \geq 0$, we get

$$\mathbf{v}^{\mathsf{T}} A \mathbf{v} - \frac{1}{\xi_{\mathrm{T}}} \mathbf{v}^{\mathsf{T}} (\mathbf{d}_{\mathrm{L}} \mathbf{d}_{\mathrm{L}}^{\mathsf{T}} + \mathbf{d}_{\mathrm{R}} \mathbf{d}_{\mathrm{R}}^{\mathsf{T}}) \mathbf{v} \geq 0. \tag{59}$$

Comparing with (17), we deduce that $h\gamma = 1/\xi_{\mathrm{T}}$.

### C.2 Proof of Lemma 3.6

We define $\mathbf{w} = S\mathbf{v} + M^{-1}\mathbf{e}_\mathrm{L}\rho_\mathrm{L} + M^{-1}\mathbf{e}_\mathrm{R}\rho_\mathrm{R}$ and use the relations in (30) to compute

$$\mathbf{w}^\mathsf{T} M\mathbf{w} = \mathbf{v}^\mathsf{T} A\mathbf{v} + 2\rho_\mathrm{L}\mathbf{v}^\mathsf{T}\mathbf{d}_\mathrm{L} + 2\rho_\mathrm{R}\mathbf{v}^\mathsf{T}\mathbf{d}_\mathrm{R} + \rho_\mathrm{L}^2 q_\mathrm{L} + 2\rho_\mathrm{L}\rho_\mathrm{R} q_\mathrm{C} + \rho_\mathrm{R}^2 q_\mathrm{R} \tag{60}$$

where $q_\mathrm{L,R,C}$ are defined in (31) and where $\rho_\mathrm{L,R}$ are any scalars. It is assumed that $M > 0$ and that $\mathbf{w}^\mathsf{T} M\mathbf{w} \geq 0$. Note that the right-hand-side of (60) has the same form as (54), but with $\xi_\mathrm{L,R,C}$ replaced by $q_\mathrm{L,R,C}$. Thus, by following the same procedure, we obtain the relation corresponding to (59), namely

$$\mathbf{v}^\mathsf{T} A\mathbf{v} - \frac{1}{q_\mathrm{T}}\mathbf{v}^\mathsf{T}\left(\mathbf{d}_\mathrm{L}\mathbf{d}_\mathrm{L}^\mathsf{T} + \mathbf{d}_\mathrm{R}\mathbf{d}_\mathrm{R}^\mathsf{T}\right)\mathbf{v} \geq 0$$

with $q_\mathrm{T}$ defined in (31). Comparing with (17) we see that $h\gamma = 1/q_\mathrm{T}$.

### C.3 Proof of Lemma 3.7

In [13], it was shown that $\widetilde{q}_\mathrm{L,R}$ and $\widetilde{q}_\mathrm{C}$ in (32) can be computed as

$$\widetilde{q}_\mathrm{L} = \mathbf{d}_\mathrm{L}^\mathsf{T} K_0\mathbf{d}_\mathrm{L}, \qquad \widetilde{q}_\mathrm{R} = \mathbf{d}_\mathrm{R}^\mathsf{T} K_0\mathbf{d}_\mathrm{R}, \qquad \widetilde{q}_\mathrm{C} = \mathbf{d}_\mathrm{L}^\mathsf{T} K_0\mathbf{d}_\mathrm{R} = \mathbf{d}_\mathrm{R}^\mathsf{T} K_0\mathbf{d}_\mathrm{L}, \tag{61}$$

with $K_0$ defined (using our notation from (21)) as

$$K_0 = \begin{bmatrix} 0 & \vec{0}^\mathsf{T} \quad 0 \\ \vec{0} & \multirow{2}{*}{$\begin{bmatrix} \bar{A} & \vec{a}_\mathrm{R} \\ \vec{a}_\mathrm{R}^\mathsf{T} & a_\mathrm{R} \end{bmatrix}^{-1}$} \\ 0 & \end{bmatrix}.$$

Now, we want to show that the quantities in (61) are equal to the ones in (25). Applying the formula for inverses of block matrices to the above definition of $K_0$, and thereafter using the relation for $a_\mathrm{R}$ in (46), we obtain

$$\begin{aligned} K_0 &= \frac{1}{a_\mathrm{R} - \vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{R}} \begin{bmatrix} 0 & \vec{0}^\mathsf{T} & 0 \\ \vec{0} & (a_\mathrm{R} - \vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1}\vec{a}_\mathrm{R})\bar{A}^{-1} + \bar{A}^{-1}\vec{a}_\mathrm{R}\vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1} & -\bar{A}^{-1}\vec{a}_\mathrm{R} \\ 0 & -\vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1} & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \vec{0}^\mathsf{T} & 0 \\ \vec{0} & \bar{A}^{-1} & \vec{0} \\ 0 & \vec{0}^\mathsf{T} & 0 \end{bmatrix} + \ell \begin{bmatrix} 0 \\ -\bar{A}^{-1}\vec{a}_\mathrm{R} \\ 1 \end{bmatrix} \begin{bmatrix} 0 & -\vec{a}_\mathrm{R}^\mathsf{T}\bar{A}^{-1} & 1 \end{bmatrix}. \end{aligned} \tag{62}$$

Comparing (62) with (23) and (45), we note that $K_0 = G_2 + \mathbf{x}\mathbf{x}^\mathsf{T}/\ell$. Inserting this into (61), and thereafter using (50) and that $\mathbf{d}_\mathrm{L,R}^\mathsf{T}\mathbf{1} = 0$ and $\mathbf{d}_\mathrm{L,R}^\mathsf{T}\mathbf{x} = 1$, yields

$$\widetilde{q}_\mathrm{L} = -\mathbf{b}_\mathrm{L}^\mathsf{T}\mathbf{d}_\mathrm{L}, \qquad \widetilde{q}_\mathrm{R} = \mathbf{b}_\mathrm{R}^\mathsf{T}\mathbf{d}_\mathrm{R}, \qquad \widetilde{q}_\mathrm{C} = -\mathbf{b}_\mathrm{L}^\mathsf{T}\mathbf{d}_\mathrm{R} = \mathbf{b}_\mathrm{R}^\mathsf{T}\mathbf{d}_\mathrm{L},$$

that is exactly the same relations as in (25).

## D Explicit Inverses of the Second Derivative Operator

We provide the explicit expressions of $\bar{A}^{-1}$, $\mathbf{b}_\mathrm{L,R}$, $\xi_\mathrm{L,R}$ and $\xi_\mathrm{C}$ for the (2,0), (2,1) and (4,2) order accurate narrow-stencil operators and the (2,0) order accurate wide-stencil operator. By the notation "(2,0) order accurate operator", we refer to a matrix $D_2$ which has order 2 in the interior finite difference stencil and order 0 at the boundaries.

### D.1 The Narrow-Stencil (2,0) Order Operator

The simplest possible example of a second derivative operator $D_2$ fulfilling the SBP-properties in (13) is the narrow-stencil (2,0) order operator, and its corresponding matrix $\widetilde{A}$ was inverted already in [14] for the special case $\alpha_{\mathrm{L,R}} = 1$, $\beta_{\mathrm{L,R}} = 0$ and $\tau_{\mathrm{L,R}} = 0$. It is given below, together with its associated $\mathbf{d}_{\mathrm{L,R}}$ vectors.

$$D_2 = \frac{1}{h^2}\begin{bmatrix} 0 & 0 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 0 & 0 \end{bmatrix}, \quad \mathbf{d}_{\mathrm{L}} = \frac{1}{h}\begin{bmatrix} -1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \quad \mathbf{d}_{\mathrm{R}} = \frac{1}{h}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \\ 1 \end{bmatrix}. \tag{63}$$

The operator $D_2$ is also associated with $H = h \operatorname{diag}\left(\frac{1}{2}, 1, 1, \dots, 1, 1, \frac{1}{2}\right)$, and using (13) we obtain the $(n+1) \times (n+1)$ matrix $A$ given below. The $(n-1) \times (n-1)$ matrix $\bar{A}$ is identified using (21). Gauss–Jordan elimination then leads to $\bar{A}^{-1}$ as

$$A = \frac{1}{h}\begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}, \qquad \bar{A}^{-1} = h\begin{bmatrix} 1-\frac{1}{n} & 1-\frac{2}{n} & \cdots & \frac{1}{n} \\ 1-\frac{2}{n} & 2(1-\frac{2}{n}) & \cdots & \frac{2}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{2}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix}.$$

Inserting $\bar{A}^{-1}$ from above into (23), and using that $x_i = ih$, yields

$$(G_2)_{i,j} = \begin{cases} x_j(1 - x_i/\ell), & 0 \le j \le i \le n, \\ x_i(1 - x_j/\ell), & 0 \le i \le j \le n. \end{cases} \tag{64}$$

Note the striking similarity to the continuous Green's function in Remark 3.4. Next, by noticing the structure of $\mathbf{d}_{\mathrm{L,R}}$ in (63) and identifying the first and last columns of $\bar{A}^{-1}$ as $h(\vec{1} - \vec{x}/\ell)$ and $h\vec{x}/\ell$ we can compute $G_2\mathbf{d}_{\mathrm{L,R}}$ and consequently $\mathbf{b}_{\mathrm{L,R}}$ in (23) as

$$G_2\mathbf{d}_{\mathrm{L}} = \begin{bmatrix} 0 \\ \vec{1} - \vec{x}/\ell \\ 0 \end{bmatrix}, \qquad G_2\mathbf{d}_{\mathrm{R}} = -\begin{bmatrix} 0 \\ \vec{x}/\ell \\ 0 \end{bmatrix}, \qquad \mathbf{b}_{\mathrm{L}} = \mathbf{e}_{\mathrm{L}}, \qquad \mathbf{b}_{\mathrm{R}} = \mathbf{e}_{\mathrm{R}}.$$

Furthermore, inserting these $\mathbf{b}_{\mathrm{L,R}}$ and $\mathbf{d}_{\mathrm{L,R}}$ from (63) into (25), we obtain

$$\xi_{\mathrm{L}} = \xi_{\mathrm{R}} = 1/h, \qquad \xi_{\mathrm{C}} = 0.$$

### D.2 The Narrow-Stencil (2,1) Order Operator

The narrow-stencil (2,1) order operator (see Section C.1 in [25]), have the same matrices $H$ and $A$ as the (2,0) order operator, and hence its $G_2$ is given by (64). However, the difference matrices $\mathbf{d}_{\mathrm{L,R}}$ differ, for the (2,1) order operator they are

$$\mathbf{d}_{\mathrm{L}}^{\mathrm{T}} = \frac{1}{h}\begin{bmatrix} -\frac{3}{2} & 2 & -\frac{1}{2} & 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad \mathbf{d}_{\mathrm{R}}^{\mathrm{T}} = \frac{1}{h}\begin{bmatrix} 0 & \cdots & 0 & 0 & \frac{1}{2} & -2 & \frac{3}{2} \end{bmatrix}.$$

We can compute $G_2\mathbf{d}_\mathrm{L}$ as

$$G_2\mathbf{d}_\mathrm{L} = h\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1-\frac{1}{n} & 1-\frac{2}{n} & \cdots & \frac{1}{n} & 0 \\ 0 & 1-\frac{2}{n} & 2(1-\frac{2}{n}) & \cdots & \frac{2}{n} & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \frac{1}{n} & \frac{2}{n} & \cdots & 1-\frac{1}{n} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \frac{1}{h}\begin{bmatrix} -\frac{3}{2} \\ 2 \\ -\frac{1}{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{2}-\frac{1}{n} \\ 1-\frac{2}{n} \\ \vdots \\ \frac{1}{n} \\ 0 \end{bmatrix}$$

and repeating the procedure for $G_2\mathbf{d}_\mathrm{R}$ and thereafter using (23), we arrive at

$$\mathbf{b}_\mathrm{L} = \begin{bmatrix} 1 \; -\frac{1}{2} \; 0 \cdots 0 \; 0 \end{bmatrix}^\mathsf{T}, \qquad \mathbf{b}_\mathrm{R} = \begin{bmatrix} 0 \; 0 \cdots 0 \; -\frac{1}{2} \; 1 \end{bmatrix}^\mathsf{T}.$$

Finally, we use (25) to compute

$$\xi_{\mathrm{L,R}} = 2.5/h, \qquad \xi_\mathrm{C} = 0,$$

where $\xi_\mathrm{C} = 0$ holds for $n \geq 4$.

### D.3 The Narrow-Stencil (4,2) Order Operator

The operator $D_2$ with fourth order interior accuracy and diagonal norm $H$, see Section C.2 in [25], is associated with the difference operators

$$\mathbf{d}_\mathrm{L}^\mathsf{T} = \frac{1}{h}\left[\tfrac{-11}{6} \; 3 \; \tfrac{-3}{2} \; \tfrac{1}{3} \; 0 \cdots 0\right], \qquad \mathbf{d}_\mathrm{R}^\mathsf{T} = \frac{1}{h}\left[0 \cdots 0 \; \tfrac{-1}{3} \; \tfrac{3}{2} \; -3 \; \tfrac{11}{6}\right]. \tag{65}$$

Using (13) and identifying the interior of $A$ according to (21), we obtain

$$\bar{A} = \frac{1}{h}\begin{bmatrix} \frac{59}{24} & -\frac{59}{48} & 0 & & & & & & \\ -\frac{59}{48} & \frac{55}{24} & -\frac{59}{48} & \frac{1}{12} & & & & & \\ 0 & -\frac{59}{48} & \frac{59}{24} & -\frac{4}{3} & \frac{1}{12} & & & & \\ & \frac{1}{12} & -\frac{4}{3} & \frac{5}{2} & -\frac{4}{3} & \frac{1}{12} & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{12} & -\frac{4}{3} & \frac{5}{2} & -\frac{4}{3} & \frac{1}{12} & \\ & & & & \frac{1}{12} & -\frac{4}{3} & \frac{59}{24} & -\frac{59}{48} & 0 \\ & & & & & \frac{1}{12} & -\frac{59}{48} & \frac{55}{24} & -\frac{59}{48} \\ & & & & & & 0 & -\frac{59}{48} & \frac{59}{24} \end{bmatrix}.$$

We are now looking for a matrix $\bar{G}$ such that $\bar{G} = \bar{A}^{-1}$, and follow the same procedure as in "Section A.2" of Appendix. We make the ansatz

$$\bar{G} = \begin{bmatrix} \vec{g}_1 \; \vec{g}_2 \; \ldots \; \vec{g}_{n-1} \end{bmatrix}, \qquad \vec{g}_j = \begin{bmatrix} g_{1,j} \; g_{2,j} \; \ldots \; g_{n-1,j} \end{bmatrix}^\mathsf{T}.$$

For $\bar{A}\bar{G} = \bar{I}$ to hold, $\bar{A}\vec{g}_j = \vec{e}_j$ must be fulfilled for all $j = 1, 2, \ldots, n-1$, where the vector $\vec{e}_j = [0 \; \ldots \; 0 \; 1 \; 0 \; \ldots \; 0]^\mathsf{T}$ is non-zero only in its $j$th element. From the mid rows of $\bar{A}\vec{g}_j$, given the inner structure of $\bar{A}$, we thus need

$$g_{i-2,j} - 16g_{i-1,j} + 30g_{i,j} - 16g_{i+1,j} + g_{i+2,j} = 12h\delta_{i,j}, \qquad \begin{array}{l} \forall i = 4, 5, \ldots, n-4, \\ \forall j = 1, 2, \ldots, n-1, \end{array}$$

where $\delta_{i,j}$ is the Kronecker delta. Hence, the fourth order linear homogeneous recurrence relation $g_{i-2,j}-16g_{i-1,j}+30g_{i,j}-16g_{i+1,j}+g_{i+2,j}=0$ has to be fulfilled by almost all $g_{i,j}$. The explicit solution to this recursive relation has the form $g_{i,j}=c_1+c_2i+c_3\psi^i+c_4\psi^{-i}$, where $\psi=7+\sqrt{48}\approx 13.9$ and where $c_{1,2,3,4}$ are $j$-dependent constants. To be precise, $g_{i,j}$ has this form for $2\le i\le n-2$, and we need two versions of the $j$-dependent constants, that is $g_{i,j}=c_1^u+c_2^ui+c_3^u\psi^i+c_4^u\psi^{-i}$ for $2\le i\le j$ and $g_{i,j}=c_1^l+c_2^li+c_3^l\psi^i+c_4^l\psi^{-i}$ for $j\le i\le n-2$. For each $j=2,3,\dots,n-2$, we thus have eight unknown constants $c^u_{1,2,3,4}$ and $c^l_{1,2,3,4}$, as well as the two remaining unknowns $g_{1,j}$ and $g_{n-1,j}$. These are determined by the three first and the three last rows in the requirement $\bar{A}\vec{g}_j=\vec{e}_j$, which gives us six conditions. From the rows $i=j-1, j, j+1$, we get three more conditions and in addition, we demand that the two versions of $g_{j,j}$ are identical. Altogether, this leads to a $10\times 10$ system of equations which we solve using Gauss–Jordan elimination. The boundary columns $j=1$ and $j=n-1$ must be treated separately, in a similar manner. All in all, these steps lead to the elements of the inverse $(\bar{A}^{-1})_{i,j}=g_{i,j}$ as

$$(\bar{A}^{-1})_{i,j}=\kappa_{i,j}+\begin{cases} x_j(1-x_i/\ell), & 1\le j\le i\le n-1\\ x_i(1-x_j/\ell), & 1\le i\le j\le n-1,\end{cases}$$

which is thus similar to the second order version of $\bar{A}^{-1}$, plus an additional term $\kappa_{i,j}$. This additional correction term is, for $2\le i,j\le n-2$, given by

$$\kappa_{i,j}=\begin{cases} -h\dfrac{\mathcal{P}_j\mathcal{P}_{n-i}}{\mathcal{Q}_n}, & 2\le j\le i\le n-2,\\ -h\dfrac{\mathcal{P}_i\mathcal{P}_{n-j}}{\mathcal{Q}_n}, & 2\le i\le j\le n-2,\end{cases}$$

where

$$\mathcal{P}_i=\frac{(51-2\psi^{-1})\psi^{i-2}-(51-2\psi)\psi^{2-i}}{\psi-\psi^{-1}},$$
$$\mathcal{Q}_n=\frac{\psi^{n-4}(2\psi^{-1}-51)^2-\psi^{4-n}(2\psi-51)^2}{\psi-\psi^{-1}}.$$

Note that $\mathcal{Q}_n\neq 0$ (unless $n\approx 3.7$), so there is no risk of division by zero. Moreover, for $i,j=1$ or $i,j=n-1$ we have

$$\kappa_{1,j}=-h\frac{\mathcal{P}_{n-j}}{\mathcal{Q}_n},\qquad \kappa_{n-1,j}=-h\frac{\mathcal{P}_j}{\mathcal{Q}_n},\qquad 2\le j\le n-2,$$
$$\kappa_{i,1}=-h\frac{\mathcal{P}_{n-i}}{\mathcal{Q}_n},\qquad \kappa_{i,n-1}=-h\frac{\mathcal{P}_i}{\mathcal{Q}_n},\qquad 2\le i\le n-2,$$

and

$$\kappa_{1,1}=\kappa_{n-1,n-1}=-h\frac{\mathcal{P}_{n-2}}{2\mathcal{Q}_n}-h\frac{11}{118},\qquad \kappa_{1,n-1}=\kappa_{n-1,1}=-h\frac{\mathcal{P}_2}{2\mathcal{Q}_n}.$$

**Table 2** The parameters $h\xi_{\mathrm{L,R}}$ and $h\xi_{\mathrm{C}}$ in the (4,2) order case evaluated explicitly

| $n$ | $h\xi_{\mathrm{L,R}}$ | $h\xi_{\mathrm{C}}$ |
|---|---|---|
| 8 | 3.986350339808304 | 0.000041141179445 |
| 9 | 3.986350339313381 | 0.000002953803786 |
| 10 | 3.986350339310831 | 0.000000212073570 |
| 11 | 3.986350339310817 | 0.000000015226197 |
| 12 | 3.986350339310817 | 0.000000001093192 |

From (23) we have that the interior of $G_2$ is given by $\bar{A}^{-1}$ described above. Next, we use $\mathbf{d}_{\mathrm{L}}$ from (65) to compute $G_2\mathbf{d}_{\mathrm{L}}$ and thereafter (23) again, to compute $\mathbf{b}_{\mathrm{L}}$ as

$$(\mathbf{b}_{\mathrm{L}})_i = \begin{cases} 1 & i = 0 \\ -\frac{85}{118} + \frac{17}{2}\frac{\mathcal{P}_{n-2}}{\mathcal{Q}_n} & i = 1 \\ 17\frac{\mathcal{P}_{n-i}}{\mathcal{Q}_n} & i = 2, 3, \dots, n-2, \\ \frac{17}{\mathcal{Q}_n} & i = n-1 \\ 0 & i = n \end{cases} \qquad \lim_{n\to\infty} \mathbf{b}_{\mathrm{L}} = \begin{bmatrix} 1 \\ -0.5532\dots \\ 0.3342\dots \\ 0.0239\dots \\ \vdots \\ 0 \end{bmatrix}, \tag{66}$$

where we have used that $\mathcal{Q}_n + 2\mathcal{P}_{n-3} = 51\mathcal{P}_{n-2}$. Then, $\mathbf{b}_{\mathrm{R}}$ is given by $(\mathbf{b}_{\mathrm{R}})_i = (\mathbf{b}_{\mathrm{L}})_{n-i}$. We also compute the scalars from (25), as

$$\xi_{\mathrm{L}} = \xi_{\mathrm{R}} = \frac{1}{h}\left(\frac{2417}{354} - \frac{17^2\mathcal{P}_{n-2}}{2\mathcal{Q}_n}\right), \qquad \xi_{\mathrm{C}} = \frac{1}{h}\frac{17^2}{\mathcal{Q}_n}.$$

Evaluating $h\xi_{\mathrm{L,R}}$ and $h\xi_{\mathrm{C}}$ explicitly for some values of $n$, see Table 2, we see that these numbers corresponds exactly (to machine precision) to $\widetilde{q}_{\mathrm{L}}h$ and $\widetilde{q}_{\mathrm{C}}h$ tabulated in [13]. This serves as a numerical verification of Lemma 3.7 and indirectly of Theorem 3.1.

### D.4 The Wide-Stencil (2,0) Order Operator

The wide-stencil (2,0) order accurate operator $D_2$, which is obtained by squaring the (2,1) order accurate operator $D_1$ from (33), is given below together with $\mathbf{d}_{\mathrm{L,R}} = D_1^{\mathsf{T}}\mathbf{e}_{\mathrm{L,R}}$

$$D_2 = \frac{1}{h^2}\begin{bmatrix} \frac{1}{2} & -1 & \frac{1}{2} & & & & \\ \frac{1}{2} & -\frac{3}{4} & 0 & \frac{1}{4} & & & \\ \frac{1}{4} & 0 & -\frac{1}{2} & 0 & \frac{1}{4} & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \frac{1}{4} & 0 & -\frac{1}{2} & 0 & \frac{1}{4} \\ & & & \frac{1}{4} & 0 & -\frac{3}{4} & \frac{1}{2} \\ & & & & \frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}, \quad \mathbf{d}_{\mathrm{L}} = \frac{1}{h}\begin{bmatrix} -1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{d}_{\mathrm{R}} = \frac{1}{h}\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

The operator is also associated with the same $H = h\mathrm{diag}\left(\frac{1}{2}, \quad 1, \quad 1, \quad \dots, \quad 1, \quad 1, \quad \frac{1}{2}\right)$ as the other operators with second order accuracy, and from this we can compute the $(n+1)\times(n+1)$ matrix $A$. Identifying the parts of $A$ according to (21), gives us the $(n-1)\times(n-1)$

matrix $\bar{A}$. The inverse of this matrix $\bar{A}$ is

$$\bar{A}^{-1} = 2h \begin{bmatrix} 1-\frac{1}{n} & 0 & 1-\frac{3}{n} & \cdots \\ 0 & 2(1-\frac{2}{n}) & 0 & \cdots \\ 1-\frac{3}{n} & 0 & 3(1-\frac{3}{n}) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

that is the discrete Green's function in (23) becomes

$$(G_2)_{i,j} = \begin{cases} x_j(1-x_i/\ell)(1+(-1)^{i+j}), & 0 \le j \le i \le n, \\ x_i(1-x_j/\ell)(1+(-1)^{i+j}), & 0 \le i \le j \le n. \end{cases}$$

Thus the discrete Green's function produced by the wide operator oscillate, jumping between 0 and 2 times the exact value. Next, using (23) we obtain the vectors

$$\mathbf{b}_{\mathrm{L}}^{\mathsf{T}} = \left[1 \ -(1-\tfrac{1}{n}) \ 1-\tfrac{2}{n} \ -(1-\tfrac{3}{n}) \ \ldots \ (-1)^n\tfrac{2}{n} \ -(-1)^n\tfrac{1}{n} \ 0\right],$$
$$\mathbf{b}_{\mathrm{R}}^{\mathsf{T}} = \left[0 \ -(-1)^n\tfrac{1}{n} \ (-1)^n\tfrac{2}{n} \ \ldots \ -(1-\tfrac{3}{n}) \ 1-\tfrac{2}{n} \ -(1-\tfrac{1}{n}) \ 1\right].$$

Last, we compute the (2,0) order wide-stencil version of (25), as

$$\xi_{\mathrm{L}} = \xi_{\mathrm{R}} = \frac{2}{h} - 1/\ell, \qquad \xi_{\mathrm{C}} = -(-1)^n/\ell.$$

In the wide-stencil case, $q_{\mathrm{L,R}} = \mathbf{e}_{\mathrm{L,R}}^{\mathsf{T}} H^{-1}\mathbf{e}_{\mathrm{L,R}} = 2/h$ and $q_{\mathrm{C}} = \mathbf{e}_{\mathrm{L,R}}^{\mathsf{T}} H^{-1}\mathbf{e}_{\mathrm{R,L}} = 0$ can be computed directly. We recall that $\widetilde{q}_{\mathrm{L,R,C}} = \xi_{\mathrm{L,R,C}}$ and note that $\widetilde{q}_{\mathrm{L,R}} \neq q_{\mathrm{L,R}}$ and $\widetilde{q}_{\mathrm{C}} \neq q_{\mathrm{C}}$, but still $\widetilde{q}_{\mathrm{T}} = q_{\mathrm{T}} = 2/h$. Compare with the discussion in Sect. 3.4.

## References

1. Almquist, M., Wang, S., Werpers, J.: Order-preserving interpolation for summation-by-parts operators at nonconforming grid interfaces. SIAM J. Sci. Comput. **41**(2), 1201–1227 (2019)
2. Appelö, D., Kreiss, G.: Application of a perfectly matched layer to the nonlinear wave equation. Wave Motion **44**(7), 531–548 (2007)
3. 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)
4. Beyn, W.-J.: Discrete Green's functions and strong stability properties of the finite difference method. Appl. Anal. **14**(2), 73–98 (1982)
5. Brandén, H., Holmgren, S., Sundqvist, P.: Discrete fundamental solution preconditioning for hyperbolic systems of PDE. J. Sci. Comput. **30**(1), 35–60 (2007)
6. Carpenter, M.H., Gottlieb, D., Abarbanel, S.: Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes. J. Comput. Phys. **111**(2), 220–236 (1994)
7. 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)
8. Chung, F., Yau, S.-T.: Discrete Green's functions. J. Comb. Theory Ser. A **91**(1), 191–214 (2000)
9. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen differenzengleichungen der mathematischen physik. Math. Ann. **100**(1), 32–74 (1928)
10. Deeter, C.R., Springer, G.: Discrete harmonic kernels. J. Math. Mech. **14**(3), 413–438 (1965)
11. Del Rey Fernández, D.C., Boom, P.D., Zingg, D.W.: A generalized framework for nodal first derivative summation-by-parts operators. J. Comput. Phys. **266**, 214–239 (2014)
12. Del Rey Fernández, D.C., 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)
13. Eriksson, S.: A dual consistent finite difference method with narrow stencil second derivative operators. J. Sci. Comput. **75**(2), 906–940 (2018)

14. Eriksson, S., Nordström, J.: Analysis of the order of accuracy for node-centered finite volume schemes. Appl. Numer. Math. **59**(10), 2659–2676 (2009)
15. Gong, J., Nordström, J., Stable: Accurate and Efficient Interface Procedures for Viscous Problems. Technical Report 2006-19, Department of Information Technology, Uppsala University, Uppsala, Sweden (2006)
16. Grote, M., Huckle, T.: Parallel preconditioning with sparse approximate inverses. SIAM J. Sci. Comput. **18**(3), 838–853 (1997)
17. Gustafsson, B., Kreiss, H.-O., Oliger, J.: Time-Dependent Problems and Difference Methods. Wiley, New York (2013)
18. 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)
19. Hicken, J.E., Zingg, D.W.: Summation-by-parts operators and high-order quadrature. J. Comput. Appl. Math. **237**(1), 111–125 (2013)
20. Kreiss, H.-O., Lorenz, J.: Initial-Boundary Value Problems and the Navier–Stokes Equations. Academic Press, Boston (1989)
21. Kreiss, H.-O., Scherer, G.: Finite element and finite difference methods for hyperbolic partial differential equations. In: De Boor, C. (ed.) Mathematical Aspects of Finite Elements in Partial Differential Equation. Academic Press, New York (1974)
22. Linders, V., Nordström, J., Frankel, S.H.: Properties of Runge–Kutta-summation-by-parts methods. J. Comput. Phys. **419**, 109684 (2020)
23. Mattsson, K., Almquist, M.: A solution to the stability issues with block norm summation by parts operators. J. Comput. Phys. **253**, 418–442 (2013)
24. Mattsson, K., Ham, F., Iaccarino, G.: Stable and accurate wave-propagation in discontinuous media. J. Comput. Phys. **227**(19), 8753–8767 (2008)
25. Mattsson, K., Nordström, J.: Summation by parts operators for finite difference approximations of second derivatives. J. Comput. Phys. **199**(2), 503–540 (2004)
26. Mattsson, K., Svärd, M., Shoeybi, M.: Stable and accurate schemes for the compressible Navier–Stokes equations. J. Comput. Phys. **227**(4), 2293–2316 (2008)
27. Ruggiu, A.A., Nordström, J.: Eigenvalue analysis for summation-by-parts finite difference time discretizations. SIAM J. Numer. Anal. **58**(2), 907–928 (2020)
28. Stetter, H.J.: Instability and non-monotonicity phenomena in discretizations to boundary-value problems. Numer. Math. **12**(2), 139–145 (1968)
29. Strand, B.: Summation by parts for finite difference approximation for d/dx. J. Comput. Phys. **110**(1), 47–67 (1994)
30. Svärd, M., Nordström, J.: A stable high-order finite difference scheme for the compressible Navier–Stokes equations: no-slip wall boundary conditions. J. Comput. Phys. **227**(10), 4805–4824 (2008)
31. Svärd, M., Nordström, J.: Review of summation-by-parts schemes for initial-boundary-value problems. J. Comput. Phys. **268**, 17–38 (2014)
32. Wang, S., Kreiss, G.: Convergence of summation-by-parts finite difference methods for the wave equation. J. Sci. Comput. **71**(1), 219–245 (2017)