Provably non-stiff implementation of weak coupling conditions for hyperbolic problems

Ossian O'Reilly¹ · Jan Nordström^{2,3}

Received: 15 July 2020 / Revised: 14 December 2021 / Accepted: 14 December 2021 /

Published online: 5 January 2022

© The Author(s) 2022

Abstract

In the context of coupling hyperbolic problems, the maximum stable time step of an explicit numerical scheme may depend on the design of the coupling procedure. If this is the case, the coupling procedure is sensitive to changes in model parameters independent of the Courant–Friedrichs–Levy condition. This sensitivity can cause artificial stiffness that degrades the performance of a numerical scheme. To overcome this problem, we present a systematic and general procedure for weakly imposing coupling conditions via penalty terms in a provably non-stiff manner. The procedure can be used to construct both energy conservative and dissipative couplings, and the user is given control over the amount of dissipation desired. The resulting formulation is simple to implement and dual consistent. The penalty coefficients take the form of projection matrices based on the coupling conditions. Numerical experiments demonstrate that this procedure results in both optimal spectral radii and superconvergent linear functionals.

Mathematics Subject Classification 65M99

1 Introduction

Many scientific and engineering problems require numerical methods that couple different wave phenomena. Numerous applications can be found within computational

✉ Jan Nordström
jan.nordstrom@liu.se

¹ Formerly at Southern California Earthquake Center, University of Southern California, Los Angeles, CA 90089, USA

² Department of Mathematics, Applied Mathematics, Linköping University, 581 83 Linköping, Sweden

³ Department of Mathematics and Applied Mathematics, University of Johannesburg, P.O. Box 524, Auckland Park, Johannesburg 2006, South Africa

electrodynamics, computational geophysics, and computational fluid dynamics. Typical examples include the Euler equations coupled to the elastic wave equation, and electromagnetic wave interactions with scattering across material interfaces. A difficulty that can arise when developing numerical methods for these types of coupled problems is that the resulting discretization may suffer from stiffness. In such a case, an explicit time stepping procedure may require a significantly smaller time step than what is expected from each separate subproblem. For example, within fluid-structure interaction, the impact of the coupling procedure is a well-known problem. The so-called added mass instability is an effect of coupling light fluids to thin, dense structures immersed in fluid. This problem has led to a plethora of different coupling strategies, e.g., see [2,6,32]. Potential stiffness issues related to large density contrasts also appear in linear settings when coupling acoustic and elastic wave equations. Stiffness can also appear in other settings, e.g., for frictional sliding interfaces [17], and narrow cracks [24].

The imposition of coupling conditions is closely related to the imposition of boundary conditions where stiffness can occur for similar reasons. In this paper, we will for ease of presentation switch between these two use cases. We focus on weakly imposed boundary and coupling conditions by including additional source terms in the governing equations, so-called penalty terms. The role of the penalty terms is to force the numerical solution on the interface towards satisfying the actual coupling conditions [4,30]. The penalty term can be interpreted as specific weightings of the coupling conditions in combination with the equations.

There is some flexibility in choosing penalty weights, and the choice can have a striking impact on a scheme's performance. Well-designed penalty terms can be the difference between having a scheme that works in practice or a suboptimal scheme that runs orders of magnitude slower. Unfortunately, it is quite easy to construct suboptimal penalty terms that cause these problems, even though they are semi-discretely stable. Coupling procedures that overcome these problems use either upwind numerical fluxes based on solving the Riemann problem [34], or a specific characteristic treatment [17,20,21]. In each of the above cases, the penalty terms introduce some amount of potentially unwanted additional dissipation.

Adding certain types of artificial dissipation may result in improved accuracy [16,21]. However, dissipation may be suboptimal for wave propagation problems since that may prevent the use of symplectic or staggered time-stepping methods that have improved accuracy and stability properties [33]. For instance, a benefit of the fourth order staggered Runge–Kutta time-stepping scheme introduced in [11] is that it has a factor 16 smaller truncation error and a factor of two larger stability region along the imaginary axis compared to the classical fourth order Runge–Kutta scheme. Unfortunately, the staggered Runge–Kutta scheme is unstable even for a small amount of artificial dissipation [11,33]. An energy conservative weak coupling procedure for coupling acoustic materials across nonconforming grids is presented in [8]. However, this coupling procedure causes stiffness for problems with large density contrasts. To the best of our knowledge, there are no known penalty procedures that are both energy conservative and non-stiff. In this paper, we present a systematic procedure for constructing both energy conservative and dissipative penalty terms that are provably non-stiff.

We argue that numerical schemes should be provably stable and non-stiff. A powerful method for proving semi-discrete stability of a scheme is the energy method [13]. This method gives sufficient conditions for semi-discrete stability, but not necessary ones. However, when the energy method fails, it is usually an indication that there is an underlying instability present that may be triggered under certain circumstances. Without a proof of stability, there are no guarantees that the output produced by a numerical scheme is reliable. A proof of stability also helps to rule out bugs and can serve as a strict test of the implementation. For instance, one can randomly initialize all solution fields and model parameters. Afterwards, one numerically computes the discrete energy rate in the implementation and asserts that it is non-positive.

Similar arguments can be made for having a provably non-stiff scheme. A proof of non-stiffness gives a sufficient condition in order to prevent stiffness from occurring due to sensitivity in model parameters. We will show that if the matrix norm of each penalty matrix can be bounded by the CFL condition, it is provably non-stiff. Since this is only a sufficient condition, a lack of proof does not imply that the scheme is stiff. However, failure to bound the matrix norm could be an indication that the numerical scheme is sensitive to certain parameter values. A proof of non-stiffness also helps to rule out bugs in the implementation. For instance, if the simulation of a provably stable and non-stiff scheme becomes unstable due to a change in a model parameter independent of the CFL-condition, it must be caused by an incorrect implementation. In addition, since the proof of non-stiffness constrains the parameters to some degree, this analysis can limit the potential parameter space that one must probe to select optimal parameter values.

The key developments in this work are applicable to any numerical scheme in Summation-By-Parts (SBP) form that weakly imposes boundary or coupling conditions via Simultaneous-Approximation-Terms (SAT). Numerical schemes that are in SBP-SAT form lead to a proof of stability via the energy method and apply to finite difference [5, 19, 28], finite volume [22], spectral collocation [3], discontinuous Galerkin [9], and flux reconstruction schemes [26]. See [7, 30] for reviews. When developing SBP-SAT schemes, one can focus on formulating and analyzing the continuous problem. Once this analysis is complete one can proceed by discretizing the continuous formulation and obtain a provably stable scheme without much difficulty, as shown in [20]. This aspect allows us to construct penalty terms in the continuous setting, leaving out specific implementation details for the variety of numerical schemes that belong to the SBP-SAT family. Hence, the focus of this paper is on constructing and analyzing penalty terms in a continuous setting without placing emphasis on particular discretization or implementation details.

This paper is organized as follows. We begin in Sect. 2 by discussing a motivating example: the coupling of wave equations. This example demonstrates how naively selected penalty weights cause stiffness. Section 3 presents the general theory for constructing non-stiff penalty terms. In particular, it is shown that the penalty terms are connected to a projection matrix. Section 4 shows that penalty terms constructed via a projection matrix formula automatically result in a dual consistent formulation. Section 5 introduces a diagnostic test to flag penalty formulations suffering from stiffness, and shows that certain penalty terms obtained by the projection matrix formula

Fig. 1 Problem setup for coupling the wave equation across an interface Γ

are provably non-stiff. Section 6 revisits the motivating example and applies the general theory to construct both non-stiff, energy conservative, and energy dissipative penalty terms. Section 7 introduces numerical experiments that exemplify our theoretical developments. In particular, we present a challenging 2D air-water interface problem for the wave equation. Finally, we discuss our findings in Sect. 8.

2 A motivating example

Artificial stiffness can easily emerge in coupled problems [10, 17]. In the decoupled case, each problem has some maximum stable time step $\Delta t_{max}^{(i)}$, for $i = 1, 2$ when using a specific explicit time integrator. On the other hand, if the two subdomains are coupled together, one might obtain a different Δt_{max} that can be much larger than either $\Delta t_{max}^{(1)}$ or $\Delta t_{max}^{(2)}$. This phenomenon stems from a suboptimal selection of penalty weights in the coupling procedure. We demonstrate this problem by coupling two wave equations at an interface Γ (see Fig. 1).

The wave equation for each side of the interface is written as a first order hyperbolic system of equations in Cartesian coordinates,

$$\partial_t q^{(i)} + A_x^{(i)} \partial_x q^{(i)} + A_y^{(i)} \partial_y q^{(i)} = f^{(i)}, \quad (x, y) \in \Omega^{(i)}, \quad t \geq 0, \quad (1)$$

where $q^{(i)} = (p^{(i)} / \rho_i c_i, v_x^{(i)}, v_y^{(i)})$, $i = 1, 2$. The vector $q^{(i)}$ collects the pressure $p^{(i)}$ and the components of the velocity field $v^{(i)} = (v_x^{(i)}, v_y^{(i)})$, decomposed with respect to the Cartesian basis. Moreover,

$$A_x^{(i)} = \begin{bmatrix} 0 & c_i & 0 \\ c_i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_y^{(i)} = \begin{bmatrix} 0 & 0 & c_i \\ 0 & 0 & 0 \\ c_i & 0 & 0 \end{bmatrix}.$$

The material parameters $\rho_1 > 0$ and $\rho_2 > 0$ are the respective densities on each side of the interface. Similarly, c_1 and c_2 are the respective wave speeds. The forcing/penalty terms $f^{(i)}$ are responsible for weakly imposing coupling conditions in each subdomain.

2.1 Penalty terms

A general formulation of the penalty terms is

$$f^{(i)} = \mathcal{L}(\Sigma^{(i)} L^T u), \quad (2)$$

where $u = (q^{(1)}, q^{(2)})$. The role of the *lifting operator* $\mathcal{L}$ is to make sure that the penalty term acts on the boundary only [1, 27]. In one dimension, it is analogous to the Dirac delta function $\delta(x - x_0)$, where x_0 is the boundary point. In higher dimensions, it is defined by

$$\int_{\Omega} v^T \mathcal{L}(u) d\Omega = \int_{\partial\Omega} v^T u ds, \quad (3)$$

for smooth and vector-valued functions u and v , and where $\partial\Omega$ is an appropriate boundary part of the domain Ω . This relationship is similar to the divergence theorem in that it relates a volume/surface integral (left-hand side) to a surface/line integral (right-hand side). The matrix L^T is the boundary operator, determined by the coupling conditions, and $\Sigma^{(i)}$ are penalty matrices. The penalty matrices describe how to form linear combinations of the coupling conditions for each equation.

2.2 Well-posedness

For the problem to be well-posed, there are three general requirements. See [20] for further details.

Definition 1 The problem (1) subject to the weak boundary/coupling conditions (2) is well-posed if:

1. A solution exists,
2. The solution is unique,
3. The solution is bounded.

For existence, there cannot be too many boundary conditions, for uniqueness there cannot be too few boundary conditions, and for a bound, the boundary condition must have appropriate form. In this work, we weakly impose the minimum number of appropriate boundary/coupling conditions that bound the solution.

2.2.1 Coupling conditions

One set of well-posed and physically relevant coupling conditions for the wave equation are

$$p^{(1)} - p^{(2)} = 0, \quad v^{(1)} \cdot n - v^{(2)} \cdot n = 0 \quad (x, y) \in \Gamma. \quad (4)$$

In (4), $n = (n_x, n_y)$ is the unit normal on the interface Γ , directed towards $\Omega^{(2)}$. The coupling conditions (4) can be expressed in matrix vector form $L^T u$, as done in (2), where

$$L^T u = \begin{bmatrix} p^{(1)} - p^{(2)} \\ v_n^{(1)} - v_n^{(2)} \end{bmatrix}, \quad L^T = \begin{bmatrix} \rho_1 c_1 & 0 & 0 & -\rho_2 c_2 & 0 & 0 \\ 0 & n_x & n_y & 0 & -n_x & -n_y \end{bmatrix}, \quad (5)$$

and $v_n^{(1)} = v^{(1)} \cdot n$, $v_n^{(2)} = v^{(2)} \cdot n$.

As previously mentioned, the penalty terms are formed by taking a linear combination of the coupling conditions (4) and distributing them among the equations, yielding

$$f^{(i)} = \mathcal{L} \left(\Sigma^{(i)} L^T u \right) = \mathcal{L} \left(\begin{bmatrix} \sigma_{11}^{(i)} (p^{(1)} - p^{(2)}) + \sigma_{12}^{(i)} (v_n^{(1)} - v_n^{(2)}) \\ \sigma_{21}^{(i)} (p^{(1)} - p^{(2)}) + \sigma_{22}^{(i)} (v_n^{(1)} - v_n^{(2)}) \\ \sigma_{31}^{(i)} (p^{(1)} - p^{(2)}) + \sigma_{32}^{(i)} (v_n^{(1)} - v_n^{(2)}) \end{bmatrix} \right), \quad (6)$$

where

$$\Sigma^{(i)} = \begin{bmatrix} \sigma_{11}^{(i)} & \sigma_{12}^{(i)} \\ \sigma_{21}^{(i)} & \sigma_{22}^{(i)} \\ \sigma_{31}^{(i)} & \sigma_{32}^{(i)} \end{bmatrix}.$$

2.3 The energy method

The weights in the penalty matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are determined by bounding the solution via the energy method [13].

Let $\|u\| = \sqrt{\int_{\Omega} u^T u d\Omega}$ denote the L_2 -norm. By differentiating $\|q^{(i)}\|^2$ with respect to t and substituting $\partial_t q^{(i)}$ using (1) it follows that

$$\frac{d\|q^{(i)}\|^2}{dt} = - \int_{\Omega^{(i)}} 2(q^{(i)})^T \left(A_x^{(i)} \partial_x q^{(i)} + A_y^{(i)} \partial_y q^{(i)} - f^{(i)} \right) d\Omega. \quad (7)$$

By applying Gauss's theorem and the definition of the lifting operator (3), the volume integral in (7) can be converted into the following surface integral,

$$\frac{d\|q^{(i)}\|^2}{dt} = - \int_{\Gamma} (q^{(i)})^T \tilde{A}^{(i)} q^{(i)} - 2(q^{(i)})^T \Sigma^{(i)} L^T u ds, \quad (8)$$

where $\tilde{A}^{(1)} = A_x^{(1)} n_x + A_y^{(1)} n_y$, $\tilde{A}^{(2)} = -A_x^{(2)} n_x - A_y^{(2)} n_y$. Equation (8) shows that the energy $\|q^{(i)}\|$ is conserved in the interior, and changes in the energy rate are purely determined by fluxes across the interface. Note that in (8), the energy rate contribution from exterior boundaries have been neglected.

By summing over $i = 1, 2$ in (8) and scaling each term by the positive parameters α and β , we get

$$\alpha \frac{d\|q^{(1)}\|^2}{dt} + \beta \frac{d\|q^{(2)}\|^2}{dt} = - \int_{\Gamma} u^T \hat{A}u - 2u^T \hat{\Sigma} L^T u ds, \quad (9)$$

where

$$\hat{A} = \text{diag} \left(\alpha \tilde{A}^{(1)}, \beta \tilde{A}^{(2)} \right) \text{ and } \hat{\Sigma} = \begin{bmatrix} \alpha \Sigma^{(1)} \\ \beta \Sigma^{(2)} \end{bmatrix}.$$

Following [8], we can prove.

Proposition 1 *The solution of the problem (1) is bounded and the energy (9) is conserved if the penalty matrices in (6) are chosen as*

$$\sigma_{11}^{(i)} = \sigma_{22}^{(i)} = \sigma_{32}^{(i)} = 0, \quad (10)$$

$$\sigma_{12}^{(i)} = c_i k_i, \quad \sigma_{21}^{(i)} = \frac{n_x}{\rho_i} l_i, \quad \sigma_{31}^{(i)} = \frac{n_y}{\rho_i} l_i, \quad (11)$$

where k_i and l_i are real parameters that satisfy the constraints

$$1 - k_1 - l_1 = 0, \quad k_1 - l_2 = 0, \quad l_1 - k_2 = 0. \quad (12)$$

Proof A direct calculation yields,

$$u^T \hat{A}u = \alpha (q^{(1)})^T \tilde{A}^{(1)} q^{(1)} + \beta (q^{(2)})^T \tilde{A}^{(2)} q^{(2)} = 2 \left(\frac{\alpha}{\rho_1} p^{(1)} v_n^{(1)} - \frac{\beta}{\rho_2} p^{(2)} v_n^{(2)} \right).$$

Choosing $\alpha = \rho_1$ and $\beta = \rho_2$, followed by expanding (9), yields

$$\sum_{i=1}^2 \frac{\rho_i}{2} \frac{d\|q^{(i)}\|^2}{dt} = - \int_{\Gamma} p^{(1)} v_n^{(1)} - p^{(2)} v_n^{(2)} - \mathcal{R} ds, \quad (13)$$

where

$$\begin{aligned} \mathcal{R} = & \left(\sum_{i=1}^2 \rho_i \left(\sigma_{11}^{(i)} \frac{p^{(i)}}{\rho_i c_i} + \sigma_{21}^{(i)} v_x^{(i)} + \sigma_{31}^{(i)} v_y^{(i)} \right) \right) (p^{(1)} - p^{(2)}) \\ & + \left(\sum_{i=1}^2 \rho_i \left(\sigma_{12}^{(i)} \frac{p^{(i)}}{\rho_i c_i} + \sigma_{22}^{(i)} v_x^{(i)} + \sigma_{32}^{(i)} v_y^{(i)} \right) \right) (v_n^{(1)} - v_n^{(2)}). \end{aligned} \quad (14)$$

Finally, the right-hand side in (13) vanishes if (10)–(12) holds. $\square$

2.4 Stiffness

Proposition 1 states sufficient conditions for obtaining an energy conservative coupling. However, the constraints (12) imply that there is one free parameter. The problem we deal with in this paper is: how should this parameter be chosen? It might be tempting to try to make all parameters share the same magnitude. For instance, as was done in [8], one can choose

$$k_1 = k_2 = l_1 = l_2 = \frac{1}{2}. \quad (15)$$

However, as we will see next, this choice causes stiffness (as do many other choices).

Before we proceed, we need to discuss the spatial discretization of the coupled wave equation problem. The discretization of (1) in space leads to linear system of ordinary differential equations:

$$\frac{d\mathbf{q}}{dt} = M_h \mathbf{q}, \quad \mathbf{q} = \begin{bmatrix} \mathbf{p} \\ \mathbf{v} \end{bmatrix}, \quad (16)$$

where $\mathbf{p}$ and $\mathbf{v}$ discretize the pressure and velocity fields, and M_h is a sparse matrix that holds the discretization of the spatial operators and all penalty terms; applied to each side of the interface (see Sect. 5.2 for more details). Since we are only interested in understanding how the penalty parameters influence Δt_{max} , the specific spatial discretization technique is not important (e.g., finite difference, finite volume, discontinuous Galerkin, etc.).

The maximum stable time step of an explicit time integrator can be estimated as

$$\Delta t_{max} = CFL \frac{h}{c_{max}}. \quad (17)$$

In (17), h is the grid spacing (taken to be constant and the same in each subdomain), and $c_{max} = \max(c_1, c_2)$. To prevent stiffness from occurring, Δt_{max} must be no smaller than either $\Delta t_{max}^{(1)}$ and $\Delta t_{max}^{(2)}$, associated with the respective uncoupled cases. The naive parameter choice (15) violates this condition. For this choice, the spectral radius of M_h , denoted $\rho(M_h)$, depends on the density ratio and that causes stiffness (see Fig. 2). For example, if $\rho_1/\rho_2 \approx 100$ then Δt_{max} must be reduced by a factor of ≈ 10 compared to $\min(\Delta t_{max}^{(1)}, \Delta t_{max}^{(2)})$. Our ambition in this paper is to find a procedure that removes the kink in Fig. 2 and minimizes the spectral radius.

3 General treatment

As made apparent by the previous example, the design of penalty terms can strongly influence the efficacy of the numerical scheme. One option is to perform numerical experiments to find optimal penalty weights, but that can be cumbersome and imprac-

Fig. 2 The maximum stable time step Δt_{max} as a function of the density ratio ρ_1/ρ_2 (on a log-log scale). This plot was generated by discretizing the wave equation in (18) using $c = c_1 = c_2$, fourth order SBP finite difference operators [28], and 100 grid points in each subdomain

tical. In this section, we present a general approach that can automatically leads to non-stiff penalty weights.

We again consider the problem (1) but for simplicity focus on one subdomain, and now with general and symmetric A_x and A_y matrices:

$$\partial_t q + A_x \partial_x q + A_y \partial_y q = f, \quad (x, y) \in \Omega, \quad t \geq 0. \quad (18)$$

Since we are only considering one subdomain, all superscripts have been dropped. Our procedure for deriving penalty terms is inspired by the procedure developed in [17, 18]. We start with a modified ansatz of the penalty term,

$$f = \mathcal{L}(A\delta q), \quad A = A_x n_x + A_y n_y. \quad (19)$$

In this formulation, $\delta q = \delta q(q)$ is an unknown penalty vector that depends on q such that a minimum number of boundary conditions are weakly imposed.

3.1 The energy method

Next, we introduce the bilinear form

$$\Phi(q, v) = \oint_{\partial\Omega} q^T A v ds. \quad (20)$$

Then, the energy method applied to (18) can now be written as

$$\frac{d}{dt} \|q\|^2 = -\Phi(q, q) + 2\Phi(q, \delta q).$$

We can reformulate the above by completing the square, yielding

$$\frac{d}{dt} \|q\|^2 = -\Phi(q - \delta q, q - \delta q) + \Phi(\delta q, \delta q). \quad (21)$$

The solution is bounded if the right-hand side of (21) is non-positive. However, here we enforce the stronger requirement that each term on the right-hand side is *individually* non-positive (not just their sum):

$$-\Phi(q - \delta q, q - \delta q) \leq 0 \quad \text{and} \quad \Phi(\delta q, \delta q) \leq 0, \quad (22)$$

We will see that the two conditions in (22) uniquely determine δq .

3.2 Diagonalization

Since A is symmetric, it is diagonalizable by a complete set of orthogonal eigenvectors, $A = X\Lambda X^T$, with $X^{-1} = X^T$. Here, X is a matrix that contains the eigenvectors arranged as column vectors, and Λ is diagonal matrix containing the corresponding eigenvalues. The eigenvalue matrix Λ is split into positive $\Lambda^+ > 0$ and negative parts $\Lambda^- < 0$, and $X = [X^+, X^-]$. Without loss of generality, we assume that A contains no zero eigenvalues (if it does, they are ignored).

We also define the transformation $w = \sqrt{|\Lambda|}X^T q$, leading to

$$w^+ = \sqrt{\Lambda^+}(X^+)^T q \quad \text{and} \quad w^- = \sqrt{|\Lambda^-|}(X^-)^T q. \quad (23)$$

This way of defining w^+ and w^- simplifies the upcoming presentation. For the other variable, δq , we define similar transforms, i.e.,

$$\delta w^+ = \sqrt{\Lambda^+}(X^+)^T \delta q, \quad \text{and} \quad \delta w^- = \sqrt{|\Lambda^-|}(X^-)^T \delta q. \quad (24)$$

When the eigendecomposition is applied to the quadratic form $\Phi(q, q)$, we find

$$\Phi(q, q) = \oint_{\partial\Omega} q^T A q ds = \oint_{\partial\Omega} (w^+)^T w^+ - (w^-)^T w^- ds. \quad (25)$$

3.3 Boundary conditions

Recall that the goal is to determine the penalty vector δq such that (22) holds subject to the boundary conditions.

3.3.1 Strong boundary conditions

For strong boundary conditions, it is natural to impose (see [20])

$$w^- = R w^+, \quad (x, y) \in \partial\Omega, \quad (26)$$

where R is a rectangular matrix that imposes a minimum number of boundary conditions. From (25), we get

$$\Phi(q, q) = \oint_{\partial\Omega} (w^+)^T (w^+) - (w^-)^T (w^-) ds = \oint_{\partial\Omega} (w^+)^T (I^+ - R^T R) w^+ ds.$$

For strong boundary conditions, $\delta q = 0$, the energy rate (21) becomes

$$\frac{d}{dt} \|q\|^2 = -\Phi(q, q). \quad (27)$$

It follows that the right-hand side in (27) is non-positive if $R^T R \leq I^+$, where I^+ is the identity matrix related to the size of w^+ . Clearly, this condition is satisfied when R is sufficiently small.

3.3.2 The weak primal condition

To satisfy $\Phi(q - \delta q, q - \delta q) \geq 0$ in (22), we proceed in the same way and specify

$$w^- - \delta w^- = R(w^+ - \delta w^+), \quad (x, y) \in \partial\Omega. \quad (28)$$

If $R^T R \leq I^+$, we obtain

$$\Phi(q - \delta q, q - \delta q) = \oint_{\partial\Omega} (w^+ - \delta w^+)^T (I^+ - R^T R) (w^+ - \delta w^+) ds \geq 0. \quad (29)$$

3.3.3 The weak dual condition

As seen in (22), we also need $\Phi(\delta q, \delta q) \leq 0$. To obtain that, we specify

$$\delta w^+ = \delta R \delta w^-, \quad (x, y) \in \partial\Omega, \quad (30)$$

where δR is another rectangular matrix. If $\delta R^T \delta R \leq I^-$, we obtain

$$\Phi(\delta q, \delta q) = - \oint_{\partial\Omega} (\delta w^-)^T (I^- - (\delta R)^T (\delta R)) \delta w^- ds \leq 0. \quad (31)$$

Interestingly, the condition $(\delta R)^T (\delta R) \leq I^-$ corresponds to the well-posedness condition for the dual problem in (18). This is further discussed in Appendix 4.

3.4 Well-posedness

The conditions (28) and (30) determine δq uniquely under the conditions given in the following Proposition.

Proposition 2 *Let $R, \delta R$ be the matrices in (28), (30) satisfying (29), (31). Then, if $\det(I^- - R\delta R) \neq 0$,*

$$\delta q = \left(X^+ \sqrt{(\Lambda^+)^{-1}} \delta R + X^- \sqrt{|\Lambda^-|^{-1}} \right) (I^- - R\delta R)^{-1} (w^- - R w^+). \quad (32)$$

Proof The starting point is the eigenvector transformation (24). Since $X^T = X^{-1}$, the inverse transformation is

$$\delta q = X^+ \sqrt{(\Lambda^+)^{-1}} \delta w^+ + X^- \sqrt{|\Lambda^-|^{-1}} \delta w^-.$$

Inserting (30) into the above yields

$$\delta q = \left(X^+ \sqrt{(\Lambda^+)^{-1}} \delta R + X^- \sqrt{|\Lambda^-|^{-1}} \right) \delta w^-.$$

Also, inserting (30) into (28) yields

$$w^- = (I^- - R\delta R) \delta w^- + R w^+.$$

If $\det(I^- - R\delta R) \neq 0$, then $\delta w^- = (I^- - R\delta R)^{-1} (w^- - R w^+)$. Inserting this final expression into δq results in the formula (32). $\square$

Proposition 2 shows that δq is formulated in terms of the boundary condition $w^- - R w^+$. Hence, this formulation imposes a minimum number of boundary conditions, as required for well-posedness, and we have proven the following result.

Proposition 3 *The problem (18) subject to the weak boundary conditions (19) is well-posed if (26), (28) and (30) hold, and*

$$R^T R \leq I^+ \quad \text{and} \quad (\delta R)^T \delta R \leq I^-. \quad (33)$$

3.5 Projection matrix formula

There is an alternative procedure to determine δq , obtained via $\delta q = P q$, where P is a projection matrix. By definition, a projection matrix satisfies the idempotency property $P = P P$.

Proposition 4 *Let*

$$L = X^- \sqrt{|\Lambda^-|} - X^+ \sqrt{\Lambda^+} R^T, \quad \delta L = X^- \sqrt{|\Lambda^-|^{-1}} + X^+ \sqrt{(\Lambda^+)^{-1}} \delta R. \quad (34)$$

Then $\delta q = Pq$, where P is the projection matrix

$$P = \delta L(L^T \delta L)^{-1} L^T. \quad (35)$$

Proof We start by deriving the formula for P and then show that this matrix is a projection matrix. Note that δq in (32) can be written as

$$\delta q = \delta L(I^- - R\delta R)^{-1} L^T q.$$

Hence, it remains to show that $L^T \delta L = I^- - R\delta R$. The result immediately follows from eigenvector orthogonality: $(X^+)^T X^- = 0$, $(X^-)^T X^+ = 0$, $(X^+)^T X^+ = I^+$, and $(X^-)^T X^- = I^-$.

For P to be a projection matrix it must satisfy $P = PP$. We have

$$PP = \left(\delta L(L^T \delta L)^{-1} L^T \right) \left(\delta L(L^T \delta L)^{-1} L^T \right) = P.$$

□

By applying Proposition 4, the penalty term (19) can be formulated as

$$f = \mathcal{L}(APq).$$

4 The dual problem

We show how the general formulation presented in the previous section is connected to the dual problem.

4.1 The strong dual problem

Define the functional

$$J(u, v) = \int_0^T \int_{\Omega} u v d\Omega dt.$$

The dual of the primal problem (18) is derived from the condition $J(q^*, f) = J(f^*, q)$, where f^* is the PDE describing the dual problem (to be determined). By inserting (18) into $J(q^*, f)$ and applying integration by parts in time and space, we get

$$\begin{aligned} J(q^*, f) &= J(q^*, \partial_t q + A_x \partial_x q + A_y \partial_y q) \\ &= J(f^*, q) + \int_0^T \Phi(q, q^*) dt + \int_{\Omega} \Psi(q, q^*) d\Omega, \end{aligned} \quad (36)$$

where $\Phi(q, q^*)$ is defined in (20), $\Psi(q, q^*) = [q^T q^*]_0^T$, and

$$f^* = -(\partial_t q^* + A_x \partial_x q^* + A_y \partial_y q^*).$$

The terms Φ and Ψ arising from integration by parts must vanish for the dual problem to be well-defined. The term Ψ vanishes due to the initial condition $u(x, 0) = 0$ and end condition $u^*(x, T) = 0$.

Next, we derive the dual boundary conditions. Following Sect. 3.2, by diagonalizing and splitting A into positive and negative parts, we get

$$\Phi(q, q^*) = \oint_{\partial\Omega} (w^+)^T w^{+*} - (w^-)^T w^{-*} ds. \quad (37)$$

where $w^{+*} = (X^+)^T q^*$, and $w^{-*} = (X^-)^T q^*$. The strong primal boundary condition (26) leads to

$$\Phi(q, q^*) = \oint_{\partial\Omega} (w^+)^T (w^{+*} - R^T w^{-*}) ds.$$

Hence, to make Φ vanish, the dual boundary conditions must be

$$w^{+*} = R^T w^{-*}, \quad (x, y) \in \partial\Omega. \quad (38)$$

By introducing the time reversal transformation $\tau = T - t$, the dual problem becomes

$$\partial_\tau q^* - A_x \partial_x q^* - A_y \partial_y q^* = f^*, \quad (x, y) \in \Omega, \quad \tau \geq 0, \quad (39)$$

subject to the boundary conditions (38).

Proposition 5 *The dual problem (38)–(39) is well-posed if and only if the primal problem (18), (26) is well-posed.*

Proof The energy rate for the dual problem (39) with $f^* = 0$ is

$$\frac{d\|q^*\|^2}{d\tau} = \Phi(q^*, q^*).$$

By imposing the strong dual boundary conditions (38), we get

$$\Phi(q^*, q^*) = - \oint_{\partial\Omega} (w^{-*})^T (I^- - R R^T) w^{-*} ds.$$

Hence, the dual problem is well-posed if $R R^T \leq I^-$. On the other hand, the primal problem is well-posed if $R^T R \leq I^+$ (see Proposition 3). We can show that the two conditions are the same using the singular value decomposition (SVD) $R = U \Sigma V^T$, where U and V are real orthogonal matrices, and Σ is a diagonal matrix. By

applying the SVD to R , we have $RR^T = U\Sigma^2U^T$, $R^TR = V\Sigma^2V^T$. Thus, both the primal and dual problems are well-posed if $\Sigma \leq I$. $\square$

4.2 The weak dual problem

Since the analysis of the weak dual problem is very similar to the weak primal problem, some details are omitted. To weakly impose the boundary conditions (38), we set the right-hand side of (39) to

$$f^* = -\mathcal{L}(A\delta q^*). \quad (40)$$

In (40), δq^* is an unknown penalty vector. This vector is determined in the same manner as the penalty vector δq of the primal problem. When the dual problem (39) uses the penalty term (40), the energy method leads to

$$\frac{d}{d\tau} \|q^*\|^2 = \Phi(q^* - \delta q^*, q^* - \delta q^*) - \Phi(\delta q^*, \delta q^*),$$

To bound the solution, we require

$$\Phi(q^* - \delta q^*, q^* - \delta q^*) \leq 0 \quad \text{and} \quad \Phi(\delta q^*, \delta q^*) \geq 0. \quad (41)$$

First, consider $\Phi(q^* - \delta q^*, q^* - \delta q^*)$ and specify

$$w^{*+} - \delta w^{*+} = R^*(w^{*-} - \delta w^{*-}), \quad (x, y) \in \partial\Omega. \quad (42)$$

By inserting (42) into (37) and requiring $RR^T \leq I^-$, we get

$$\begin{aligned} \Phi(q^* - \delta q^*, q^* - \delta q^*) &= - \oint_{\partial\Omega} (w^{*-} - \delta w^{*-})^T (I^- - RR^T) (w^{*-} - \delta w^{*-}) ds \\ &\leq 0. \end{aligned}$$

Next, consider $\Phi(\delta q^*, \delta q^*)$ and specify

$$\delta w^{*-} = \delta R^* \delta w^{*+}, \quad (x, y) \in \partial\Omega. \quad (43)$$

By inserting (43) into (37) and requiring $(\delta R^*)^T \delta R^* \leq I^+$, we get

$$\Phi(\delta q^*, \delta q^*) = \oint_{\partial\Omega} (\delta w^{*+})^T (I^+ - (\delta R^*)^T \delta R^*) \delta w^{*+} ds \geq 0.$$

The results established thus far are summarized in the following proposition.

Proposition 6 *The problem (39) subject to the weak boundary conditions (40) is well-posed if (38), (42)–(43) uniquely hold, and*

$$RR^T \leq I^- \quad \text{and} \quad (\delta R^*)^T (\delta R^*) \leq I^+. \quad (44)$$

4.3 Dual consistency

To achieve superconvergent linear functionals, it is important that the weak primal and dual problems are related to each other via an appropriate choice of penalty terms, f and f^* [21]. By substituting f and f^* in (36) using the penalty terms (19) and (40), we get

$$\Phi(q^*, \delta q) = -\Phi(q, \delta q^*) + \Phi(q, q^*). \quad (45)$$

To arrive at (45), we have neglected the integrals over t and set the initial and end conditions to zero. As shown in the following proposition, (45) establishes a relationship between the penalty parameters δR and δR^* that is analogous to the relationship between the primal (26) and dual (38) boundary conditions.

Proposition 7 *The weak primal and dual problems (see Proposition 3 and 6) satisfy (45) if*

$$\delta R^* = (\delta R)^T. \quad (46)$$

Proof See Appendix A for proof. $\square$

5 Bounded matrix norms

For spatial discretizations with all eigenvalues located in the correct half plane, the maximum stable time step is determined by the stability region of the particular time stepping scheme. For the spatial discretization to be robust, it must handle any choice of model parameters without causing unwanted growth of the spectral radius, which can be costly. To avoid computing the spectral radius, our goal in this section is to *a priori* analyze how the matrix norm will scale (up to a constant factor) for the spatial discretization of a particular model problem. Since this analysis relies on symbolic computation, it can identify model parameters that cause suboptimal scaling of the matrix norm. If the scaling is the same as the expected CFL-condition obtained from the uncoupled problem, we say that discretization is non-stiff. This definition will be made precise soon. In more detail, we focus on the following objectives:

1. To develop an easy-to-use diagnostic test that indicates how the penalty parameters influence the matrix norm of the spatial discretization.
2. To prove that the projection matrix formula (35) results in penalty terms that are non-stiff, in the sense discussed below.

5.1 Matrix analysis

Before proceeding, we briefly summarize key concepts and results from matrix analysis. For a matrix $A \in \mathbb{R}^{m \times m}$, the 2-norm is defined by the induced vector norm

$$\|A\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{R}^m \\ \mathbf{x} \neq 0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}, \quad (47)$$

where $\|\mathbf{x}\|_2 = (\sum_{i=1}^m x_i^2)^{1/2}$. The Frobenius norm is

$$\|A\|_F = \sqrt{\text{trace}(A^T A)}. \quad (48)$$

The 2-norm and Frobenius norm satisfy the following properties:

$$\varrho(A) \leq \|A\|_2 \leq \|A\|_F. \quad (49)$$

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2, \quad \|AB\|_F \leq \|A\|_F \|B\|_F, \quad (50)$$

$$\|A \otimes B\|_2 = \|A\|_2 \|B\|_2, \quad \|A \otimes B\|_F = \|A\|_F \|B\|_F, \quad (51)$$

where $\otimes$ is the Kronecker product, and $\varrho(A)$ is the spectral radius of A . See e.g., Golub and Van Loan [12] for further properties and proofs of (49)–(51).

5.2 Model problem

As the model problem, we consider the 1D hyperbolic PDE

$$u_t + A_x u_x = \mathcal{L}(\Sigma L^T)u \quad (52)$$

and discretize it in space with SBP operators,

$$\frac{d\mathbf{u}}{dt} = M_h \mathbf{u}, \quad M_h = -A_x \otimes D_x + \mathcal{L}_h \otimes \Sigma L^T. \quad (53)$$

Here, $\mathbf{u} = [\mathbf{u}_0 \ \mathbf{u}_1 \ \dots \ \mathbf{u}_m]^T$, $\mathbf{u}_i = [u_{i0} \ u_{i1} \ \dots \ u_{iN}]$, $u_{ij} = u_i(x_j, t)$, where $x_j = jh$ and h is the grid spacing for an equidistant grid. The matrix D_x is defined by $D_x = H_x^{-1} Q_x$, where $H_x = H_x^T > 0$. The matrix Q_x satisfies the summation-by-parts property:

$$Q_x + Q_x^T = \mathbf{e}_N \mathbf{e}_N^T - \mathbf{e}_0 \mathbf{e}_0^T.$$

This property is key for proving energy stability of a SBP-based scheme. The vectors $\mathbf{e}_N$ and $\mathbf{e}_0$ are typically chosen as $\mathbf{e}_0 = [1 \ 0 \ \dots \ 0]^T$, $\mathbf{e}_N = [0 \ \dots \ 0 \ 1]^T$. Their role is to extract the boundary values u_{i0} and u_{iN} from $\mathbf{u}$. For further details about SBP-based discretizations, see [7,30], and the references therein.

To simplify the forthcoming analysis, we assume that we only need to apply a penalty term on the left boundary, such that $\mathcal{L}_h$ in (53) becomes

$$\mathcal{L}_h = H_x^{-1} \mathbf{e}_0 \mathbf{e}_0^T.$$

Moreover, we assume that D_x is a consistent approximation of the first derivative such that it differentiates a constant exactly: $D_x \mathbf{1} = 0$ for $\mathbf{1} = [1 \ 1 \ \dots \ 1]^T$.

Before we can prove that the discretization (53) is *non-stiff* for appropriately chosen penalty parameters, we give the following definition of non-stiffness.

Definition 2 The semi-discrete approximation (53) of (52) is *non-stiff* if there exists a positive constant γ , independent of the model parameters, such that

$$h\varrho(M_h) \leq \gamma\varrho(A_x).$$

5.3 Diagnostic test

Our objective is to decide how $h\|M_h\|_F$ (or the 2-norm) scales for a particular choice of penalty terms without performing any numerical computations. This capability is useful because it enables one to perform a diagnostic test and assess the robustness of a particular weak coupling procedure without having to implement it and compute the spectral radius for a range of model parameters. If the test passes, then the weak coupling procedure is provably non-stiff. If the test fails, it indicates that the coupling procedure is sensitive to certain choices of model parameters. Our test consists of the following four steps.

1. Determine Σ such that the problem is energy stable and the penalty parameters σ_{ij} are parameterized by the model parameters (e.g., a_{ij} in A_x , and l_{ij} in L).
2. Compute the spectral radius $\varrho(A_x)$
3. Compute the norm of the penalty matrix $\|\Sigma L^T\|_F$.
4. Check if there exists a constant γ , independent of the model parameters, such that: $\|\Sigma L^T\|_F \leq \gamma\varrho(A_x)$.
5. If the test fails, repeat it for the 2-norm.
6. If the test still fails, redesign the coupling procedure to be provably non-stiff.

We recommend checking the Frobenius norm first, because it is much easier to compute than the 2-norm. To compute the 2-norm one must symbolically compute the singular values of the penalty matrix.

The diagnostic test is motivated by the following assumption and Proposition.

Assumption 1 For finite sized (fixed N) matrices, D_x and $\mathcal{L}_h$, there exists constants $s_{d,l}, r_{d,l} > 0$ such that

$$r_d \leq h\|D_x\|_2 \leq s_d, \quad r_l \leq h\|\mathcal{L}_h\|_2 \leq s_l. \quad (54)$$

Proposition 8 The matrix M of the spatial discretization (53) satisfies

$$r_l\|\Sigma L^T\|_2 \leq h\|M_h\|_2 \leq s_d\varrho(A_x) + s_l\|\Sigma L^T\|_2. \quad (55)$$

Proof See Appendix B for proof. $\square$

Remark 1 If there are multiple penalty terms present in M_h , we use the reverse triangle inequality on each term to obtain a lower bound, i.e.,

$$\|A + B\| \geq \|A\| - \|B\|.$$

Likewise, we use the triangle inequality to obtain an upper bound.

5.3.1 Suboptimal scaling

As an illustration, consider the motivating example presented in Sect. 2.

1. From (6), the penalty matrix on each side of the interface is

$$\Sigma^{(i)} L^T = \begin{bmatrix} 0 & c_i k_i n_x & c_i k_i n_y & 0 & c k_i n_x & c k_i n_y \\ \frac{Z_1}{\rho_i} l_i n_x & 0 & 0 & -\frac{Z_2}{\rho_i} l_i n_x & 0 & 0 \\ \frac{Z_1}{\rho_i} l_i n_y & 0 & 0 & -\frac{Z_2}{\rho_i} l_i n_y & 0 & 0 \end{bmatrix}.$$

2. The spectral radius is $\varrho(A_x^{(i)}) = c_i$, and $\max \varrho(A_x^{(i)}) \leq c_{\max}$, where $c_{\max} = \max(c_1, c_2)$.
3. The Frobenius norm of each penalty matrix is

$$\begin{aligned} \|\Sigma^{(1)} L^T\|_F &= \frac{1}{\rho_1} \sqrt{c_1^2 \rho_1^2 (2k_1^2 + l_1^2) + c_2^2 \rho_2^2 l_1^2}, \\ \|\Sigma^{(2)} L^T\|_F &= \frac{1}{\rho_2} \sqrt{c_2^2 \rho_2^2 (2k_2^2 + l_2^2) + c_1^2 \rho_1^2 l_2^2}. \end{aligned}$$

4. For the naive choice of parameters (15), we get

$$\|\Sigma^{(1)} L^T\|_F \leq \frac{c_{\max}}{2} \sqrt{3 + \frac{\rho_2^2}{\rho_1^2}} \quad \text{and} \quad \|\Sigma^{(2)} L^T\|_F \leq \frac{c_{\max}}{2} \sqrt{3 + \frac{\rho_1^2}{\rho_2^2}}.$$

The presence of ρ_2/ρ_1 causes linear growth in $h\|M\|_F$ for $\rho_2/\rho_1 \ll 1$ and $\rho_2/\rho_1 \gg 1$.

5. Since we have failed to bound the penalty matrices in the Frobenius norm, we proceed to check the 2-norm:

$$\begin{aligned} \|\Sigma^{(1)} L^T\|_2 &= \max \left(\frac{c_1}{2\sqrt{2}}, \frac{\sqrt{c_1^2 \rho_1^2 + c_2^2 \rho_2^2}}{2\rho_1} \right) \quad \text{and} \\ \|\Sigma^{(2)} L^T\|_2 &= \max \left(\frac{c_2}{2\sqrt{2}}, \frac{\sqrt{c_1^2 \rho_1^2 + c_2^2 \rho_2^2}}{2\rho_2} \right). \end{aligned}$$

The presence of ρ_1/ρ_2 also causes linear growth in $h\|M\|_2$. Hence, we cannot rule out that the spectral radius scales suboptimally as a function of the model parameters. In fact, the numerical experiment presented in Fig. 2 shows unwanted growth of the spectral radius in ρ_1/ρ_2 .

6. Since we have failed to prove that this coupling is non-stiff, we will redesign it using the projection matrix formula.

5.4 Provably non-stiff penalty treatment

As is shown in the following proposition, the projection matrix formula yields a useful estimate for well-posed linear hyperbolic problems. For certain choices of the penalty parameters δR , the method is provably non-stiff.

Proposition 9 *Let $\Sigma L^T = AP$, where A is given in (19) and P is computed using the projection matrix formula (35). Then*

$$\|\Sigma L^T\|_2 \leq \frac{(1 + \|R\|_2)(1 + \|\delta R\|_2)}{\|I^- - R\delta R\|_2} \rho(A). \quad (56)$$

Proof See Appendix C for proof. $\square$

If the denominator $\|I^- - R\delta R\|_2$ approaches zero, the norm of $h\|M\|_2$ scales suboptimally. There are at least two choices of δR that prevent the denominator from vanishing for all R . For the energy dissipative choice $\delta R = 0$, (56) yields

$$\|\Sigma L^T\|_2 \leq (1 + \|R\|_2)\rho(A) \leq 2\rho(A).$$

For the energy conservative choice $\delta R = -R^T$, (56) yields

$$\|\Sigma L^T\|_2 \leq \frac{(1 + \|R\|_2)^2}{\|I^- + RR^T\|_2} \rho(A) \leq 4\rho(A).$$

In each case, we used $\|R\|_2 \leq 1$, which is required by well-posedness.

6 Revisiting the motivating example

To demonstrate how to derive dual consistent and non-stiff penalty terms by example, we apply it to the motivating example presented in Sect. 2.

We summarize our procedure and carry out each step at a time.

1. (The penalty formulation) Formulate the penalty terms such that the energy rate can be written similarly to (21). Since we are studying the coupled problem (1), we need to scale the energy rate as done in (9),

$$\alpha \frac{d}{dt} \|q^{(1)}\|^2 + \beta \frac{d}{dt} \|q^{(2)}\|^2 = -\Phi(u, u) + 2\Phi(u, \delta u), \quad \Phi(u, v) = \oint_{\Gamma} u^T A v ds,$$

where $u = (q^{(1)}, q^{(2)})$, $A = \text{diag}(\alpha \tilde{A}^{(1)}, \beta \tilde{A}^{(2)})$, and δu is to be determined.

2. (Diagonalization) Diagonalize A and construct w^+ and w^- using (23).
3. (Coupling conditions) Rewrite the coupling conditions (4) in the form of $w^- = R w^+$ subject to the constraint $R^T R \leq I^+$ (see Proposition 3).
4. (Penalty parameters) Choose some δR subject to the constraint $(\delta R)^T (\delta R) \leq I^-$ (see Proposition 3).

5. (Projection matrix formula) Determine $\delta u = Pu$ using the projection matrix formula defined in Proposition 4.
6. (Matrix norms) Following Sect. 5, compute the matrix norms $\|AP/\alpha\|_F$ and $\|AP/\beta\|_F$ to verify that the formulation is provably non-stiff.

1. We formulate the penalty terms for the respective sides of the interface as

$$f^{(1)} = \mathcal{L} \left(\frac{1}{\alpha} [I \ 0] A \delta u \right) \quad \text{and} \quad f^{(2)} = \mathcal{L} \left(\frac{1}{\beta} [0 \ I] A \delta u \right), \quad (57)$$

This formulation of the penalty terms leads to

$$\int_{\Omega^{(1)}} \alpha (q^{(1)})^T f^{(1)} d\Omega + \int_{\Omega^{(2)}} \beta (q^{(2)})^T f^{(2)} d\Omega = \int_{\Gamma} u^T A \delta u ds.$$

We use $\alpha = \rho_1$ and $\beta = \rho_2$, as these values are necessary for obtaining an energy balance when imposing the coupling conditions (4).

2. Next, following Sect. 3.2, we diagonalize A to obtain

$$\int_{\Gamma} u^T A u ds = \int_{\Gamma} (w^+)^T w^+ - (w^-)^T (w^-) ds. \quad (58)$$

In (58), $w^+ = (w^{+(1)}, w^{+(2)})$, $w^- = (w^{-(1)}, w^{-(2)})$, where

$$w^{\pm(1)} = \frac{\sqrt{Z_1}}{\sqrt{2}} \left(\frac{p^{(1)}}{Z_1} \pm v_n^{(1)} \right), \quad w^{\pm(2)} = \frac{\sqrt{Z_2}}{\sqrt{2}} \left(\frac{p^{(2)}}{Z_2} \mp v_n^{(2)} \right),$$

$Z_1 = \rho_1 c_1$, and $Z_2 = \rho_2 c_2$.

3. The coupling conditions (4) need to be converted into the form $w^- = R w^+$. After some algebra, one obtains the solution

$$R(Z_1, Z_2) = \frac{1}{Z_1 + Z_2} \begin{bmatrix} Z_2 - Z_1 & 2\sqrt{Z_1 Z_2} \\ 2\sqrt{Z_1 Z_2} & Z_1 - Z_2 \end{bmatrix}. \quad (59)$$

We verify that the matrix (59) strictly satisfy (29). A direction calculation yields $R^T R = I_2$, where I_2 is the 2×2 identity matrix. Hence, R is an orthogonal matrix.

4. Next, we choose $\delta R = -R^T$. This choice results in energy conservation by (29) and (31), since $R^T R = I_2$ and therefore $\delta R (\delta R)^T = I_2$.

5. We use the projection matrix formula in Proposition 4 for $\delta R = -R^T$, to obtain the penalty vector

$$\delta u = Pu = \frac{1}{Z_1 + Z_2} \begin{bmatrix} p^{(1)} - p^{(2)} \\ Z_2 n_x \left(v_n^{(1)} - v_n^{(2)} \right) \\ Z_2 n_y \left(v_n^{(1)} - v_n^{(2)} \right) \\ p^{(2)} - p^{(1)} \\ Z_1 n_x \left(v_n^{(2)} - v_n^{(1)} \right) \\ Z_1 n_y \left(v_n^{(2)} - v_n^{(1)} \right) \end{bmatrix}. \quad (60)$$

After inserting (60) into (57), the final penalty terms become

$$f_{ec}^{(i)} = \mathcal{L} \left(\frac{1}{\rho_i} A P u \right) = \mathcal{L} \left(\frac{1}{\rho_i (Z_1 + Z_2)} I^{(i)} \begin{bmatrix} Z_1 Z_2 \left(v_n^{(1)} - v_n^{(2)} \right) \\ Z_1 n_x \left(p^{(1)} - p^{(2)} \right) \\ Z_1 n_y \left(p^{(1)} - p^{(2)} \right) \\ Z_1 Z_2 \left(v_n^{(1)} - v_n^{(2)} \right) \\ Z_2 n_x \left(p^{(1)} - p^{(2)} \right) \\ Z_2 n_y \left(p^{(1)} - p^{(2)} \right) \end{bmatrix} \right), \quad (61)$$

where $I^{(1)} = [I \ 0]$, $I^{(2)} = [0 \ I]$.

6. Following the steps laid out in the diagnostic test in Sect. 5, we arrive at

$$\begin{aligned} \left\| \frac{1}{\rho_1} A P \right\|_F &\leq \frac{c_{max}}{2} \sqrt{\frac{\rho_1^2 + 3\rho_2^2}{(\rho_1 + \rho_2)^2}} \leq \frac{\sqrt{3}}{2} c_{max}, \quad \left\| \frac{1}{\rho_2} A P \right\|_F \\ &\leq \frac{c_{max}}{2} \sqrt{\frac{3\rho_1^2 + \rho_2^2}{(\rho_1 + \rho_2)^2}} \leq \frac{\sqrt{3}}{2} c_{max}, \end{aligned}$$

since $\rho_1 > 0$, $\rho_2 > 0$. Hence, this formulation is provably non-stiff in the sense of Definition 2.

Alternately, if we repeat steps 1-5 for the energy dissipative choice $\delta R = 0$, the penalty terms become

$$f_{ed}^{(i)} = \mathcal{L} \left(\frac{1}{\rho_i (Z_1 + Z_2)} I^{(i)} \begin{bmatrix} Z_1 \left(Z_2 (v_n^{(1)} - v_n^{(2)}) - (p^{(1)} - p^{(2)}) \right) \\ Z_1 n_x \left(-Z_2 (v_n^{(1)} - v_n^{(2)}) + (p^{(1)} - p^{(2)}) \right) \\ Z_1 n_y \left(-Z_2 (v_n^{(1)} - v_n^{(2)}) + (p^{(1)} - p^{(2)}) \right) \\ Z_2 \left(Z_1 (v_n^{(2)} - v_n^{(1)}) + (p^{(2)} - p^{(1)}) \right) \\ Z_2 n_x \left(Z_1 (v_n^{(2)} - v_n^{(1)}) + (p^{(2)} - p^{(1)}) \right) \\ Z_2 n_y \left(Z_1 (v_n^{(2)} - v_n^{(1)}) + (p^{(2)} - p^{(1)}) \right) \end{bmatrix} \right). \quad (62)$$

6.1 General penalty term formulation

By inspecting (61)–(62) we can formulate the following general penalty terms.

$$f^{(i)} = \mathcal{L} \left(\frac{1}{\rho_i} I^{(i)} \begin{bmatrix} Z_1 k_1 (v_n^{(1)} - v_n^{(2)}) \\ n_x k_2 (p^{(1)} - p^{(2)}) \\ n_y k_2 (p^{(1)} - p^{(2)}) \\ Z_2 k_2 (v_n^{(1)} - v_n^{(2)}) \\ n_x k_1 (p^{(1)} - p^{(2)}) \\ n_y k_1 (p^{(1)} - p^{(2)}) \end{bmatrix} + \frac{\zeta}{\rho_i} I^{(i)} \begin{bmatrix} -k_2 (p^{(1)} - p^{(2)}) \\ -n_x k_1 Z_1 (v_n^{(1)} - v_n^{(2)}) \\ -n_y k_1 Z_1 (v_n^{(1)} - v_n^{(2)}) \\ k_1 (p^{(2)} - p^{(1)}) \\ n_x k_2 Z_2 (v_n^{(2)} - v_n^{(1)}) \\ n_y k_2 Z_2 (v_n^{(2)} - v_n^{(1)}) \end{bmatrix} \right), \quad \zeta \geq 0. \quad (63)$$

In (63), ζ controls the amount of energy dissipation, and k_1, k_2 are listed in Proposition 1, and they are here defined as

$$k_1 = \frac{Z_2^m}{Z_1^m + Z_2^m} \quad \text{and} \quad k_2 = \frac{Z_1^m}{Z_1^m + Z_2^m}, \quad (64)$$

where m is a free parameter. This formulation recovers all other formulations.

- The naive choice (15) corresponds to

$$m = 0 \quad \text{and} \quad \zeta = 0. \quad (65)$$

- The energy conservative choice, $\delta R = -R^T$, corresponds to

$$m = 1 \quad \text{and} \quad \zeta = 0. \quad (66)$$

- The energy dissipative choice, $\delta R = 0$, corresponds to

$$m = 1 \quad \text{and} \quad \zeta = 1. \quad (67)$$

To illustrate the results, we revisit the spectral radius investigation presented in Sect. 2.4. The energy conservative and dissipative choices (66)–(67) each cause the spectral radius to be independent of the density ratio, as shown in Fig. 3.

7 Numerical experiments

In this section, we conduct several numerical experiments. These experiments demonstrate the importance of having a scheme with bounded matrix norm for robustness, and show the predictive ability of Propositions 8–9. In addition, we show that linear functionals are superconverging. Finally, we investigate advantages and disadvantages of developing a near-energy conservative scheme (in both space and time) of the wave equation in two spatial dimensions.

Fig. 3 The normalized spectral radius $h_Q(M_h)/c$ as a function of the density ratio ρ_1/ρ_2 (on a log-log scale)

7.1 Matrix norm behavior

Consider the wave equation in a single domain on the unit interval $0 \leq x \leq 1$, subject to the boundary conditions

$$w^-(0, t) = 0, \quad (68)$$

$$p(1, t) = \alpha Z v(1, t), \alpha > 0. \quad (69)$$

The left boundary condition (68) is implemented by taking $R = 0$ and $\delta R = 0$ in the projection matrix formula (Proposition 4). By Proposition 3, the problem is well-posed and energy dissipative. We consider two different well-posed implementations of the right boundary condition, described next.

7.1.1 Naive choice

The right boundary condition can be implemented naively by applying the penalty term

$$f_{x=1} = \mathcal{L}(\Sigma(p - \alpha Z v)),$$

where $\int u^T \mathcal{L}(v) dx = u(1, t)v(1, t)$. The penalty matrix is chosen as $\Sigma = [0, 1/\rho]^T$, which results in the energy dissipation rate

$$\rho \frac{\|q\|^2}{dt} = -\alpha Z v^2(1, t) \leq 0.$$

Fig. 4 Behavior of $\|M_h\|_2$ as a function of the boundary parameter α in (69). Lower bounds (LB) and upper bounds (UB) are given by Proposition 8

(The energy rate contribution from the left boundary has been neglected.) We write the penalty term as $f_{x=1} = \mathcal{L}(\Sigma L^T q)$, where

$$\Sigma L^T = \begin{bmatrix} 0 & 0 \\ c & -\alpha c \end{bmatrix}.$$

By Proposition 8, this penalty term causes supoptimal scaling, since:

$$\|\Sigma L^T\|_F = c\sqrt{1 + \alpha^2}.$$

Hence, for $\alpha \gg 1$ the growth rate of $h\|M_h\|$ is linear in α . We confirm this prediction by numerically computing $\|M_h\|_2$ as well as lower and upper bounds on it by applying Proposition 8. We discretize in space by applying the fourth order SBP operators [28] using 100 grid points. The constants in Proposition 8 are numerically computed to be:

$$s_d = h\|D_x\|_2 \approx 2.359 \quad r_l = s_l = h\|\mathcal{L}^h\|_2 \approx 2.824.$$

Figure 4 show that the norm $h\|M_h\|_2$ grows linearly as predicted when α increases.

7.1.2 Projection matrix formula

The right boundary condition (69) can also be implemented by applying the projection matrix formula (Proposition 4). In this case, $R = (\alpha - 1)/(1 + \alpha)$, whereas δR is a free, but bounded parameter. We investigate how δR influences $h\|M_h\|_2$ for a fixed $\alpha = 10$ and vary δR within its stability limit $-1 \leq \delta R \leq 1$. Figure 5 shows that any

Fig. 5 Behavior of $\|M_h\|_2$ as a function of the penalty parameter δR . Lower bounds (LB) and upper bounds (UB) are given by Proposition 8

$\delta R \leq 0$ has negligible impact on $h\|M_h\|_2$. On the other hand, when $\delta R \rightarrow 1$, $h\|M_h\|_2$ grows rapidly. Proposition 9 explains this growth because $\|I^- - R\delta R\| \ll 1$ when $\delta R \rightarrow 1$.

7.2 Superconvergence and dual consistency

SBP-SAT schemes that are dual consistent exhibit superconvergence for linear functionals. As explained in Sect. 4, a necessary condition for dual consistency is that the weak primal boundary conditions are related to the weak dual boundary conditions in a certain way. Interestingly, Proposition 7, shows that all penalty terms constructed using the projection matrix formula satisfy this condition. We numerically verify this result by constructing both energy conservative and energy dissipative penalty treatments that result in superconvergent linear functionals.

Consider again the wave equation in a single domain $0 \leq x \leq 1$. We use the method of manufactured solutions to construct the solution

$$p(x, t) = \cos(kt) \sin(kx), \quad v(x, t) = -\sin(kt) \cos(kx). \quad (70)$$

The manufactured solution (70) satisfies the wave equation subject to the initial and boundary conditions:

$$\begin{aligned} p(x, 0) &= \sin(kx), & v(x, 0) &= 0, \\ p(0, t) &= 0, & v(1, t) &= -\sin(kt) \cos(k). \end{aligned} \quad (71)$$

Table 1 Errors and convergence rates when using an energy dissipative penalty parameter value $\delta R = 0$

N	p		v		$F(p)$		$F(v)$	
	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate
20	-0.05193	–	-0.08724	–	-1.681	–	-1.51	–
39	-1.138	3.608	-1.013	3.076	-2.225	1.806	-2.103	1.97
77	-2.305	3.877	-2.122	3.683	-3.468	4.131	-3.305	3.993
153	-3.413	3.679	-3.13	3.348	-4.691	4.064	-4.515	4.021
305	-4.448	3.439	-4.063	3.102	-5.907	4.039	-5.723	4.012
609	-5.431	3.265	-4.971	3.014	-7.101	3.966	-6.932	4.017

Table 2 Errors and convergence rates when using the energy conservative penalty value $\delta R = 1$

N	p		v		$F(p)$		$F(v)$	
	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate	$\log_{10}$ error	Rate
20	-0.1154	–	-0.07907	–	-3.098	–	-1.426	–
39	-1.129	3.368	-0.9517	2.899	-2.273	-2.741	-2.109	2.271
77	-2.264	3.771	-1.891	3.12	-3.489	4.041	-3.291	3.926
153	-3.133	2.887	-2.856	3.206	-4.645	3.838	-4.505	4.032
305	-4.012	2.918	-3.801	3.138	-5.916	4.221	-5.743	4.114
609	-4.912	2.99	-4.723	3.063	-7.054	3.783	-6.946	3.995

These boundary conditions correspond to $R = -1$ at both $x = 0$ and $x = 1$. The manufactured solution (70) results for example in the linear functionals

$$\begin{aligned}
 F(p) &= \int_0^1 p dx = \frac{\cos(k) - 1}{k} \cos(kt), \\
 F(v) &= \int_0^1 v dx = -\frac{\sin(k)}{k} \sin(kt).
 \end{aligned} \tag{72}$$

As before, the boundary conditions are implemented using the projection matrix formula. We compare the energy dissipative choice $\delta R = 0$ against the energy conservative choice $\delta R = -1$. Besides this parameter choice, all other settings are the same. In (70), we use $k = 8\pi$. We advance in time using a fourth order Runge–Kutta method until the final time $T = 1.2$ using the timestep $\Delta t = h/4$ (the wave speed c is set to $c = 1$ in this case). This experiment uses fourth order SBP operators, and the expected convergence rate is third order in the variables p , and v according to [29,31]. The linear functionals (72) should superconverge at fourth order rate [14]. Tables 1 and 2 lists the errors and convergence rates for the energy dissipative and energy conservative penalty parameter choices, respectively. As expected, both choices result in superconverging functionals.

7.3 Energy conservation versus energy dissipation

For energy conservative problems, it may be beneficial to design a numerical scheme that conserves energy in a semi-discrete sense. Weak boundary conditions that have dissipation ($\gamma > 0$ in (63)) cause the spectrum of the spatial discretization to have eigenvalues with a real part. The amount of dissipation controls the distance of the real part of the eigenvalues to the imaginary axis. Since some Runge-Kutta methods have a stability region that is wider along the imaginary axis compared to the real axis, the presence of dissipation can have a negative impact on the maximum stable time step. Staggered time-stepping methods, such as leap-frog and the fourth order staggered Runge-Kutta method [11] have stability regions that are primarily confined to the imaginary axis. On the other hand, they have smaller truncation error and larger stability region along the imaginary axis compared to their non-staggered counterparts. For instance, the fourth order staggered Runge-Kutta method has about a factor of two larger stability region along the imaginary axis, and a factor of 16 smaller truncation error, compared to the classical fourth order Runge-Kutta scheme. However, to allow for dissipation may improve the accuracy of a numerical scheme. To understand how these benefits and drawbacks influence performance, we compare the computational efficiency of energy conserving discretizations advanced in time with the fourth order staggered Runge-Kutta (SRK4) scheme against energy dissipating discretizations advanced in time using the fourth order, 5-4 solution 3, 2N low-storage Runge-Kutta method (RK4) [15].

To investigate the performance and demonstrate the importance of how the coupling parameters are chosen, we present a challenging application problem featuring a light-dense media interface. The interface is located at $y = 0$ and the light medium rests on top of the dense medium. The material properties in the light medium are $\rho_1 = 1$, $c_1 = 1$, and they are $\rho_2 = 800$, $c_2 = 4$ in the dense medium. These parameter values result in an impedance contrast of $Z_2/Z_1 = 3200$, which is representative of a water-air interface. This large impedance contrast makes the problem computationally challenging. Figure 6 shows a coarse grid simulation of the pressure field for a fixed time. An explosive source initiated in the air medium sends out waves that both reflect against and transmit across the air-water interface (black horizontal line).

The air and water media are each discretized by the fourth order SBP staggered finite difference operators presented in [23]. Initially, all fields are set to zero. To initiate the simulation, we use a singular source term with a prescribed source time function. This source term is written as $\delta(x - x_s)\delta(y - y_s)g(t)$ and acts in the right-hand side of the pressure equation. The source is positioned in close vicinity of the interface, at $(x_s, y_s) = (0.0, 0.05)$. The delta functions $\delta(x - x_s)$ and $\delta(y - y_s)$ are discretized in a line by line manner to fourth order accuracy by imposing moment conditions, see [25] for details. As the source time function, $g(t)$, we use the Ricker wavelet

$$g(t) = \left(1 - 2\pi^2 f_p^2 (t - t_0)^2\right) e^{-\pi^2 f_p^2 (t - t_0)^2},$$

where $f_p = 6.4$ and $t_0 = 0.25$. These settings result in $\lambda_{\min}/h \approx 6$ grid points per minimum wavelength in the air medium. This estimate defines $\lambda_{\min} = c_{\min}/f_{\max}$,

Fig. 6 Pressure wave field at $t = 0.4$, computed on a coarse grid ($\lambda_{min}/h \approx 6$) with the energy dissipative coupling (67)

where f_{max} is the frequency at 5% of the peak amplitude of the Ricker wavelet in the Fourier domain.

Since we are mostly interested in understanding how the coupling treatment influences accuracy at the interface, we measure the error along it. Due to the discontinuous nature of the weak coupling conditions, the error is defined by the jump in a quantity that is continuous across the interface. We choose the jump in pressure. To approximate the solution anywhere along the interface, we use cubic Lagrange interpolation. In particular, we focus on measuring the error in the middle of the interface, at $(x_r, y_r) = (0.0, 0.0)$.

7.3.1 Accuracy

Before investigating the computational efficiency of each method, we compare the accuracy of the energy conservative and energy dissipative couplings obtained by the projection matrix formula (35). When solving the air-water interface problem on the coarse grid with the energy conservative coupling (66), numerical artifacts emerge. These artifacts are large-amplitude spurious oscillations that propagate along the interface. Figure 7a shows the presence of these oscillations at the same time and on the same grid as shown in Fig. 6. While these spurious oscillations vanish with grid refinement and appear to be confined to the interface, they prevent one from performing coarse grid computations that demand a reasonably (less than 10% error) accurate solution at the interface. Fortunately, we can suppress these spurious oscillations by selecting a different δR in the projection matrix formula.

To appreciate what causes the spurious oscillations to develop, recall that the interface problem couples a light medium to a dense medium. This problem has a large impedance contrast $Z_2/Z_1 \gg 1$ across the interface. In this limit, we study the reflection and transmission of plane wave solutions that are incident normal to the interface.

Fig. 7 Qualitative accuracy comparison of energy conserving couplings. The energy conservative coupling (66) causes noticeable spurious oscillations on coarse grids

The following solution solves (1),

$$\begin{aligned} p^{(1)} &= -Z_1 e^{i\tilde{k}_1(y+c_1t)} + a_R Z_1 e^{i\tilde{k}_1(-y+c_1t)}, \\ v_x^{(1)} &= 0, \quad v_y^{(1)} = e^{i\tilde{k}_1(y+c_1t)} + a_R e^{i\tilde{k}_1(-y+c_1t)} \\ p^{(2)} &= -Z_2 a_T e^{i\tilde{k}_2(y+c_2t)}, \quad v_x^{(2)} = 0, \quad v_y^{(2)} = a_T e^{i\tilde{k}_2(y+c_2t)}, \end{aligned}$$

where $\tilde{k}_1, \tilde{k}_2$ are wavenumbers that satisfy the dispersion relation $\omega = \tilde{k}_1 c_1 = \tilde{k}_2 c_2$, and a_R, a_T are reflection and transmission coefficients, determined by the coupling conditions (4). The reflection and transmission coefficients are:

$$a_R = \frac{Z_1 - Z_2}{Z_1 + Z_2} \quad \text{and} \quad a_T = \frac{2Z_1}{Z_1 + Z_2}.$$

In the limit when $Z_2/Z_1 \rightarrow \infty$, we get $a_R = -1, a_T = 0$, and there is no transmission into the dense medium. An interpretation of this limit is that a wave propagating in the light medium senses the interface as a rigid wall. On the other hand, a wave propagating in the dense medium senses the interface as a free surface, and transmits into the light medium with double the amplitude. If we neglect any transmission, the problem decouples into two problems with the boundary conditions

$$v_n^{(1)} = 0 \quad (\text{light}), \quad p^{(2)} = 0 \quad (\text{dense}), \quad (x, y) \in \Gamma.$$

The only way to implement these boundary conditions in an energy conserving manner is to set

$$m \rightarrow \infty \quad \text{and} \quad \zeta = 0, \quad (73)$$

in (63)–(64). The choice (73) results in $k_1 = 1$, $k_2 = 0$, and (63) yields

$$f^{(1)} = \mathcal{L} \left(\frac{1}{\rho_1} \begin{bmatrix} Z_1 (v_n^{(1)} - v_n^{(2)}) \\ 0 \\ 0 \end{bmatrix} \right), \quad f^{(2)} = \mathcal{L} \left(\frac{1}{\rho_2} \begin{bmatrix} 0 \\ n_x (p^{(1)} - p^{(2)}) \\ n_y (p^{(1)} - p^{(2)}) \end{bmatrix} \right), \quad (74)$$

where $v_n^{(2)} = p^{(1)} = 0$ in the decoupled case.

In the limit when $Z_2/Z_1 \rightarrow \infty$, the energy conservative coupling (66) approaches the penalty terms of (74) at a linear rate. We speculate that this rate is too slow when there is no artificial dissipation present to suppress the spurious oscillations. Instead, we choose δR so that it corresponds to (74) by directly taking the limit (we do not modify Z_2 , Z_1 in the problem itself)

$$\delta R = \lim_{Z_2/Z_1 \rightarrow \infty} -R^T(Z_1, Z_2) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (75)$$

Obviously, if $Z_1 > Z_2$ we would have taken the opposite limit $Z_2/Z_1 \rightarrow 0$ instead. Note that the choice (75) preserves the energy conservation property because δR is still an orthogonal matrix. Since we used the projection matrix formula to derive this coupling, it is also both dual consistent and provably non-stiff.

This simple modification of the penalty parameter results in a dramatic improvement in accuracy of the coarse grid simulation (Fig. 7b). However, the dissipative choice $\delta R = 0$ in Fig. 6 is the most accurate option.

Figure 8 shows the error in pressure at the center grid point of the interface as a function of time for each type of coupling. The modified energy conservative coupling (73), SRK4-EC2, reduces the maximum amplitude of the error by a factor of 18 more than the coupling (66), SRK4-EC1. While this simple modification resulted in a dramatic improvement for this particular problem, we have not investigated other problems with moderate impedance contrasts.

7.3.2 Time step selection

We determine the maximum stable time step $\Delta t_{max} = CFLh/c_{max}$ for each type of coupling and time stepping method. In each case, we maximize CFL by performing several coarse grid computations using the bisection method. Table 3 lists CFL numbers for each method up to three decimal places.

As expected from the 1D investigations, the naive coupling, RK4-NA, results in more than an order magnitude reduction in ν compared to the other couplings. Interestingly, SRK4-EC1 and SRK4-EC2 result in a factor of four larger CFL compared

(a) $\lambda/h \approx 6$. (b) $\lambda/h \approx 12$.**Fig. 8** Error in pressure in the middle of the interface. See Table 3 for label descriptions

to RK4-ED. We believe this improvement is due to two reasons: 1. SRK4 has a factor of ≈ 2 larger stability region along the imaginary axis compared to RK4. 2. The dissipative part of (67) leads to eigenvalues with real parts that are ≈ 2 smaller than the RK4 stability envelope along the imaginary axis. By decreasing ζ in (63) it should be possible to increase CFL, but most likely at the expense of decreasing the accuracy in the solution.

Table 3 CFL numbers for each type of time stepping method and coupling used in the investigation

Label	Time stepping method	Coupling	Description	CFL
RK4-NA	Runge–Kutta 4	See Eq. (65)	Energy conservative ($m = 0, \zeta = 0$)	0.068
SRK4-EC1	Staggered Runge–Kutta 4	See Eq. (66)	Energy conservative ($m = 1, \zeta = 0$)	1.548
SRK4-EC2	Staggered Runge–Kutta 4	See Eq. (73)	Energy conservative ($m \rightarrow \infty, \zeta = 0$)	1.548
RK4-ED	Runge–Kutta 4	See Eq. (67)	Energy dissipative ($m = 1, \zeta = 1$)	0.412

7.3.3 Computational efficiency

With each method tuned to run at its time step stability limit, we compare their computational efficiency (error as a function of computational time). The norm of the error is the maximum amplitude of the error in time in the middle of the interface (see Fig. 8). Figure 9 shows the computational efficiency of each method. The naive method, RK4-NA, performs the worst, running more than an order of magnitude slower than the other methods. The computational efficiency of SRK4-EC is limited by the large errors caused by the inaccurate coupling. On coarser grids, SRK4-EC2 performs significantly faster than RK4, but the trend reverses on sufficiently fine grids. RK4-ED is the most accurate method on all grids. While not shown, we also tested RK4-EC2 and that resulted in an almost indistinguishable difference in error compared to SRK4-EC2. To model the speedup of SRK4-EC2 compared to RK4-ED, we apply a cubic polynomial least square fit to each data series. Using this model, we observe that the maximum performance increase is about 150 % in favor of SRK4-EC2 for 10% error. Due to the improved accuracy of the dissipative coupling, this performance increase rapidly diminishes. At an error level of 0.01% the performance increase is completely saturated. In conclusion, the energy conservative method outperforms the energy dissipative method on coarse grids due to better stability properties. On the other hand, the energy dissipative method outperforms the energy conservative method on fine grids due to better accuracy properties (see Fig. 10).

To determine computational timings, we implemented each method in a similar manner using C++ and CUDA. The computational timing experiments were measured on a Nvidia GTX 2080 Ti card. No efforts were made to optimize the performance of any of the implementations.

8 Conclusions

When weakly coupling hyperbolic partial differential equations, one must select certain penalty parameters that describe how the coupling conditions are weighted. Within the SBP community, it is well-established how to constrain these parameters to give a proof of semi-discrete stability of a numerical scheme via the energy method. However,

Fig. 9 Computational efficiency of the SRK4 and RK4 schemes

Fig. 10 Modeled speedup of SRK4-EC2 compared to RK4-ED (speedup > 1 SRK4-EC2 is faster than RK4-ED)

the energy method alone cannot fully constrain all of the parameters. The remaining parameters must be carefully selected because they can have a striking impact on stiffness and accuracy of the numerical scheme.

We showed the importance of this claim by simulating the interaction of waves across an air-water interface in an energy conserving manner. For one set of parameter values, the coupling treatment caused the numerical scheme to be an order of

magnitude stiffer than expected from the CFL condition from the uncoupled case. Another set of parameter values prevented stiffness, but instead caused a degradation in accuracy. This accuracy degradation showed the development of spurious waves in the vicinity of the interface. In this study, we explained what causes stiffness and developed a general coupling procedure that is provably non-stiff and accurate.

To overcome stiffness in an automated manner, we presented a general formulation for designing penalty terms. This general formulation results in stable and dual consistent schemes that are provably non-stiff. In this formulation, the penalty terms are related to projection matrices that are guaranteed to exist as long as a determinant condition is satisfied. In the limit when the determinant approaches zero, the matrix norm grows without bounds. We gave two examples for which the determinant never vanishes; these examples are associated with either an energy conservative or an energy dissipative coupling.

A potential advantage of the energy conservative coupling compared to the energy dissipative coupling is that it is compatible with energy conservative time stepping schemes. For this reason, we compared the computational efficiency of a fourth order staggered Runge–Kutta with the energy conservative coupling against a fourth order Runge–Kutta with the energy dissipative coupling. We showed that energy conserving methods can outperform energy dissipative methods on coarse grids because they allow larger time steps. However, energy dissipative methods outperform energy conservative methods on fine grids because they have better accuracy properties. For the air–water interface problem, we observed a factor of 2.5 speedup of the energy conservative method compared to the energy dissipative method on a coarse grid.

There is one important aspect that we did not address: what mechanism of the penalty parameters controls accuracy? For the energy-conservative penalty parameters for the air–water interface problem, we saw that a simple change in a parameter value resulted in an order of magnitude reduction in the error level on the coarsest grid. We linked this accuracy improvement to what the coupling treatment must become in the limit that the impedance contrast approaches infinity. Ideally, we would like to have a simple diagnostic test, analogous to the stiffness test, that can establish if a set of parameters will cause an accurate interface treatment or not. It would also be useful to have an automated and general procedure that can select these parameters, thereby avoiding the cumbersome work of manually deriving and testing parameters on a case by case basis.

Funding Open access funding provided by Linköping University. Ossian O'Reilly was supported by the Southern California Earthquake Center (Contribution No. 10148), with contributions from National Science Foundation Cooperative Agreements EAR-1600087, United States Geological Survey Cooperative Agreement G17AC00047, NSF-ACI Award 1450451, and Pacific Gas and Electric. Jan Nordström was supported by Vetenskapsrådet (Award Number 2018-05084 VR), Sweden.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code availability Codes for reproducing the results in this work are available at github.com/ooreilly/sbp.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

A Proof of Proposition 7

Proof Recall (45),

$$\Phi(q^*, \delta q) = -\Phi(q, \delta q^*) + \Phi(q, q^*).$$

By linearity,

$$\Phi(q - \delta q, q^* - \delta q^*) = \Phi(q, q^*) - \Phi(q, \delta q^*) - \Phi(\delta q, q^*) + \Phi(\delta q, \delta q^*),$$

i.e.,

$$\Phi(q, q^*) = \Phi(q - \delta q, q - \delta q^*) + \Phi(q, \delta q^*) + \Phi(\delta q, q^*) - \Phi(\delta q, \delta q^*). \quad (76)$$

Inserting (76) into (45) yields

$$\Phi(q - \delta q, q^* - \delta q^*) = \Phi(\delta q, \delta q^*). \quad (77)$$

We diagonalize each side of (77) individually. By the primal (28) and dual conditions (42), the left-hand side in (77) yields

$$\begin{aligned} \Phi(q - \delta q, q^* - \delta q^*) &= \oint_{\partial\Omega} (w^+ - \delta w^+)^T (w^{*+} - \delta w^{*+}) \\ &\quad - (w^- - \delta w^-)^T (w^{*-} - \delta w^{*-}) ds \\ &= \oint_{\partial\Omega} (w^+ - \delta w^+)^T (R^* - R^T) (w^{*-} - \delta w^{*-}) ds. \end{aligned} \quad (78)$$

Likewise, by the primal (30) and dual conditions (43), the right-hand side in (77) yields

$$\begin{aligned} \Phi(\delta q, \delta q^*) &= \oint_{\partial\Omega} (\delta w^+)^T (\delta w^{*+}) - (\delta w^-)^T (\delta w^{*-}) ds \\ &= \oint_{\partial\Omega} (\delta w^-)^T ((\delta R)^T - \delta R^*) \delta w^{*+} ds. \end{aligned} \quad (79)$$

In general, the only possibility for (78) to be equal to (79) is if (46) holds. $\square$

B Proof of Proposition 8

Proof (i) Upper bound: By the triangle inequality and Kronecker product property (51), we obtain

$$\begin{aligned}\|M_h\|_2 &= \|-A_x \otimes D_x + \mathcal{L}_h \otimes \Sigma L^T\|_2 \leq \|A_x \otimes D_x\|_2 + \|\mathcal{L}_h \otimes \Sigma L^T\|_2 \\ &= \|A_x\|_2 \|D_x\|_2 + \|\mathcal{L}_h\|_2 \|\Sigma L^T\|_2\end{aligned}$$

Since A_x is symmetric, $\|A_x\|_2 = \varrho(A_x)$, and from (54), the upper bound in (55) follows.

(ii) Lower bound: By definition, any vector $z \neq 0$ yields a lower bound on $\|M_h\|_2$:

$$\|M_h\|_2 = \sup_{\substack{x \in \mathbb{R}^m \\ x \neq 0}} \frac{\|M_h x\|_2}{\|x\|_2} \geq \frac{\|M_h z\|_2}{\|z\|_2}.$$

In particular, if D_x is a consistent approximation of the derivative, the choice $z = \mathbf{1} \otimes x'$ implies that $D_x e = 0$, and therefore

$$\frac{\|M_h z\|_2}{\|z\|_2} = \frac{\|(-A_x \otimes D_x + \mathcal{L}_h \otimes \Sigma L^T)(e \otimes x')\|_2}{\|e \otimes x'\|} = \frac{\|\mathcal{L}_h \otimes \Sigma L^T(e \otimes x')\|_2}{\|e \otimes x'\|}.$$

By (51), (54), and since $x' \neq 0$ is arbitrary,

$$\frac{\|\mathcal{L}_h e \otimes \Sigma L^T x'\|_2}{\|e \otimes x'\|_2} \geq \frac{r_l}{h} \sup_{\substack{x' \in \mathbb{R}^m \\ x' \neq 0}} \frac{\|\Sigma L^T x'\|_2}{\|x'\|_2} = \frac{r_l}{h} \|\Sigma L^T\|_2.$$

□

C Proof of Proposition 9

Proof By the projection matrix formula (35) and the submultiplicative norm property (50), we have

$$\|\Sigma L^T\|_2 = \|AP\|_2 \leq \underbrace{\|(L^T \delta L)\|_2^{-1}}_{\zeta_1} \underbrace{\|A \delta L\|_2}_{\zeta_2} \underbrace{\|L\|_2}_{\zeta_3}. \quad (80)$$

We will bound each of three factors in the right-hand side of (80) individually. By the definition of P in (35), ζ_1 satisfies

$$\zeta_1 = \|(L^T \delta L)\|_2^{-1} = \frac{1}{\|I - R \delta R\|_2} \quad (81)$$

By eigenvector orthogonality and orthonormality, ζ_2 satisfies

$$\begin{aligned}\zeta_2 = \|\Lambda \delta L\|_2 &= \left\| \left(X^+ \Lambda^+ (X^+)^T + X^- \Lambda^- (X^-)^T \right) \right. \\ &\quad \times \left. \left(X^- \sqrt{|\Lambda^-|^{-1}} - X^+ \sqrt{(\Lambda^+)^{-1}} \delta R \right) \right\|_2 \\ &= \|X^- \sqrt{|\Lambda^-|} - X^+ \sqrt{\Lambda^+} \delta R\|_2 \leq \sqrt{\varrho(A)} (1 + \|\delta R\|_2). \quad (82)\end{aligned}$$

Similarly, ζ_3 satisfies

$$\zeta_3 = \|L\|_2 \leq \sqrt{\varrho(A)} (1 + \|R\|_2). \quad (83)$$

By inserting (81)–(83) into (80), the bound (56) follows. $\square$

References

1. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39**(5), 1749–1779 (2002). <https://doi.org/10.1137/S0036142901384162>
2. Banks, J., Henshaw, W., Sjögreen, B.: A stable FSI algorithm for light rigid bodies in compressible flow. *J. Comput. Phys.* **245**, 399–430 (2013). <https://doi.org/10.1016/j.jcp.2013.02.050>
3. Carpenter, M.H., Gottlieb, D.: Spectral methods on arbitrary grids. *J. Comput. Phys.* **129**(1), 74–86 (1996). <https://doi.org/10.1006/jcph.1996.0234>
4. 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). <https://doi.org/10.1006/jcph.1994.1057>
5. 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). <https://doi.org/10.1006/jcph.1998.6114>
6. Causin, P., Gerbeau, J., Nobile, F.: Added-mass effect in the design of partitioned algorithms for fluid-structure problems. *Comput. Methods Appl. Mech. Eng.* **194**(42), 4506–4527 (2005). <https://doi.org/10.1016/j.cma.2004.12.005>
7. Fernández, D.C.D.R., Hicken, J.E., Zingg, D.W.: Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations. *Comput. Fluids* **95**, 171–196 (2014). <https://doi.org/10.1016/j.compfluid.2014.02.016>
8. Gao, L., Fernandez, D.C.D.R., Carpenter, M., Keyes, D.: SBP-SAT finite difference discretization of acoustic wave equations on staggered block-wise uniform grids. *J. Comput. Appl. Math.* **348**, 421–444 (2019). <https://doi.org/10.1016/j.cam.2018.08.040>
9. Gassner, G.J.: A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods. *SIAM J. Sci. Comput.* **35**(3), A1233–A1253 (2013). <https://doi.org/10.1137/120890144>
10. Ghasemi, F., Nordström, J.: Coupling requirements for multiphysics problems posed on two domains. *SIAM J. Numer. Anal.* **55**(6), 2885–2904 (2017). <https://doi.org/10.1137/16M1087710>
11. Ghrist, M., Fornberg, B., Driscoll, T.A.: Staggered time integrators for wave equations. *SIAM J. Numer. Anal.* **38**(3), 718–741 (2000). <https://doi.org/10.1137/S0036142999351777>
12. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, vol. 3. JHU press, Baltimore (2012)
13. Gustafsson, B., Kreiss, H.O., Oliger, J.: *Time-Dependent Problems and Difference Methods*. Wiley, New York (2013). <https://doi.org/10.1002/9781118548448.fmatter>
14. 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). <https://doi.org/10.1137/100790987>
15. Kennedy, C.A., Carpenter, M.H., Lewis, R.: Low-storage, explicit Runge–Kutta schemes for the compressible Navier–Stokes equations. *Appl. Numer. Math.* **35**(3), 177–219 (2000). [https://doi.org/10.1016/S0168-9274\(99\)00141-5](https://doi.org/10.1016/S0168-9274(99)00141-5)

16. Kopriva, D.A., Nordström, J., Gassner, G.J.: Error boundedness of discontinuous Galerkin spectral element approximations of hyperbolic problems. *J. Sci. Comput.* **72**(1), 314–330 (2017). <https://doi.org/10.1007/s10915-017-0358-2>
17. Kozdon, J.E., Dunham, E.M., Nordström, J.: Interaction of waves with frictional interfaces using summation-by-parts difference operators: weak enforcement of nonlinear boundary conditions. *J. Sci. Comput.* **50**(2), 341–367 (2012). <https://doi.org/10.1007/s10915-011-9485-3>
18. Kozdon, J.E., Dunham, E.M., Nordström, J.: Simulation of dynamic earthquake ruptures in complex geometries using high-order finite difference methods. *J. Sci. Comput.* **55**, 92–124 (2013)
19. 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 Equations*, Academic Press, pp. 195–212 (1974). <https://doi.org/10.1016/B978-0-12-208350-1.50012-1>
20. Nordström, J.: A roadmap to well posed and stable problems in computational physics. *J. Sci. Comput.* **71**(1), 365–385 (2017). <https://doi.org/10.1007/s10915-016-0303-9>
21. Nordström, J., Ghasemi, F.: The relation between primal and dual boundary conditions for hyperbolic systems of equations. *J. Comput. Phys.* **401**, 109032 (2020). <https://doi.org/10.1016/j.jcp.2019.109032>
22. Nordström, J., Forsberg, K., Adamsson, C., Eliasson, P.: Finite volume methods, unstructured meshes and strict stability for hyperbolic problems. *Appl. Numer. Math.* **45**(4), 453–473 (2003). [https://doi.org/10.1016/S0168-9274\(02\)00239-8](https://doi.org/10.1016/S0168-9274(02)00239-8)
23. O'Reilly, O., Petersson, N.A.: Energy conservative SBP discretizations of the acoustic wave equation in covariant form on staggered curvilinear grids. *J. Comput. Phys.* **411**, 109386 (2020). <https://doi.org/10.1016/j.jcp.2020.109386>
24. O'Reilly, O., Dunham, E.M., Nordström, J.: Simulation of wave propagation along fluid-filled cracks using high-order summation-by-parts operators and implicit-explicit time stepping. *SIAM J. Sci. Comput.* **39**(4), B675–B702 (2017). <https://doi.org/10.1137/16M1097511>
25. Petersson, N.A., O'Reilly, O., Sjögreen, B., Bydlon, S.: Discretizing singular point sources in hyperbolic wave propagation problems. *J. Comput. Phys.* **321**, 532–555 (2016). <https://doi.org/10.1016/j.jcp.2016.05.060>
26. Ranocha, H., Öffner, P., Sonar, T.: Summation-by-parts operators for correction procedure via reconstruction. *J. Comput. Phys.* **311**, 299–328 (2016). <https://doi.org/10.1016/j.jcp.2016.02.009>
27. Sudirham, J.J., Vandervegt, J., van Damme, R.: A study on discontinuous Galerkin finite element methods for elliptic problems. *Tech. Memorandum Dept. Appl. Math* **1690**, 1–21 (2003)
28. Strand, B.: Summation by parts for finite difference approximations for d/dx . *J. Comput. Phys.* **110**(1), 47–67 (1994). <https://doi.org/10.1006/jcph.1994.1005>
29. Svärd, M., Nordström, J.: On the order of accuracy for difference approximations of initial-boundary value problems. *J. Comput. Phys.* **218**(1), 333–352 (2006). <https://doi.org/10.1016/j.jcp.2006.02.014>
30. Svärd, M., Nordström, J.: Review of summation-by-parts schemes for initial-boundary-value problems. *J. Comput. Phys.* **268**, 17–38 (2014). <https://doi.org/10.1016/j.jcp.2014.02.031>
31. Svärd, M., Nordström, J.: On the convergence rates of energy-stable finite-difference schemes. *J. Comput. Phys.* **397**, 108819 (2019). <https://doi.org/10.1016/j.jcp.2019.07.018>
32. van Brummelen, E.H.: Added mass effects of compressible and incompressible flows in fluid–structure interaction. *J. Appl. Mech.* (2009). <https://doi.org/10.1115/1.3059565>
33. Verwer, J.G.: On time staggering for wave equations. *J. Sci. Comput.* **33**(2), 139–154 (2007). <https://doi.org/10.1007/s10915-007-9146-8>
34. Wilcox, L.C., Stadler, G., Burstedde, C., Ghattas, O.: A high-order discontinuous Galerkin method for wave propagation through coupled elastic-acoustic media. *J. Comput. Phys.* **229**(24), 9373–9396 (2010). <https://doi.org/10.1016/j.jcp.2010.09.008>