Department of Mathematics

On the Convergence Rates of Energy-Stable Finite-Difference Schemes

Magnus Svärd and Jan Nordström

LiTH-MAT-R--2017/14--SE
ON THE CONVERGENCE RATES OF ENERGY-STABLE
FINITE-DIFFERENCE SCHEMES

MAGNUS SVÄRD AND JAN NORDSTRÖM

Abstract. We consider initial-boundary value problems, with a $k$th derivative in time and a highest spatial derivative of order $q$, and their semi-discrete finite difference approximations. With an internal truncation error of order $p \geq 1$, and a boundary error of order $r \geq 0$, we prove that the convergence rate is: $\min(p, r + q)$.

The assumptions needed for these results to hold are: i) The continuous problem is linear and well-posed (with a smooth solution). ii) The numerical scheme is consistent, nullspace consistent and energy stable.

1. Introduction

In the finite-difference community the relation between truncation errors and convergence rates has been a long-standing subject for research. The reason is that for methods with order of accuracy higher than one ($p > 1$) at interior points, it is impossible to prove stability when the truncation error for the boundary scheme is also of order $p$. To prove stability, a boundary truncation error of order $r < p$ is inevitable and the question is what the convergence rate of the numerical solution will be.

For a stable scheme, it is clear that the convergence rate is at least $r$ and for schemes satisfying an energy estimate one can show that the rate in $L^2$ has to be at least $r + 1/2$. (See e.g. [GKO95].) However, higher convergence rates are often observed in practice. A number of questions emerge:

• Will the convergence rate drop to $r + 1/2$ with sufficient grid refinement?
• Can it be proven that the convergence rate is better than $r + 1/2$?
• What are the weakest conditions that give an improved convergence rate?
• What are the optimal rates?

The two first questions were decisively answered in the well-known papers [Gus75, Gus81]. Precise conditions that imply a convergence of $r + 1$ (which is often observed for first-order hyperbolic equations) were given. The analysis used the Laplace-transform technique (normal-mode analysis) and the conditions, although precise, are not easily evaluated for a high-order scheme.

For parabolic equations, a convergence rate of $\min(p, r + 2)$ is often observed in practice and in [ADG00] it was proven that the rate is at least $\min(p, r + 3/2)$.

The next step was taken in [SN06] where it was shown that the convergence rate of the boundary errors could be improved by the same order as the highest spatial derivative (denoted $q$). That is, for a parabolic equation with a second-order derivative ($q = 2$), the rate is $\min(p, r + 2)$. For the biharmonic equation, with a fourth derivative ($q = 4$), it is $\min(p, r + 4)$ and so on. The requirements for this improved rate was that the scheme was stable in $L^2$ and $L^\infty$ (pointwise). These conditions are significantly more tractable than the conditions emerging from the Laplace transform analysis. Nevertheless, proving an $L^\infty$ bound is usually beyond reach for most schemes.

Date: October 16, 2017.
Returning to the questions above, we conclude that there are cases where a higher rate can be expected from a theoretical viewpoint and they are not merely an illusion due to poor grid refinement. In many cases, the theoretical rates match the observed rates and thus appear to be sharp. However, there is no proof of the latter.

All the different convergence theories rely on an available energy estimate, which is required to prove that the errors generated by the internal scheme converges with order \( p \). However when analyzing the boundary errors, the energy stability is often not used. Instead, the scheme is proven stable by means of Laplace transform analysis and the boundary error convergence is inferred from that. (This is the method in \([\text{Gus75, Gus81, GKO95}]\).) In \([\text{SN06}]\) another approach is taken. Schemes that satisfy certain stability conditions are shown to converge at improved rates. The Laplace transform analysis is not used as a tool to prove stability but only to analyze convergence rates. This is the approach we are taking here as well.

Herein, we prove that the convergence rate is \( \min(p, r + q) \), under the assumption that the scheme is energy stable, consistent and nullspace consistent. Furthermore, we present arguments supporting that the rates \( \min(p, r + q) \) are sharp. This implies that the convergence rates are the same as in \([\text{SN06}]\) but the \( L^\infty \)-stability assumption is made redundant, leading to significantly more tractable conditions.

2. Preliminaries

2.1. Definitions. Throughout this paper, we consider one-dimensional linear Partial Differential Equations (PDEs) on a spatial domain, \( \Omega = (0, 1) \). Hence, we define the \( L^2 \) inner product of two real-valued functions as,

\[
(u, v) = \int_0^1 uv \, dx.
\]

In particular note that this definition implies that \( (1, v) \) is the mean value of \( v \). Furthermore, \( (u, u) = \|u\|^2 \) is the \( L^2 \)-norm.

Let \( u_{kt} \) denote the \( k \)th derivative in time, where in particular \( k = 0 \) denotes the non-differentiated variable. The general linear Initial-Boundary Value Problem (IBVP) we are interested in is,

\[
\begin{align*}
  &u_{kt} = Pu, \quad 0 < x < 1, \quad 0 < t \leq T, \quad k \geq 1, \\
  &L_0 u(0, t) = g_0, \\
  &L_1 u(1, t) = g_1, \\
  &u_j(x, 0) = f_j(x), \quad j = 0, ..., k - 1.
\end{align*}
\]

We assume that (1) is a constant coefficient problem.

**Definition 2.1.** If the problem (1) satisfies an energy estimate, \( \|u_{(k-1)t}\| \leq \text{Constant} \), with \( L_0, L_1 \) enforcing the fewest possible boundary conditions, we say that the problem is well-posed.

**Remark 2.2.** Our definition of well-posedness ensures that a unique solution exists.

Furthermore, we require that the initial and boundary data are smooth and compatible such that the problem (1) has a smooth solution. (Specifically this means that all \( f_j \) are bounded in \( L^2 \). In fact, \( f_j \in L^p \) \( p = 1...\infty \) since the domain is bounded.)

**Remark 2.3.** High-order convergence rates require sufficient smoothness of the functions \( f_j, g_0, g_1 \). The smoothness assumption can be relaxed to a certain number of smooth derivatives where the number depends on the order of the principal part and the order of accuracy.
We say that the initial-boundary value problem (1) satisfies an energy estimate if, after multiplication by \( u_{(k-1)t} \) and integration in time and space, an estimate of \( \| u_{(k-1)t} \| \) for all \( t \in (0,T] \), is obtained.

Next we introduce a discrete spatial domain \( \Omega_N \), which consists of the equally distributed points \( x_i = ih, i = 0, \ldots, N \) where \( h = 1/N \) is the grid spacing. Furthermore, let \( v(t) \) denote the approximate solution vector where \( v_i(t) \) is the semi-discrete solution at \( x_i \). Similarly, we define the grid functions \( [f_j]_i = f_j(x_i) \). We introduce a general semi-discrete approximation of the well-posed linear initial-boundary value problem (1),

\[
\begin{align*}
\mathbf{v}_{kt} &= P_h \mathbf{v} + B_h \mathbf{g}, \\
\mathbf{v}_{jt}(0) &= f_j, \quad j = 0, \ldots, k-1.
\end{align*}
\]

The boundary conditions are built into (2). Their contributions are included in \( P_h \) (\( \mathbf{v} \) dependent part) and \( B_h \mathbf{g} \) (data).

**Definition 2.4.** Assume that the problem (1) is well-posed and satisfies an energy estimate. Furthermore, assume that (2) satisfies an analogous energy estimate that bounds \( \| \mathbf{v}_{(k-1)t} \| \), for all \( t \in (0,T] \), in a discrete \( L^2 \)-norm. Then, we say that (2) is energy stable.

To develop the theory, it will be convenient to work with the homogeneous version of problem (1) where the boundary data, \( g_0 = g_1 = 0 \). We state it as

\[
\begin{align*}
\mathbf{u}_{kt} &= P \mathbf{u}, \quad 0 < x < 1, \quad 0 < t \leq T, \quad k \geq 1, \\
L_0 u(0, t) &= 0, \\
L_1 u(1, t) &= 0, \\
\mathbf{u}_{jt}(x, 0) &= f_j(x), \quad j = 0, \ldots, k-1,
\end{align*}
\]

and its corresponding semi-discretization

\[
\begin{align*}
\mathbf{v}_{kt} &= P_h \mathbf{v}, \\
\mathbf{v}_{jt}(0) &= f_j, \quad j = 0, \ldots, k-1.
\end{align*}
\]

We make a few remarks on the homogeneous and inhomogeneous problems.

- It is straightforward to transform (1) into (3) (augmented with a forcing function) by the transformation, \( u = \mathbf{v} + \Phi \) where \( \Phi \) is a smooth function satisfying the boundary conditions. The only condition is that \( g_{0,1} \) are smooth, which we have assumed. (See [GKO95] and equation (35) below where we carry out the transformation.)
- The previous point implies that if one problem is well-posed, so is the other. Similarly, if either of the semi-discrete schemes is energy stable, then so is the other. (In the sense that when data is smooth the two forms are equivalent since a bounded forcing function does not affect well-posedness nor stability.)

### 2.2. Energy stable schemes

The development of Summation-By-Parts (SBP) finite-difference schemes and the Simultaneous Approximation Term (SAT) methodology enables straightforward construction of energy stable finite difference schemes. We will give a brief introduction to SBP-SAT schemes and refer to the review articles [SN14, FHZ14] for more information.

**Definition 2.5.** An SBP first-derivative approximation on \( \Omega_N \) is defined by

\[
\mathbf{D}_1 \mathbf{v} = H^{-1}Q \mathbf{v},
\]

where \( Q + Q^T = B = \text{diag}(-1, 0, 0, \ldots, 1) \).
$H$ is a symmetric positive-definite matrix with elements of size $O(h)$. The matrix $H$ defines an inner product $(u, v) = u^T H v$ and an $L^2$-equivalent discrete norm, $(v, v) = \|v\|^2$. At interior points, $H$ is diagonal (with diagonal element $h$). Near the boundary $H$ need not be diagonal. (All matrices are of size $(N + 1) \times (N + 1).$)

A second-derivative SBP approximation is given by

$$D_2 v = H^{-1}(-\bar{A} + BS)v,$$

where $\bar{A}$ is symmetric positive semi-definite and $Sv$ is a first-derivative approximation. It suffices that $S$ is defined on the boundaries; it can be zero elsewhere.

For more information on SBP operators, we refer to [Mat14, MN04, Str94].

Remark 2.6. The condition on the matrix $\bar{A}$ can sometimes be weakened to $\bar{A} + \bar{A}^T$ being symmetric positive definite for parabolic equations. Here, we require symmetry of $\bar{A}$, since it is necessary for e.g. the wave equation.

In fact, the definitions of both the first-derivative and second-derivative approximations can be stated on a generic form which is common for SBP operators approximating any derivative:

Definition 2.7. Let $H$ be defined as in Definition 2.5. Then an SBP $m$th-order derivative approximation takes the form:

$$D_m = H^{-1}(A + R), \ m = 1, 2, 3, ...$$

If $m$ is odd $A$ is skew-symmetric and if $m$ is even, $(-1)^{m/2} A$ is symmetric positive definite. The matrix $R$ produces approximations of the suite of boundary terms.

As an example of Definition 2.7, consider the first-derivative approximation $D_1$, where $A$ is the skew-symmetric part of $Q$ and $R = \frac{1}{2} B$. The analogy with the derivative is seen in,

$$\langle u, u_x \rangle = \int_0^1 uu_x dx = -\frac{1}{2}(u(0, t)^2 - u(1, t)^2),$$

$$\langle u, D_1 u \rangle = u^T H^{-1}(A + R) u = u^T A u + \frac{1}{2} u^T B u = -\frac{1}{2}(u_0^2 - u_N^2).$$

It is straightforward to make the same analogy for $D_2$. Note that for higher derivatives, $R$ contains several boundary terms on each boundary due to the multiple partial integrations needed to reach a skew-symmetric or symmetric form.

In general the order of the truncation errors of the SBP approximations are $r$ in a few points adjacent to the boundaries and $p > r$ at the interior points. In the special case where $H$ is diagonal, $r = p/2$.

The SAT method ([CGA94, CNG99]) is one systematic method (but not the only one, see [Ols95a, Ols95b]) to implement boundary conditions in conjunction with SBP schemes such that energy estimates can be derived. We present the SAT technique by means of a simple example. Consider

$$u_t + u_x = 0, \quad 0 < x < 1,$$

$$u(0, t) = g(t),$$

$$u(x, 0) = f(x).$$

This problem satisfies the energy estimate

$$\|u(\cdot, t)\|^2 + \int_0^t u(1, t)^2 dt = \|f\|^2 + \int_0^t g(t)^2 dt.$$
A semi-discretization is given by,
\[ v_t + H^{-1}Qv = -H^{-1}e_0(v_0 - g(t)), \]
\[ v(0) = f \]
where \( e_0 = (1, 0, ..., 0)^T \) has length \( N + 1 \). The right-hand side is the weak enforcement of the boundary condition termed SAT. This scheme satisfies the analogous estimate
\[ \|v(t)\|^2 + \int_0^t v_N(t)^2 + (v_0 - g(t))^2 dt = \|f\|^2 + \int_0^t g(t)^2 dt. \]
The last term on the left-hand side introduces extra numerical damping that vanishes as \( v_0 \to g(t) \).

### 2.3. Consistency of finite difference schemes.

A straightforward way to derive a consistent finite difference operator, is to require that it exactly differentiates polynomials up to a certain degree. That is,
\[ D_m x_i = 0, \quad i = 0 \ldots m - 1, \]
\[ D_m x^m = m!, \]
where the difference operator \( D_m \) approximates \( d^m / dx^m \) and \( x^i = (..., (jh)^i, ((j + 1)h)^i, ...)^T \).

We assume that \( x \) has the appropriate length and take \( 0^0 = 1 \) as a definition. If the conditions (7) and (8) are met, \( D_m \) is first-order accurate. If higher order polynomials are differentiated exactly, higher-order of accuracy is obtained. Herein, we assume that all difference schemes are derived by this procedure.

The consistency conditions (7) and (8) are equivalent to the conditions obtained from another commonly used method where as many terms in a Taylor series as possible are cancelled by appropriate choices of the coefficients in a finite difference scheme. For instance,
\[ D^2 u |_{x_i} = u_{xx}(x_i) + c_1 h^2 u_{xxxx}(x_i) + ..., \]
would be a second-order second-derivative approximation, where \( u(x) \) is a smooth function and the vector function \( u \) is its projection on the grid (i.e., \( u_i = u(x_i) \)).

**Note that the order of accuracy is always integer-valued by these two equivalent procedures.**

In the analysis of convergence rates it is important that consistency is defined as above. Although this is a natural definition, it is not the only one. In numerical analysis textbooks, consistency is often defined as,
\[ D_2 u |_{x_i} = u_{xx}(x_i) + E(h), \]
where \( E(h) \to 0 \) as \( h \to 0 \). The rationale for definition (10) is that it is the weakest definition that will satisfy Lax-Richtmyer’s Equivalence Theorem, [Str89]. Since it is weaker, it does not rule out the following approximation of \( u_{xx} \):
\[ D'_2 u |_{x_i} = u_{xx}(x_i) + h^\alpha C u. \]

According to (10), \( D'_2 \) is a consistent approximation (of order \( \alpha \)), as long as \( \alpha > 0 \) and \( C u \) is \( O(1) \). (\( C \) can e.g. be a first-derivative approximation.)

However, (11) is not consistent according to (7) and (8). The reason is that the error, \( h^\alpha C u \), is not really a truncation error appearing in a Taylor series of the error. It is an artificially added error term.

To end the discussion on consistency we point out that for finite differences it is always possible to construct schemes by exactly differentiating polynomials with increasing order and this will only lead to integer-valued truncation errors. Since
approximations like (11) always can be avoided, we will only consider schemes that satisfy (7) and (8) (or equivalently the Taylor-series method). (Foregoing the discussion, we remark that allowing approximations like (11), may induce substandard convergence rates.)

2.4. Nullspace of difference operators. Central to the analysis in this paper is the notion of nullspace of a linear operator $L$.

**Definition 2.8.** The nullspace of a linear operator $L$ is the set of elements $v$ of a vector space that satisfies $Lv = 0$. We denote the nullspace $\mathcal{N}$ and say that it is trivial if $v = 0$ is its only element.

Since this paper addresses linear PDEs, it is the nullspace of a linear differential operator that is of interest. We observe that, for any constants $c_i$, $i = 0 \ldots p - 1$. Hence, all monomials of order $1$ to $m - 1$ belong to the nullspace of $d^m/dx^m$.

Similarly, the nullspace of a discrete operator is denoted $\mathcal{N}_h$. That is $v \in \mathcal{N}_h$ if $D_m v = 0$. Note that, any derivative approximation that satisfies (7), encompasses the nullspace of the corresponding derivative operator, where we use the obvious injection $\Omega \to \Omega_N$ by $v_i = v(x_i)$.

**Definition 2.9.** A difference operator is termed nullspace consistent if 1) all nullspace modes of the corresponding derivative operator are nullspace modes of the difference operator, and if 2) no other nullspace modes of the difference operator exist. This should hold for all $h > 0$.

**Remark 2.10.** A nullspace consistent operator need not be consistent, since it may fail to satisfy (8). We also assume that all difference operators are scaled correctly. That is, the elements of an operator $D_m$ is scaled as $h^{-m}$, such that the result of $D_m x^m \sim O(1)$, regardless of whether it is consistent or not.

However, a consistent difference operator need not be nullspace consistent since it may have more nullspace modes than the corresponding derivative approximation.

**Example 2.11.** Consider the operator $d/dx$ which has the non-trivial nullspace $v(x) = c \cdot 1$, where $c$ is an arbitrary constant. A second-order accurate approximation is:

$$v'(x_i) \approx \frac{v_{i+1} - v_{i-1}}{2h}, \quad -\infty < i < \infty.$$  

Equating (12) with zero and solving the difference equation by the ansatz $v_i = r^i$ give $v_i = 1$ and $v_0 = (-1)^i$. Hence, the following vectors belong to the nullspace, $v_i = c \cdot 1 + d \cdot (-1)^i$ where $c, d$ are arbitrary constants. The first mode corresponds to $v(x) = c \cdot 1$ while the second one does not have a counterpart in the nullspace of $d/dx$. The nullspace of the difference operator is larger than the nullspace of $d/dx$. It is consistent in the sense of (7) and (8) but not nullspace consistent.

**Remark 2.12.** In view of the previous example, we introduce a simplified nomenclature for nullspaces. We characterize the nullspace by the functions that span it. That is, we say that the non-trivial nullspace of e.g. $d/dx$ is $v(x) = 1$, with the implicit meaning that the nullspace is $c \cdot 1$.

Turning to finite (non-periodic) domains, the nullspaces of the differential operator and the difference operator are often the same.
Example 2.13. A second-order SBP derivative approximation is
\[ v'(x_0) \approx \frac{v_1 - v_0}{h}, \]
\[ v'(x_i) \approx \frac{v_{i+1} - v_{i-1}}{2h}, \quad i = 1 \ldots N - 1, \]
\[ v'(x_N) \approx \frac{v_N - v_{N-1}}{2h}. \]
To find the nullspace of this operator, we solve the difference equation at the interior points and obtain as before, \( v_i = (-1)^i \) and \( v_i = 1 \). The latter satisfies the boundary schemes as well. So \( v_1 = 1 \) is part of the nullspace of (13) (which it should be by consistency). However, \( v_i = (-1)^i \) does not satisfy either of the boundary schemes. Hence, the boundary closures of the SBP scheme removes the vector in the discrete nullspace that does not coincide with the nullspace \( v(x) = 1 \) of \( d/dx \). □

Example 2.14. By contrast, the following approximation does not remove the nullspace-inconsistent mode.
\[ v'(x_0) \approx \frac{v_2 - v_0}{2h}, \]
\[ v'(x_i) \approx \frac{v_{i+1} - v_{i-1}}{2h}, \quad i = 1 \ldots N - 1, \]
\[ v'(x_N) \approx \frac{v_N - v_{N-2}}{2h}. \]
This derivative approximation is not in SBP form since it lacks the necessary symmetry properties to allow an energy estimate. Hence, it cannot be used to form an energy stable scheme. This is not a coincidence. Below, we show that nullspace consistency and the SBP property are interlinked. □

Since all SBP schemes have the same structure, a repeated interior stencil and finitely many stencils forming the boundary closures, we can formulate a procedure for determining the nullspace of an SBP operator.

1. Solve the internal scheme, equated with zero, as a difference equation, to obtain a set of internal solutions.
2. Test which of the internal solutions that belong to the nullspace of all boundary stencils.

Alternatively, one can numerically compute the eigenvalues and eigenvectors of the operator. The eigenvectors associated with zero eigenvalues are the nullspace modes.

Example 2.15. Consider the standard second-order first-derivative SBP operator (13). Here, \( H = h \cdot \text{diag}(1/2, 1, 1, \ldots, 1/2) \), \( R = \text{diag}(-1/2, 0, 0, \ldots, 0, 1/2) \) and
\[ A = \begin{pmatrix}
0 & 1/2 & \cdots & \\
-1/2 & 0 & 1/2 & \\
& \ddots & \ddots & \\
& & -1/2 & 0 & 1/2 & \\
& & & \ddots & -1/2 & 0
\end{pmatrix}. \]
As before, the nullspace modes of the interior scheme (interior part of the matrix \( A \)) are 1 and \((-1)^i\). At the boundary neither is a nullspace mode. Hence, \( A \) has no nullspace modes. Furthermore, \( H^{-1}A \) is a second-order approximation at the interior and inconsistent at the boundary points. The operator \( R \), makes \( H^{-1}(A + R) \) consistent. It does so by using the minimal number of non-zero elements (here, one at each boundary). Hence, the number of unknowns are only sufficient to satisfy the consistency requirements making it first-order accurate at the boundary and also
nullspace consistent. No other nullspace modes are introduced and the operator is nullspace consistent.

\[ \square \]

Remark 2.16. To the best of our knowledge, all SBP difference operators are nullspace consistent with their respective derivatives. (For instance, it is true for the standard \( D_2 \) operators.)

2.5. Nullspace of finite difference schemes. Having defined the nullspace of SBP difference operators, we continue to investigate the nullspace in the presence of boundary conditions.

Definition 2.17. The spatial nullspace, \( N \), of the initial-boundary value problem (1) are all functions, \( w \), that satisfy

\[
Pw = 0, \quad L_0 w = 0, \quad L_1 w = 0.
\]

Definition 2.18. The spatial nullspace, \( N_h \), of the semi-discretization (2) of (1), are the grid functions \( w \), that satisfy,

\[
P_h w = 0.
\]

Note that \( P_h \) includes the boundary schemes that enforce the homogeneous boundary conditions. The spatial nullspace will often be referred to simply as the nullspace. Furthermore, if the nullspaces of (14) and (15) are the same (in the sense of Def. 2.9), we say that \( P_h \) is nullspace consistent.

Many times, the boundary conditions of well-posed problems will render the nullspace trivial, as in the following example.

Example 2.19. \( u_t + u_x = 0, \quad 0 < x < 1 \)

\[
u(0, t) = g(t), \quad u(x, 0) = f(x).
\]

The nullspace of \( u_x \) is \( u = 1 \). However, since \( u(0, t) = 0, \quad u = 1 \) is not in the nullspace of the boundary condition. (Recall, that according to Definitions 2.17 and 2.18 the nullspace is given by the homogeneous problem.) Hence, the boundary condition makes the nullspace of the initial-boundary value problem trivial.

Next, we consider the 2nd-order SBP-SAT approximation of (16).

\[
\begin{align*}
(v_0)_t + \frac{v_1 - v_0}{h} &= -\frac{2}{h}(v_0 - g), & (g = 0), \\
\frac{v_{i+1} - v_i}{2h} &= 0, & i = 1...N-1, \\
\frac{v_N - v_{N-1}}{h} &= 0.
\end{align*}
\]

The right-hand side is the SAT that enforces the boundary condition.

Considering the nullspace of the spatial operator, we note that \( v_i = 1 \) satisfies the internal and right boundary scheme. However, at the left boundary

\[
\frac{v_1 - v_0}{h} + \frac{2}{h}v_0 \neq 0,
\]

if \( v_1 = v_0 = 1 \). Hence, the constant is not in the nullspace. The approximation is both consistent and nullspace consistent.

Even in cases when the boundary conditions remove the nullspace modes of the PDE, there may be non-trivial nullspace modes present in the corresponding semi-discrete scheme.
Example 2.20. Consider the wave equation
\[ u_{tt} = u_{xx}, \quad 0 < x < 1, \]
\[ u(0, t) = 0, \]
\[ u(1, t) = 0. \]

The nullspace of the spatial part is trivial. (The constant is removed from the nullspace by the boundary conditions.) An SBP-SAT discretization (where we ignore the right boundary) takes the form,
\[ v_{tt} = D^2 v + H^{-1}(S^T E_0 - \frac{\tau}{h} E_0) v = M v \]
where \([E_0]_{11} = 1\) and elsewhere zero, and \(\tau\) is a parameter. For stability, \(\tau\) must be chosen to make the matrix \(M\) negative semi-definite. For most stable choices of \(\tau\), \(M\) is negative definite, i.e., there is no zero eigenvalue and hence no nullspace mode. Then the scheme is nullspace consistent. However, for the marginal value of \(\tau\) that leads to stability, \(M\) becomes negative semi-definite. It has one zero eigenvalue. Hence, the nullspace of the scheme is larger than the nullspace for the IBVP, i.e., nullspace consistency is violated. However, the scheme is still stable and consistent.

The example shows that nullspace consistency does not follow immediately from nullspace consistency of the individual derivative approximations. It must be tested for the complete semi-discrete scheme. □

The nullspace modes of the differential operator need not only be polynomials. Consider the PDE \(u_t = (\partial^2_x - \beta^2)u, \beta > 0\). The spatial nullspace is given by the ODE,
\[ \partial^2_x - \beta^2 u = 0. \]

The ansatz \(u = e^{\beta x}\) results in the characteristic equation, \(r^2 - \beta^2\) and the nullspace modes are: \(e^{\beta x}\) and \(e^{-\beta x}\).

Approximating (18) by an SBP scheme, e.g. \(D^2 - \beta^2\), does not result in nullspace consistency. The nullspace modes are not polynomials and hence they are not exactly approximated by the scheme. (They will only be zero to design order.)

This observation might suggest that despite the SBP difference operators being nullspace consistent with their respective derivatives (due to their ability to exactly differentiate polynomials of limited degree) they might anyway be nullspace inconsistent in practice. This, however, is not the case, since exponential modes will always be annihilated by the boundary conditions.

On infinite domains, this is easily realized since some kind of trail-off condition is needed to maintain boundedness of the solution. This rules out exponential solutions. On bounded domains, which is the topic herein, there are a few ways to see that exponential nullspace modes can not be present. 1) A well-posed initial-value problem is well-posed with respect to both half-space problems and the periodic problem. A half-space problem is obtained by sending one boundary to infinity and replacing its boundary condition with a trail-off condition. Hence, all exponential nullspace modes disappear for the same reason as in the infinite domain case. 2) Laplace-transform analysis shows that all exponential modes must be bounded by a boundary conditions. Hence, they are not nullspace modes. (See [GKO95].)

We conclude that only polynomials can populate a non-trivial nullspace. Before we consider cases with non-trivial nullspaces, we will prove a crucial property of nullspaces.
2.5.1. Properties of the nullspace. Consider the problem,
\begin{equation}
\label{nullspace}
\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, \\
\frac{\partial u}{\partial x}(0, t) &= \frac{\partial u}{\partial x}(1, t) = 0, \\
u(x, 0) &= f(x).
\end{aligned}
\end{equation}

This problem has one nullspace mode, \( u = 1 \). Next, let \( \bar{f} = (1, f) \), be the mean value of the initial data and consider the problem
\begin{equation}
\begin{aligned}
\frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial x^2}, & 0 < x < 1, \\
\frac{\partial v}{\partial x}(0, t) &= \frac{\partial v}{\partial x}(1, t) = 0, \\
v(x, 0) &= f(x) - \bar{f}.
\end{aligned}
\end{equation}

It has the solution \( v = u - \bar{f} \). The underlying reason why the two solutions are connected in this way is that the problems are linear and well-posed in \( L^2 \). Since \( L^2 \) is a Hilbert space, there is an orthonormal basis and by linearity each basis function evolves independently of the other. Furthermore, since the nullspace is a subspace it is spanned by a subset of the basis of the Hilbert room. Hence, we specifically know that these modes evolve separately.

\textbf{Remark 2.21.} For \eqref{nullspace} this is well-known. The problem can be solved by separation of variables and expressed as a Fourier sum. One mode in the Fourier sum is the constant (which is in the nullspace) and hence, it evolves independently of the other modes.

This observation implies that the mean value of \( u \) in \eqref{nullspace}, is set initially and thereafter it is not affected by the PDE. The spatial operator maintains a division between the nullspace and non-nullspace part of the solution.

The above properties are important in the analysis of the convergence rates and therefore we must demonstrate that a nullspace-consistent scheme also separates the nullspace solution and the non-nullspace solutions.

Specifically, we can deduce the nullspace, \( N_h \) of \( P_h \) and hence we can split any vector into a nullspace part and its orthogonal complement, i.e., \( v = v^N + v^\perp \). By definition, \( P_h v^N = 0 \) and we wish that \( P_h v^\perp \in N^\perp \). If so, we can split \( v_t = P_h v \) into the problems \( v^N_t = 0 \) and \( v^\perp_t = P_h v^\perp \).

To prove that this is the case, consider \eqref{nullspace} with \( N \) unknowns. There is a non-singular matrix \( S \) such that \( P_h = SJS^{-1} \) where \( J \) is the Jordan matrix, with \( k \) Jordan blocks \( J_{n_i}(\lambda_i) \) of size \( n_i \) on the diagonal. (\( \sum_{i=1}^k n_i = N \). (See Thm. 3.1.11, [HJ90].) By rewriting the semi-discrete scheme as,
\begin{equation}
\begin{aligned}
\frac{d}{dt} w &= S^{-1} v, \\
S^{-1} P_h S(S^{-1} v) &= J w, \\
\end{aligned}
\end{equation}

it is evident that the modes associated with one Jordan block evolve independently of the other modes.

Given that \( P_h \) is a nullspace consistent discretization and that the continuous system \eqref{continuous} has \( n_N \) nullspace modes, we know that there are \( n_N \) zero eigenvalues of \( P_h \). Hence, one of the Jordan blocks is \( J_{n_N}(0) \).

Furthermore, since the nullspace modes can only be polynomials, it is possible to cast the eigenvectors as an orthogonal set for the nullspace modes. That is, \( J_{n_N}(0) \) is in fact diagonal. Next, we order the Jordan blocks of \( J \) such that \( J_{n_N} \) sits in the top-left corner, i.e., the first \( n_N \) components of \( w \) are nullspace modes. Hence,
\begin{equation}
\begin{aligned}
(w_{kt})_i &= 0 \cdot (w)_i, \quad i = 1 \ldots n_N \\
\end{aligned}
\end{equation}

That is \( w_i \) is a \( (k-1) \)th-order polynomial in time with its coefficients given by the initial data. Most importantly, these modes are neither affected by or affect, the non-nullspace modes.
These arguments hold for a finite \(N\). Since we are considering a finite difference scheme, we must allow that \(N \to \infty\). However, being a finite difference scheme the matrix is extended by the addition of identical stencils in the interior. These stencils extend the matrix by a row that is linearly independent of all the previous. Or equivalently, by a 1 \(\times\) 1 Jordan block (i.e. a single eigenvalue) in (20).

**Remark 2.22.** This observation implies that the number of non-diagonal Jordan blocks are at most finite and of finite size. If Jordan blocks were allowed to grow indefinitely, the time-polynomial order of a mode would become infinite. (Hence, \(L^2\) stability also precludes this possibility.)

We have proven the following proposition.

**Proposition 2.23.** Let \(P_h\) be the nullspace-consistent spatial operator in the general semi-discretization (2) that approximates the well-posed problem (1). Then:

1. Any semi-discrete solution \(v\) can be written as \(v = v^\perp + v^N\) where \(v^\perp \in N^\perp\) and \(v^N \in N\).
2. \(v^N\) can be expressed in a time-independent orthonormal \((L^2)\)-base.
3. If \(v^\perp \neq 0\), then \(P_h v = P_h v^\perp + P_h v^N = P_h v^\perp \in N^\perp\).

The proposition implies that the nullspace remains invariant when the system of ODEs, \(v_t = P_h v\), is solved. Hence, if the spatial operator is nullspace consistent the full semi-discrete scheme is also nullspace consistent. That is, there is no transfer of information between \(N\) and \(N^\perp\). The nullspace part evolves separately from the non-nullspace part.

2.5.2. Non-trivial nullspaces. In some cases, part of the nullspace remains in the initial-boundary value problem. Consider the well-posed linear scalar IBVP

\[
\begin{align*}
\text{(22)} & \quad u_t = Pu, \quad 0 < x < 1 \\
& \quad L_0 u(0, t) = 0 \\
& \quad L_1 u(1, t) = 0 \\
& \quad u(x, 0) = f(x),
\end{align*}
\]

where \(P\) is a linear differential operator of order \(q\). Let us assume that there exists a function \(v \in N\). (To reduce notation, we assume that \(N\) is one-dimensional and thereby spanned by \(v\).) That is,

\[
\begin{align*}
\text{(23)} & \quad P v = 0, \quad L_0 v(0, t) = 0, \quad L_1 v(1, t) = 0.
\end{align*}
\]

Such a \(v \in N\) is not affected by the PDE. It satisfies \(v_t = PV = 0\). Hence, \(v\) is independent of time (but not necessarily of \(x\)).

Since the problem (22) is a priori assumed to be well-posed, we know that a unique solution exists and since it is linear, we can write the solution as \(u = \alpha v + \sum_{i=1}^\infty \alpha_i v_i\) where \(\alpha, \alpha_i\) are scalars and \(v_i\) is a complete basis of \(N^\perp\). (Note that \(v\) is by definition orthogonal to all \(v_i\)'s since \(v\) is part of the nullspace.)

We can now conclude that \(\alpha\) is given by the initial data since

\[
\begin{align*}
\text{(24)} & \quad u(x, 0) = \alpha v(x) + \sum_{i=1}^\infty \alpha_i v_i(x, 0) = f(x)
\end{align*}
\]

and by orthogonality \(\alpha(v, v) = (v, f)\). Since \(v_t = 0\), \(\alpha\) is independent of \(t\). It is defined initially and thereafter it remains the same. (This is the same statement as (21).)

**Remark 2.24.** The \(\alpha_i\):s are also determined at \(t = 0\) by \(f\), but unlike \(\alpha\), they are subsequently affected by the PDE and evolve in time.
The analogous procedure is valid for the semi-discretization provided that the spatial operator is nullspace consistent.

Definition 2.25. The nullspace of the initial-boundary value problem (1) (not just the spatial part) is called the complete nullspace.

Equally, important for the analysis of convergence rates is the case with second-derivatives in time and a non-zero nullspace. That is,
\[ u_{tt} = Pu, \quad 0 < x < 1, \]
\[ L_0 u(0, t) = 0, \]
\[ L_1 u(1, t) = 0, \]
\[ u(x, 0) = f(x), \]
\[ u_t(x, 0) = h(x). \]
We assume that this problem is well-posed and that \( v \in \mathcal{N} \) of the spatial operator including the homogeneous boundary conditions. Then, \( \alpha tv + \beta v \) is in the complete nullspace of the initial-boundary value problem. We can write the solution as
\[ u = \alpha tv + \beta v + \sum_{i=1}^{\infty} \alpha_i t^i w_i. \]

The previous discussion leads to the following Theorem.

Theorem 2.26. Any mode in the nullspace of a well-posed linear homogeneous initial-boundary value problem (3), is determined by the initial data and is thereafter unaffected by the PDE. The equivalent statement holds for the approximation (4), if it is nullspace consistent.

Proof. The complete nullspace of the IBVP (3) consists of the modes \( t^z w_i, \quad z = 0, \ldots, k - 1 \) and where \( w_i \in \mathcal{N}, \quad i = 1, \ldots, n_N \). Without restriction we can consider \( w_i \) to be the orthogonal elements of a basis that spans the \( n_N \)-dimensional nullspace \( \mathcal{N} \). By letting \( v^\perp \) denote the solution in \( \mathcal{N}^\perp \), we can write
\[ u(x, t) = v^\perp + \sum_{z=0}^{k-1} \sum_{i=1}^{n_N} \alpha_{zi} t^z w_i. \]

Projecting the initial data on the nullspace
\[ (w_j, f_0) = (w_j, u(\cdot, 0)) = (w_j, \alpha_{0j} w_j) \]
allows us to conclude that all \( \alpha_{0j}, \quad j = 1, \ldots, n_N \) are uniquely defined. Similarly, taking the time derivative of (26) and equating \( (w_j, f_1) = (w_j, u_t(\cdot, 0)) \) determines all \( \alpha_{1j} \), and similarly for higher derivatives of time. Since there are \( k \) modes in time and \( k \) initial data, all modes of the complete nullspace can be determined. (Analogously for the semi-discrete scheme (4).)

Remark 2.27. Note that the theorem concerns the homogeneous problems (3) and (4). We shall later see that non-homogeneous boundary conditions and/or a non-trivial forcing function, feeds energy into the nullspace modes. Still, and as is implied by the theorem, the nullspace modes are unaffected by the operators \( P \) (or \( P_h \)).

Furthermore, in both of the two examples, (22) and (25), we can recast the problem by a change of variables such that the nullspace modes of the new problem do not carry any energy. (That is, \( \alpha = \beta = 0 \) in the example (25).) This procedure is shown in the next example which is fundamental for this paper.
Example 2.28. Assume that data are smooth and compatible, and consider
\[ u_{tt} = u_{xx}, \quad 0 < x < 1, \]
\[ u_x(0, t) = 0, \]
\[ u_x(1, t) = 0, \]
\[ u(x, 0) = f(x), \]
\[ u_t(x, 0) = h(x). \]
(27)

Here, the nullspace of \( u_{xx} \) contains the elements 1 and \( x \). The mode \( x \) is annihilated by the boundary conditions. However, the constant mode still remains in the nullspace and is unaffected by the PDE. In analogy with the previous example (25), the complete nullspace is thus \( \alpha v + \beta v \), where \( \alpha \) and \( \beta \) are determined by the initial conditions:

\[ (u(x, 0), v(x)) = (f, 1) = \int_0^1 f(x) \, dx = \beta(v, v) = \beta, \]
\[ (u_t(x, 0), v(x)) = (h, 1) = \int_0^1 h(x) \, dx = \alpha(v, v) = \alpha. \]

(Note that \( \beta \) is the mean value of \( f \) and \( \alpha \) the mean of \( h \).) Next, we split the solution into the time varying mean value \( \beta v + t\alpha v \) and a spatially varying part, \( w = u - (\beta + t\alpha) \) (where we have used that \( v = 1 \)). By linearity \( w \) satisfies

\[ w_{tt} = w_{xx}, \quad 0 < x < 1, \]
\[ w_x(0, t) = 0, \]
\[ w_x(1, t) = 0, \]
\[ w(x, 0) = f(x) - \beta, \]
\[ w_t(x, 0) = h(x) - \alpha. \]
(28)

\{var_wave\}

It is easy to see that if we carry out the same analysis for (28), we find that \( v = 1 \) is in the nullspace and hence the complete nullspace is \( \alpha't + \beta' \). However, the mean values of the initial data are now zero, i.e., \( \alpha' = \beta' = 0 \).

We include the next Lemma merely as an observation. We will not need it in the subsequent analysis but one assumption in [SN06] is that the solution is pointwise bounded and this is met by (27). (We present the continuous proof but an energy stable (nullspace consistent) semi-discrete approximation has the same property.)

Lemma 2.29. The solution \( u \) of (27) satisfies \( u \in L^\infty(0, T; H^1(0, 1) \cap L^\infty(0, 1)) \)

Proof. For (27) it is an elementary exercise to derive a bound on the energy \( E = \|u_t\|^2 + \|u_x\|^2 \). Hence, by the standard Poincare estimate we get

\[ \|u(t) - u_m\| \leq C(\|u_x(\cdot, t)\|) \]

where \( u_m \) is the mean value of \( u \). Here, \( u_m = \beta + t\alpha \) is bounded. Hence, we get an \( L^2 \) estimate of \( u \) (by the triangle inequality applied on \( \|(u - u_m) + u_m\| \)), and hence \( u \) is bounded in \( H^1 \). By Sobolev embedding, we have \( u \in L^\infty(0, T; L^\infty(0, 1)) \).

In this section we have shown that the presence of a nullspace in a well-posed IBVP means that the homogeneous PDE evolves the solution in the complement of the nullspace. The nullspace modes themselves are unaffected by the PDE and as we have shown they are set by the initial data.

In the particular example (27) this principle is demonstrated. There, it is in fact the mean value that remains unaffected and the PDE describes evolution around that mean. Hence, the nullspace mode of (27) is benign in the sense that it does not prohibit the existence of strong estimates as shown in Lemma 2.29.
3. Stability

In this section we will analyze the relation between energy based stability analysis and Laplace-transform based stability analysis. (See [GKO95] for an introduction.)

3.1. First derivative in time. We introduce the initial-boundary value problem
\begin{align*}
  u_t &= Pu, \quad 0 < x < 1 \\
  L_0 u(0, t) &= g_0 \\
  L_1 u(1, t) &= g_1 \\
  u(x, 0) &= f(x)
\end{align*}
and a general semi-discretization of (29),
\begin{align*}
  v_t &= P_h v + B_h g, \\
  v(0) &= f.
\end{align*}
To reduce notation, let \( f = 0 \). Laplace transforming the above problem leads to,
\begin{align*}
  s \hat{v} &= P_h \hat{v} + B_h \hat{g}.
\end{align*}
Furthermore, we rescale \( P_h \) such that the elements of \( P_h \) become \( O(1) \). E.g., if the principal part of \( P_h \) has a \( q \)th derivative, then (31) is multiplied by \( h^q \). (See also [SN06].) Denote the rescaled objects as \( \tilde{P}_h = h^q P_h \), \( \tilde{s} = h^q s \) and \( \tilde{B}_h = h^q B_h \). We obtain
\begin{align*}
  (\tilde{s} I - \tilde{P}_h) \hat{v} &= \tilde{B}_h \hat{g}.
\end{align*}
If \( \det(\tilde{s} I - \tilde{P}_h) \neq 0 \) for all \( \tilde{s} \) with \( \text{Re} \, \tilde{s} \geq 0 \), then we say that the problem satisfies the determinant condition and the semi-discrete problem is bounded in Laplace space. This is equivalent to saying that there is no eigenvalue \( \tilde{s}_0 \) with \( \text{Re} \, \tilde{s}_0 \geq 0 \), for the eigenvalue problem \( \tilde{P}_h \tilde{v} = \tilde{s}_0 \tilde{v} \). Hence, the transformed solution can be inverted to obtain a bounded solution in the \( x - t \)-space.

The challenge is to determine whether the determinant condition is satisfied or not. If addressed directly, it leads to a complicated algebraic problem that is usually only solvable approximately. Another more appealing route, which we used in [SN06] and that we will use herein, is to infer the determinant condition directly from known stability properties.

The Laplace-transform method for difference schemes, is well-known and found in [GKO95] where it is shown that if there is an eigenvalue with \( \text{Re} \, s \geq 0 \), then the scheme is unconditionally unstable. This is known as the Godunov-Ryabenkii condition. The proof of the Godunov-Ryabenkii condition is also well-known and we briefly give the arguments here. Consider the special case of (30) where \( f \neq 0 \) and \( g = 0 \). If there exists an \( s_0 \), and \( f \), and some \( h_0 \), such that \( v = \exp(s_0 t)f \) satisfies \( (s_0 I - \tilde{P}_h)f = 0 \). Then it follows that \( v = \exp(\tilde{s}_0 t)f \) is also a solution to (30), that grows arbitrarily fast if \( \text{Re} \, \tilde{s}_0 > 0 \).

We make following observation.

Lemma 3.1. If solutions of the semi-discrete scheme (30) are bounded in \( L^2 \), the Godunov-Ryabenkii condition is satisfied. That is, there is no eigenvalue \( \tilde{s}_0 \) of (32) with \( \text{Re} \, \tilde{s}_0 > 0 \).

Proof. If there was an eigenvalue \( \text{Re} \, \tilde{s}_0 > 0 \), the solutions with arbitrarily fast growth would violate the \( L^2 \) bound, which is a contradiction. □

Using the energy estimate even more information on the location of eigenvalues can be deduced.
Lemma 3.2. If solutions of the semi-discrete scheme (30) are bounded in $L^2$, there can be no eigenvalue $\tilde{s}_0 = i\xi_0$, $\xi_0 \neq 0$, to the problem (32).

Proof. Consider (30) with $f \neq 0$ and $g = 0$. If there is an eigenvalue $s_0 = i\xi_0$, $\xi_0 \neq 0$, then there is a solution $\exp(s_0 t)f$ and we can construct a family of solutions, $\exp(\frac{\tilde{s}_0}{h^2}t)f$. They all emanate from the prescribed initial data since $\exp(\frac{\tilde{s}_0}{h^2}t)f|_{t=0} = f$. Given that we consider a well-posed problem satisfying an energy estimate, and a numerical discretization satisfying the analogous estimate, we know that there is a unique smooth (exact) solution $u(t)$ to (29). (Here, $u(t)$ denotes the projection of the continuous $u(x,t)$ on the grid.) Hence, the semi-discrete solutions converge to $u$ in $L^2$ as $h \to 0$. However, at any arbitrary fixed time $t$

$$\|\exp(\frac{\tilde{s}_0}{h^2}t)f(\cdot) - u(t)\|,$$

(33) does not converge to zero if $\xi_0 \neq 0$. We have a contradiction. □

In cases where it can be shown that $\xi_0 = 0$ is not an eigenvalue either, it implies that the determinant condition is satisfied for (30). The determinant condition ensures that the Laplace transformed solution, $\hat{v}$ is well-defined and bounded, such that the inverse Laplace transform (for all $h \geq 0$) is well-defined and results in a bounded solution $v(t)$. (See [GKO95].) We emphasize that it is boundedness of $\hat{v}$ that is the fundamental requirement rather than the determinant condition per se. The latter is only a sufficient condition.

Nevertheless, Lemma 3.2 and 3.1 almost implies that the determinant condition is satisfied. For hyperbolic problems, the last step is taken in the next theorem.

Theorem 3.3. Assume that the problem (29) is well-posed, linear, hyperbolic with Dirichlet boundary conditions and satisfies an energy estimate. Let (30) be an energy stable and nullspace consistent semi-discrete approximation. Then the determinant condition is satisfied.

Proof. Since the determinant condition is a property of the temporal and spatial discretization, we can without restriction consider homogeneous boundary conditions.

Energy stability implies an $L^2$ estimate of the solution. By Lemma 3.2 and 3.1 there are no eigenvalues $\Re \tilde{s}_0 \geq 0$ except possibly $\tilde{s}_0 = 0$, which corresponds to a time-independent nullspace mode. For if there is an eigenvalue $s = 0$ to (30), then $P_h w = 0$ where $w$ is the associated eigenvector. However, any such $w$ is by definition in $N_h$. Since the problem is hyperbolic only the constant can be in the nullspace. Furthermore, since the scheme is energy stable and the boundary conditions are of Dirichlet type, the constant is annihilated from the nullspace. Hence there can be no eigenvalue $\tilde{s}_0 = 0$. □

Remark 3.4. The assumption that the boundary conditions are Dirichlet, excludes periodic boundary conditions for scalar problems. It also excludes boundary conditions for systems where the out-going waves (partly) specifies the in-going waves. (See [GKO95] for a thorough description.) Such boundary conditions resemble periodic boundary conditions and may allow a non-trivial nullspace. (See [CNG99] for an example of a hyperbolic system with a non-trivial nullspace.) We exclude these cases in Theorem 3.3 and 3.8 but Theorem 3.15 covers all cases.

Example 3.5. As an example of the mechanics of the proof of Theorem 3.3, consider the scheme (17). It is straightforward to show that the determinant condition is satisfied using Laplace-transform analysis (see [GKO95]). However, in the proof we infer that the determinant condition is satisfied indirectly. Partly, by Lemma
3.1 and 3.2, and partly by relating the $s = 0$ eigenvalue to the nullspace. In this example, we take a closer look at the latter part.

If $\tilde{s} = 0$ is an eigenvalue, then
\[
\frac{\hat{v}_1 - \hat{v}_0}{h} = -\frac{2}{h}(\hat{v}_0 - \hat{g}),
\]
(34)
\[
\frac{\hat{v}_{i+1} - \hat{v}_{i-1}}{2h} = 0, \quad i = 1...N-1
\]
\[
\frac{\hat{v}_N - \hat{v}_{N-1}}{h} = 0,
\]
should not have a unique solution. With the ansatz $\hat{v}_i = \sigma \kappa^i$ we solve the internal scheme and obtain $\kappa = 1, -1$. Only $\kappa = 1$ satisfies the stencil on the right boundary.

On the left boundary, we have
\[
\hat{v}_1 - \hat{v}_0 = -2(\hat{v}_0 - \hat{g}), \quad \Rightarrow \quad \sigma(\kappa - 1) = -2(\sigma - \hat{g}).
\]
With $\kappa = 1$, we have $\sigma = \hat{g}$ and a unique solution $\hat{v}_i(\tilde{s} = 0) = \hat{g}$. Hence, $\tilde{s} = 0$ is not an eigenvalue.

To relate this result to the nullspace, we observe that if there would have been two solutions to (34), $\hat{v}_1$ and $\hat{v}_2$, their difference is a nullmode. Hence, if we knew that the problem do not support nullmodes, we could rule out non-unique solutions and that $\tilde{s} = 0$ is an eigenvalue, without solving the problem in detail as we did above.

\[\square\]

Remark 3.6. Another equivalent way to address the previous problem is to set $g = 0$ and prove that only the trivial solution exist.

3.1.1. Half-space approximations. A common procedure when analyzing boundary conditions is to divide the problem in a partition of unity and thereby obtaining three simplified problems, one for each boundary and one for the interior. In this procedure, the left boundary problem is discretized by the internal scheme and the left boundary scheme. Consider the (right) half-space version of (17),
\[
\frac{\hat{v}_{i+1} - \hat{v}_{i-1}}{2h} = 0, \quad i = -\infty...N-1
\]
\[
\frac{\hat{v}_N - \hat{v}_{N-1}}{h} = 0.
\]
By the same arguments as before, we now have $\kappa = 1, -1$ and only $\kappa = 1$ satisfies the boundary stencil. Seemingly, $\hat{v}_i = C$ is in the nullspace and $\tilde{s} = 0$ is an eigenvalue and the determinant condition appears not to be satisfied. However, the trail-off condition, $\hat{v}_i \to 0, \quad i \to -\infty$ from the partition of unity, requires that $\hat{v}_i = 0$, so the problem has a unique solution.

We emphasize, that for hyperbolic problems, where a boundary can lack a boundary condition, one has to take both boundaries into account. (Note that the determinant condition is defined for the entire scheme.) In this example, it is the left boundary that removes the $\kappa = 1$ mode from the nullspace.

3.1.2. Higher derivatives in space and nullspace removal. For problems with higher derivatives in space, there may be a non-trivial nullspace mode, which implies that the naïve interpretation of the determinant condition is not satisfied since there will be an eigenvalue $\tilde{s} = 0$. Yet, we have already seen that such modes are essentially unaffected by the PDE/scheme. Hence, they should not affect stability. In the following, we shall show that this indeed is the case since the nullspace modes are determined a priori, and can be removed from the scheme during the solution process, and added again a posteriori.
Consider a well-posed and energy stable problem with a non-trivial nullspace. To reduce notation, we assume that the nullspace is one-dimensional. Since the nullspace is unaffected in homogeneous problems, the first step is to recast the PDE such that we get homogeneous boundary conditions. To this end, introduce a smooth function $\Psi$ that satisfies the boundary conditions. We make a change of variables, $u = v + \Psi$ and recast (29) into

$$v_t + P v = -\Psi_t - P \Psi = F(x, t),$$

(35) \{homogenized\} \quad L_0 v(0, t) = 0, \\
\quad L_1 v(1, t) = 0, \\
\quad v(x) = f(x) - \Psi(x, 0) = \tilde{f}(x).$$

This problem is discretized by

$$v_i = P_h v + F,$$

(36) \{semidisc2\} \quad v(0) = \tilde{f},$$

where $F_i(t) = F(x_i, t)$ and $\tilde{f}_i = \tilde{f}(x_i)$. (C.f. (30) with $g = 0$.)

Next, we divide the solution of (36) into one part that is unaffected by the PDE and one part that is affected. Assume now that $w$ is in the (one-dimensional) spatial nullspace, $N_h$, and normalized such that $\|w\| = 1$. Split $v = w\xi(t) + v'$ where $\xi(t)$ represents the unknown time evolution of the spatial nullspace and $v' \in N_h^\perp$. Moreover, introduce $F^N = w(w, F) \in N_h$ and split $F = F^N + F'$ where $F' \in N_h^\perp$. This allows us to divide the problem into the following two orthogonal equations.

$$w\xi_t = P_h(w\xi) + F^N = F^N, \quad \xi(0)w = (w, \tilde{f})w$$

(37) \{ode\} \\
$$v'_i = P_h v' + F', \quad v'(0) = \tilde{f} - \xi(0)w$$

(38) \{homogeneous\}

where we have used that $P_h(w\xi) = 0$ in (37), which follows from $w \in N_h$. From (36), we obtain $\xi(0) = (w, \tilde{f})$ in (37). Equation (37) is simply an ordinary differential equation in $\xi(t)$ that can be explicitly integrated since $F$ (and hence $F^N$) is a known function of $t$. Hence, it governs the time evolution of the nullspace of (36).

The second equation has homogeneous boundary data and by Prop. 2.23 neither the forcing function, $F'$, nor $P_h v'$ and nor $v'(0)$, has a component in the $w$ direction. Hence, the time change of $v'$ in the $w$ direction is zero, and since it is zero initially it will remain so for all time. Consequently, the number of vector components of $v'$ is one more than the number of dimensions $v'$ spans. Furthermore, since any stencil in the scheme applied on the nullspace mode results in zero, and since the stencils are not all global, the nullspace mode has to involve all components of $v'$.

(That is, it can not be spanned by a subset of the components of $v'$. $v'$ is a global (smooth) grid function.)

This implies that we can solve for one component of $v'$, say $v'_0$, using the relation $(w, v') = 0$. That is, $v'_0$ is a linear combination of $v'_i$, $i = 1,...N$. Therefore, $v'_0$ is redundant and we can remove the first equation in the semi-discrete scheme that evolves $v'_0$, i.e., the first row and column of $P_h$. We denote the reduced operator $P'_h$ and obtain,

$$v'_i = P'_h v'' + F'', \quad \{reduced\}$$

(39)

where $v''$ and $F''$ are obtained by removing the first element of $v'$ and $F'$. (This procedure is equivalent to removing the rows and columns associated with 0 eigenvalues in (20).)
Based on the reduction process described above, we make the following definition.

**Definition 3.7.** Assume that the scheme (30) has a non-trivial nullspace. Let (39) denote the reduced system where the (a priori known and bounded) nullspace modes have been removed from the unknowns. Then, we say that, (30) satisfies a generalized determinant condition, if $\det((\tilde{s}I' - \tilde{P}'_h)) \neq 0$ for all $\text{Re} \tilde{s} \geq 0$.

To obtain (39), we have merely rewritten the scheme. The solution $v''$ inherits all the stability properties from $v$. If the scheme (30) is nullspace consistent and energy stable, the only eigenvalues with $\text{Re} \tilde{s} = 0$ that could exist are (possibly multiple) $\tilde{s} = 0$, but those have now been removed from (39). Hence, the determinant condition must be satisfied for (39). Consequently, the generalized determinant condition is satisfied for the original scheme.

The previous discussion is directly applicable to parabolic problems, whose nullspaces are at most one-dimensional.

**Theorem 3.8.** Assume that the problem (29) is well-posed, linear, parabolic and satisfies an energy estimate. Let (30) be an energy stable and nullspace consistent, semi-discretization. Then the generalized determinant condition is satisfied if the boundary conditions are of Neumann- or Robin-type. The determinant condition is satisfied for Dirichlet boundary conditions.

**Proof.** Energy stability implies an $L^2$ estimate of the solution. By Lemma 3.2 and 3.1 there are no eigenvalues $\text{Re} \tilde{s}_0 \geq 0$ except possibly $\tilde{s}_0 = 0$. Energy stability and parabolicity gives the following options for boundary conditions: Dirichlet, Neumann or Robin (mixed). In the case of Dirichlet, the linear function is annihilated from the nullspace and the problem satisfies the determinant condition.

In the case of Neumann conditions, the constant solution is part of the nullspace. For Robin boundary conditions a linear function may span the nullspace. In either case, the problem takes the form (30) and we need to consider the details of the Laplace transform analysis. We reduce the problem by carrying out the steps (37) to (39). Having done so, we have merely rewritten the scheme so it has not changed its stability properties. However, the nullspace of (39) is now trivial. Since the solution is bounded in $L^2$ and there are no eigenvalues $\tilde{s} = 0$, the determinant condition is satisfied for the reduced problem (39). \qed

In the general case with a $p$th-order spatial derivative, we can have several nullspace modes. Say that we have two such nodes. These would be 1 and $x$, which would be the case if the boundary conditions had no lower derivatives than $u_{xx}$. (There could be additional boundary conditions but with higher derivatives.) In any case, it is still possible to choose a function $\Psi$ that satisfies all the boundary conditions, and thereby transform the problem to a homogeneous counterpart. This time neither 1 nor $x$ would be affected by the scheme but instead be determined by the resulting forcing function. Choosing orthogonal basis functions, $1, 1/2 - x$ (on $[0, 1]$) and by splitting the solution as $u = \xi_1(t) + (1/2 - x)\xi_2(t) + u'$, we obtain two ODEs for $\xi_1$ and $\xi_2$ respectively. In $u'$, the nullspace orthogonal part, the modes $1, 1/2 - x$ would remain constant and identically zero, and we can remove two rows and columns from the semi-discrete system to make it non-singular.

The procedure is completely general.

**Theorem 3.9.** Assume that the problem (29), with a $p$th-order spatial derivative, is well-posed, linear, and satisfies an energy estimate. Let (30) be an energy stable and nullspace consistent semi-discretization. Then the generalized determinant condition is satisfied. (If the (spatial) nullspace is trivial, the determinant condition is satisfied.)
Proof. The proof is analogous to the previous one. Energy stability implies an $L^2$ estimate of the solution. By Lemma 3.2 and 3.1 there are no eigenvalues $\Re \tilde{s}_0 \geq 0$ except possibly $\tilde{s}_0 = 0$. By energy stability, the boundary conditions can only contain derivatives of orders less than $p$ and there can be at most $p-1$ nullspace modes. These are determined by the initial data and evolved by a separate ODE (see (37)), with no influence from, or on, the spatial operator. Furthermore, the operator $P_h$ has an eigenvalue zero of multiplicity (at most) $p-1$. We can use the predetermined nullspace modes to reduce the number of unknowns and obtain a discretization,

$$v''_t + P_h'v'' = F''$$

which inherits the stability properties of the original system, i.e., $v''$ satisfies all the same estimates as $v$. This system has no eigenvalue equal to zero and satisfies a determinant condition. 

In summary, if a scheme with a first derivative in time is energy stable and nullspace consistent, the determinant condition is satisfied for a reduced system where the nullspace has been removed.

3.2. Systems. The generalization of the results in the previous section to hyperbolic and parabolic systems, or higher-order principal parts is straightforward. The arguments regarding the connection between energy stability and eigenvalues are the same. However, the different equations that form the system, need not have the same character. One such example, is the compressible Navier-Stokes equations which is incompletely parabolic. That is, one equation is hyperbolic and the others are parabolic. To reduce notation we settle with one hyperbolic and one parabolic equation and follow the derivation in [SN06].

Consider the incompletely parabolic problem

$$\{\text{incomp}\} \quad \left(\begin{array}{c} u \\ v \end{array}\right)_t + \left(\begin{array}{cc} a_{11} & a_{12} \\ a_{12} & a_{22} \end{array}\right) \left(\begin{array}{c} u \\ v \end{array}\right)_x = \left(\begin{array}{c} 0 \\ \epsilon v \end{array}\right)_x, \quad 0 < x < 1,$$

with appropriate initial and boundary conditions. We assume that the system is symmetric which is a necessary requirement for energy stability. Since it is symmetric we can choose to diagonalize either the hyperbolic or the parabolic part. Hence, for two equations, the system (40) represents a completely general example.

The semi-discrete approximation is,

$$\{\text{incomp\_semi}\} \quad \left(\begin{array}{c} u \\ v \end{array}\right)_t + \left(\begin{array}{cc} a_{11}I & a_{12}I \\ a_{12}I & a_{22}I \end{array}\right) \left(\begin{array}{c} u \\ Dv \end{array}\right) = \left(\begin{array}{c} 0 \\ \epsilon Dv \end{array}\right) + \left(\begin{array}{c} B_{h1}g_1 \\ B_{h2}g_2 \end{array}\right).$$

As usual, we assume that the scheme is energy stable and nullspace consistent. (Even if the same SBP operator is used to approximate the first derivatives of $u$ and $v$, the operators $D^u_1$ and $D^v_1$ may differ since they include the boundary treatment.) Furthermore, we note that if the problem is energy stable then by necessity, the problem is also energy stable with $a_{12} = 0$, i.e., if the problem decouples. Hence, by Theorem 3.3,

$$\{\text{hyp\_part}\} \quad \left(\tilde{s}I + a_{11}\tilde{D}^u_1\right)u = 0, \quad \tilde{s} = sh$$

can not have any eigenvalues $\Re \tilde{s} \geq 0$. That is, the determinant condition is satisfied. Specifically this means that $(\tilde{s}I + a_{11}\tilde{D}^u_1)^{-1}$ exists.

Laplace transformation of (41) leads to:

$$\begin{pmatrix} sI + a_{11}D_1^u & a_{12}D_1^v \\ a_{12}D_1^u & sI + a_{22}D_1^v - \epsilon D_2 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \begin{pmatrix} B_{h1}\tilde{g}_1 \\ B_{h2}\tilde{g}_2 \end{pmatrix}.$$
Multiplying the first equation by \( h \) and the second by \( h^2 \) yields,

\[
\begin{align*}
\{ \text{incomp-sen2} \} \quad & \quad \begin{pmatrix} \hat{s} I + a_{11} \hat{D}_t^u & a_{12} \hat{D}_t^v \\ -ha_{12} \hat{D}_t^u & s I + ha_{22} \hat{D}_t^v - \epsilon \hat{D}_2 \end{pmatrix} \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} h B^u_1 g_1 \\ h^2 B^v_1 g_2 \end{pmatrix},
\end{align*}
\]

where \( \hat{s} = h^2 s \) and \( \hat{s} = hs \). Note that these are not independent variables. In particular, when \( h = 0 \), both are 0.

First, consider the case when \( Re \hat{s}, \hat{s} \geq 0 \) excluding the case \( \hat{s} = \hat{s} = 0 \). We employ the usual energy arguments to rule out any other singularities. Next, we turn to the critical case when \( \hat{s} = \hat{s} = 0 \). If the nullspace is trivial there can be no such eigenvalue and the determinant condition is satisfied. (To connect the Laplace-transform analysis with the nullspace, we proceed with the assumption that \( g_1 = g_2 = 0 \). This is not a restriction as discussed earlier.)

On the other hand, if the nullspace is non-trivial, we proceed and investigate the nature of the singularity. To this end, we set \( \hat{s} = 0 \) in (43) and consider the first equation,

\[
(44) \quad a_{11} \hat{D}_t^u \dot{u} + a_{12} \hat{D}_t^v \dot{v} = 0.
\]

We have already shown that \( \hat{s} I + a_{11} \hat{D}_t^u \) is invertible for \( Re \hat{s} \geq 0 \). In particular, when \( \hat{s} = 0 \) it implies that \( \hat{D}_t^u \) is invertible and that the nullspace of \( \hat{D}_t^u \) is trivial.

(See the discussion on (42) above.) Hence, a nullspace mode of (44) must satisfy

\[
\dot{u} = -\frac{a_{12}}{a_{11}} (\hat{D}_t^v)^{-1} \hat{D}_t^u \dot{v}.
\]

Inserting this relation in the second equation gives,

\[
h(-\frac{a_{12}}{a_{11}} + a_{22}) \hat{D}_t^v \dot{v} - \epsilon \hat{D}_2 \dot{v} = 0.
\]

If this system is singular we can remove the nullspace mode by the standard procedure discussed above. The remaining system satisfies the determinant condition. Consequently, \( \dot{v} \) is uniquely determined and thereby \( \dot{u} \) as well. We conclude that the problem satisfies the generalized determinant condition.

3.3. Higher derivatives in time. We begin by proving a stability property.

**Lemma 3.10.** An energy-stable scheme (2) bounds all \( \| v_{j,t} \| \), \( j = 0, \ldots, k - 1 \).

**Proof.** By definition of energy stability \( \| v_{(k-1)t} \| \) is bounded. Also by definition, \( \partial_t v_{(k-2)t} = v_{(k-1)t} \). Taking the inner product with \( v_{k-2} \) gives

\[
v_{(k-2)t} H(\partial_t v_{(k-2)t}) = v_{(k-2)t} H v_{(k-1)t}.
\]

\[
\partial_t \frac{1}{2} \| v_{(k-2)t} \|^2 \leq \| v_{(k-2)t} \| \| v_{(k-1)t} \|
\]

Since \( v_{(k-1)t} \in L^\infty(0,T; L^2(\Omega)) \) from the energy estimate, \( C = \max_{t \in [0,T]} \| v_{(k-1)t}(t) \| \) is bounded. Therefore, we can proceed and integrate in time

\[
\| v_{(k-2)t}(t) \| - \| v_{(k-2)t}(0) \| \leq CT.
\]

Noting that \( v_{(k-2)t}(0) = f_{k-2} \in L^2(\Omega) \), we have \( v_{(k-2)t}(t) \in L^2(\Omega) \) for all \( t \in [0,T] \). Repeating the process for the lower time derivatives yields all bounds. \( \square \)

Consider the problem

\[
\begin{align*}
v_{tt} & = P_h v + B_h g, \\
v(0) & = f, \\
v_t(0) & = h,
\end{align*}
\]

\{sec:high_time\}

\{lemma:time_est\}
where the highest spatial derivative is \( q \geq 1 \). Energy stability implies that \( \| v_t \| \) is directly bounded by the energy estimate (See Definition 2.4) and by Lemma 3.10 also \( \| v(t) \| \) is bounded in \( L^2 \). We proceed to study the eigenvalues of (46) and use similar arguments as in the case of a first-order derivative in time to exclude all eigenvalues with \( \Re \tilde{s} > 0 \) and \( \tilde{s} = i\xi \) with \( \xi > 0 \).

We begin by showing that there are no eigenvalues \( \Re \tilde{s} > 0 \). As before we assume there is an eigenvalue \( s_0 \) for some \( h_0 \) such that \( v(t) = \exp(s_0 t) \mathbf{f} \) solves the homogeneous equation, i.e. (46) with \( g = 0 \). That is,
\[
\tilde{s}_0^2 \mathbf{f} = P_h \mathbf{f}, \quad \text{or,} \quad \tilde{s}_0^2 \mathbf{f} = \tilde{P}_h \mathbf{f},
\]
where \( \tilde{s}_0 = s_0 h_0^{q/2} \) and \( \tilde{P}_h = h^q P_h \) such that the elements of \( \tilde{P}_h \) are \( O(1) \).

Given that (46) with \( g = 0 \) is satisfied, we note that \( v = h^{q/2} \exp(s_0 t/h^{q/2}) \mathbf{f} \) is also a solution of (46) with \( g = 0 \). It is well defined, since its initial data is bounded and given by \( v(0) = h^{q/2} \mathbf{f} \) and \( v_t(0) = \tilde{s}_0 \mathbf{f} \). However, \( v_t = \tilde{s}_0 \exp(s_0 t/h^{q/2}) \mathbf{f} \) grows arbitrarily fast (from \( L^2 \) bounded initial data) as \( h \to 0 \) and hence it violates energy stability.

**Remark 3.11.** Note that it is the time-derivative of the constructed solution of (46) that grows unboundedly, not the solution as is the case in the analog argument for an equation with a first derivative in time.

Next, we turn to the possibility of purely imaginary eigenvalues, \( s_0 = i\xi_0, \xi_0 \neq 0 \). Once again we assume that there is one such eigenvalue for some \( h = h_0 \). Since the entries in the matrix \( P_h \) are real, its complex eigenvalues will always occur in conjugate pairs. Hence, there is also an eigenvalue \( -s_0 = -i\xi_0 \). Consequently, we assume that there is a solution \( v = \cos(\xi_0 t) \mathbf{f} \) that solves (46) with \( g = 0 \):
\[
v_t = P_h \mathbf{f} \quad \Rightarrow \quad -\xi_0^2 \cos(\xi_0 t) \mathbf{f} = P_h \cos(\xi_0 t) \mathbf{f} \quad \Rightarrow \quad (47) \quad \{ \text{eig wave2} \} \quad -\xi_0^2 \cos(\xi_0 t) \mathbf{f} = \tilde{P}_h \cos(\xi_0 t) \mathbf{f},
\]
where the last step is achieved by multiplying by \( h_0^q \) and using \( \tilde{s}_0 = h_0^{q/2} \xi_0 \). Next, we observe that \( v(t) = \cos(\xi_0 t/h^{q/2}) \mathbf{f} \) results in
\[
-\frac{\tilde{s}_0^2}{h^q} \cos(\frac{\tilde{s}_0}{h^{q/2}} t) \mathbf{f} = P_h \cos(\frac{\tilde{s}_0}{h^{q/2}} t) \mathbf{f}, \quad \Rightarrow \quad -\xi_0^2 \cos(\xi_0 t/h^{q/2}) \mathbf{f} = \tilde{P}_h \cos(\xi_0 t/h^{q/2}) \mathbf{f}.
\]
That is, it is equivalent to (47) and hence a solution to (46). The initial data of this solution are given by: \( v(0) = \mathbf{f}, v_t(0) = 0 \). That is, they are independent of \( h \) and bounded in \( L^2 \). We know that the scheme is stable and consistent, i.e., convergent to an exact solution \( \mathbf{u} \), and yet \( \| v(t) - \mathbf{u} \| = \| \cos(\frac{\tilde{s}_0}{h^{q/2}} t) \mathbf{f} - \mathbf{u} \| \) is not converging since at any given time \( t \), \( \mathbf{u}(t) \) is a fixed function in space and \( v(t) \) is oscillating between 0 and \( \pm \mathbf{f} \) as \( h \to 0 \). Hence, we have a contradiction. We have proven the following Lemma.

**Lemma 3.12.** If the scheme (46) is energy stable and nullspace consistent, there is no eigenvalue \( \tilde{s} \) with \( \Re \tilde{s} > 0 \) or \( \tilde{s} = i\xi, \xi \neq 0 \).

**Remark 3.13.** In the preceding argument, we use the fact that energy stability implies an \( L^2 \) convergence rate of \( \min(r + 1/2, p) \). Hence, the method need not be consistent near the boundary. Nullspace consistency with \( r = 0 \) suffices.

What is left is the possibility of an eigenvalue \( \tilde{s} = 0 \), which coincides with the nullspace. (The following arguments are analogous to those in the proof of
Theorem 3.8.) The first conclusion is that if the boundary conditions annihilate all nullspace modes, then the determinant condition is satisfied. Otherwise, the spatial nullspace consists of \( w_i \), \( i = 1 \ldots n_N \) where \( n_N \) is the dimension of the nullspace, and the complete nullspace contains \( \alpha_i w_i + t_i \beta_i w_i \). According to Theorem 2.26, the \( \alpha_i \) and \( \beta_i \) are a priori determined by the initial conditions and thereafter unaffected by the scheme for problems with homogeneous boundary conditions. Hence, we recast the semi-discrete problem into one with homogeneous boundary conditions and remove unknowns (using the nullspace modes) from the scheme to obtain a reduced scheme. (See (35) and the subsequent discussion.)

The reduced problem has a trivial nullspace and is energy stable. Hence, the determinant condition for the reduced problem is satisfied, and the generalized determinant condition is satisfied for the original problem. We summarize the results in the following theorem.

**Theorem 3.14.** If the scheme (46) is energy stable and nullspace consistent, the generalized determinant condition is satisfied. (If the nullspace is trivial, the determinant condition is satisfied.)

We can generalize this to higher temporal derivatives.

**Theorem 3.15.** Assume that the problem (1) has a \( q \)th-order principal part and that the problem is well-posed, linear, and satisfies an energy estimate. Furthermore, we assume that its semi-discrete approximation (2) is nullspace consistent and energy stable. Then:

1. There can be no eigenvalues in the right-half plane. (Godunov-Ryabenkii)
2. If the nullspace is trivial, the determinant condition is satisfied.
3. If the nullspace is non-trivial, the generalized determinant condition is satisfied.

**Proof.** The proof is analogous to the previous arguments. Assuming that, there is an eigenvalue \( s_0 \) then \( v(x) = \exp(s_0 t)f \) is a solution of the homogeneous equation, we have

\[
\tilde{s}_0^j f = \tilde{P}_h f, \quad \text{or} \quad \tilde{s}_0^j f = \tilde{\tilde{P}}_h f,
\]

where \( \tilde{s}_0 = s_0 h^{q/k} \). Then, \( v(x) = h^q(s_0)^j \exp(\frac{\tilde{s}_0 j}{k^{q/k}} t)f \) is another solution (with bounded initial data \( f_j = h^q(s_0)^j \exp(\frac{\tilde{s}_0 j}{k^{q/k}}) f, j = 0 \ldots k - 1 \).) In particular, \( f_{k-1} = (\tilde{s}_0)^{k-1} f \) is bounded and non-vanishing and \( v_{(k-1)t} = (\tilde{s}_0)^{k-1} \exp(\frac{j \tilde{s}_0}{k^{q/k}} t)f \) grows arbitrarily fast as \( h \to 0 \), violating the energy estimate.

Similarly, for \( s_0 = i\xi_0 \) and we can construct a solution that will not converge as \( h \to 0 \), contradicting the convergence theory inferred by the energy estimate. Hence, there can be no \( \Re s, \tilde{s} \geq 0 \) that makes the problem singular apart from possibly \( s, \tilde{s} = 0 \).

Analogously, \( \tilde{s} = 0 \) corresponds to nullspace modes. If the nullspace is trivial, there can be no eigenvalue \( \tilde{s} = 0 \). Hence, the determinant condition is satisfied.

If the nullspace of \( P_h \) is non-trivial, we homogenize the system (see equation (35) and onwards). One part, (37), evolves the nullspace modes independently of the PDE. The other part, (38), has no energy in the nullspace modes and we can reduce the system with as many unknowns as the number of nullspace modes. We are left with a system with a trivial nullspace and the reduced part of the solution can be uniquely determined, i.e., the generalized determinant condition is satisfied.

\[ \square \]

4. Convergence rates

In this Section, we address the convergence rates of energy-stable and nullspace-consistent schemes. To this end, we use the Laplace-transform technique and rely
on the stability results in the previous section. The reason that we need the Laplace transform is to deduce sharper convergence rates than obtainable directly from the energy method. The underlying theory is found in [GKO95] and we also refer to the seminal papers [Gus75, Gus81]. We will extend this theory to cases when the determinant condition is not satisfied due to a non-trivial nullspace.

Consider (1) semi-discretized by (2). We assume that (1) is well-posed such that a smooth solution exists. By projecting the smooth solution, \( u \), onto the grid (denoted \( u \)), it can be inserted in the scheme (2). The solution \( u \) will not satisfy (2) and the residual is the truncation error that enters the error equation as a forcing function. It follows that the error \( e = u - v \) satisfies

\[
e_{kt} = P_h e + T, \tag{48}
\]

\[
e_{jt}(0) = 0 \quad j = 0, ..., k - 1.
\]

Note that \( e \) satisfies homogeneous boundary conditions, modulo truncation errors that we incorporate in \( T \). Furthermore, the truncation error is given by the spatial discretization as,

\[
T_i = [Pu](x_i) - (P_h u)_i.
\]

We divide the truncation error as, \( T = T_i + T_b \) where \( T_b \) is zero except at the rows corresponding to the boundary schemes and hence, \( T_i \) is zero at the boundary positions and non-zero elsewhere.

**Remark 4.1.** We could allow that the initial data is satisfied to within the order of accuracy. However, the above is more natural in practice and reduces notation.

We summarize all the necessary assumptions.

**Assumption 4.2.** We assume:

- Smoothness and compatibility of data, such that the exact solution is smooth.
- The highest derivative in \( P \) is of order \( q \).
- The problem (1) satisfies an energy estimate.
- The scheme (2) is energy stable.
- The scheme (2) is nullspace consistent.
- \( T_b \sim O(h^r) \) and \( T_i \sim O(h^p) \), where \( r < p \)
- The discrete inner product is a \( p \)th-order quadrature rule.

The convergence rate is deduced by investigating the rate by which \( e \) approaches zero. We follow the usual procedure and split \( e = e_i + e_b \). Since the problem is linear, we can divide it as

\[
(e_b)_{kt} = P_h e_b + T_b, \tag{49}
\]

\[
(e_b)_{jt}(0) = 0, \quad j = 0, ..., k
\]

and the analogous problem for \( e_i \).

**Lemma 4.3.** A necessary and sufficient requirement for convergence of \( e_i \), with arbitrary bounded and smooth data, is that \( p \geq 1 \). Then the convergence rate of \( e_i \) is \( p \).

**Proof.** First, we prove sufficiency by considering \( (e_i)_{kt} = P_h e_i + T_i \). If \( p \geq 1 \), then \( (e_i)_{(k-1)\ell} \) converges with the rate \( p \). This follows from energy stability, which implies that \( \| (e_i)_{(k-1)\ell} (\ell) \| \leq \| T_i \| = O(h^p) \).

Repeating the derivations leading to (45) for the error equation leads to

\[
\partial_t \| (e_i)_{(k-2)\ell} \| \leq \| (e_i)_{(k-1)\ell} \| \leq \| T_i \|,
\]
and hence
\[ \| (e_i)^{(k-2)t}(t) \| - \| (e_i)^{(k-2)t}(0) \| \leq \int_0^T \| T_i \| \, dt. \]
Since \((e_i)^{(k-2)t}(0) = 0\), we obtain \(\| (e_i)^{(k-2)t}(t) \| = O(h^p)\). Repeating the procedure leads to \(\| (e_i)(t) \| = O(h^p)\).

Necessity follows from our basic assumption in Section 2.3 that \(p\) is an integer. Hence, if \(p \geq 1\) is not satisfied, \(p = 0\) or less. That is, the scheme is inconsistent in which case it need not converge. \(\square\)

**Example 4.4.** It is easy to construct a non-convergent scheme by violating consistency. For instance, by adding a constant to some of the coefficients of a convergent energy stable scheme thereby making it inconsistent (but still stable) and clearly not convergent. An example of this is to change (17) to
\[ (v_i)_t + \frac{v_{i+1} + ch v_i - v_{i-1}}{2h} = 0, \quad c = \text{constant} \]
which is energy stable and not convergent to solutions of \(u_t + u_x = 0\). \(\square\)

**Remark 4.5.** Note that Lemma 4.3 only states that consistency is a necessity for the internal stencil. It does not concern the boundary stencils.

Before we address the boundary errors, we need two auxiliary results.

**Lemma 4.6.** Assume that (3) semi-discretized by (4) satisfy Assumptions 4.2. Then for all \(w \in \mathcal{N}\), and the corresponding mode \(w \in \mathcal{N}_h\)
\[ \int_0^1 wPu \, dx = w^T HP_h u = 0. \]
where \(u\) is the projection of a smooth function \(u(x,t)\) on the grid. (That is, \(u\) need not be the semi-discrete solution, and nor need \(u(x,t)\) be the solution of a PDE.)

**Proof.** We prove the continuous relation. Split the function \(u\) as \(u = u_N + u_\perp\), where \(u_N \in \mathcal{N}\) and \(u_\perp \in \mathcal{N}_\perp\).
\[ \int_0^1 wPu \, dx = \int_0^1 wPu_N + \int_0^1 wPu_\perp \, dx = \int_0^1 wPu_\perp \, dx \]
since \(Pu_N = 0\) by definition of the nullspace. Furthermore, \(Pu_\perp = v \in \mathcal{N}_\perp\). By orthogonality of \(\mathcal{N}\) and \(\mathcal{N}_\perp\), it follows that the inner product
\[ \int_0^1 u_\perp v \, dx = 0. \]
The discrete proof follows is analogous. \(\square\)

**Example 4.7.** We exemplify Lemma 4.6 when \(u\) is the solution of the wave equation with homogeneous Neumann boundary condition \((u_x([0,1],t) = g([0,1]) = 0)\). Then, for the spatial operator, we have,
\[ \int_0^1 1 \cdot Pu \, dx = \int_0^1 1 \cdot u_{xx} \, dx = u_x |_{x=0}^1 = 0, \]
and we conclude that the (spatial) nullspace mode is \(w = 1\). Similarly, the semi-discrete scheme is \(u_t = H^{-1}(-\bar{A} + BS)u - H^{-1}(BSu - Bg)\) where \(g = (g_0, 0, ..., 0, g_1)^T = 0\) and \(w = 1 \in \mathcal{N}\). Furthermore,
\[ 1Hu_t = 1((-\bar{A} + BS)u - (BSu)) = 0 \]
Note that \(1\bar{A}u = 0\) since \(1\) is a nullspace mode of the symmetric matrix \(\bar{A}\). \(\square\)
Lemma 4.8. Consider (1) semi-discretized by (2). If the Assumptions 4.2 are met, then the truncation error in (48) satisfies,

$$w^THT = O(h^p).$$

for all $w \in N$.

Proof. We begin by transforming the problem into one with homogeneous boundary conditions. That is, we choose a smooth function, $\Phi$, that satisfies the boundary conditions and introduce $\chi = u + \Phi$. Then $\chi$ satisfies

$$(50) \quad \{ibvp_hom\} \chi_{kt} = P\chi + F(x,t), \quad 0 < x < 1, \quad 0 < t \leq T, \quad k \geq 1,$$

and the semi-discretization becomes

$$\{ibvp_homsemi\} v_{kt} = P_h v + F,$$

$$v_{jt}(0) = \tilde{f}_j(x), \quad j = 0, ..., k - 1.$$

The truncation error is obtained by first projecting the exact solution $\chi(x,t)$ onto the grid, here denoted $\chi_h$. From (50), we have the identity

$$(\chi(x_i,t))_{kt} = (P\chi)(x_i,t) + F(x_i,t).$$

Since it is now a grid function, we can apply the spatial scheme on $\chi$ and deduce

$$(\chi(x_i,t))_{kt} = (P_h\chi)_i + [(P\chi)(x_i,t) - (P_h\chi)_i] + F_i = (P_h\chi)_i + T_i + F_i$$

where $T_i$ is the component of the truncation error at $x_i$. (The error $e = v - \chi$ satisfies $e_{kt} = P_h e + T$.)

Next, we use the assumption that the inner product is a $p$th-order quadrature. That is, for any smooth functions $u, w$ projected onto the grid,

$$\int_0^1 w(x)u(x,t)dx = w^T H u + O(h^p).$$

(In [HZ13] it was proven that the $H$ matrix appearing in SBP schemes is a $p$th-order accurate quadrature rule.)

With $w \in N$ and using the quadrature relation, we compute

$$w^THT = w^T H [(P\chi) - P_h\chi] = \int_0^1 w(P\chi)dx + O(h^p) - w^T H(P_h\chi).$$

By Lemma 4.6, we conclude that $w^THT = O(h^p)$. \qed

Now we are ready to state the main result of this article.

Theorem 4.9. If Assumptions 4.2 hold, the convergence rate of $e_b$ in (49) is:

- $r + q$ if the nullspace of $P_h$ is trivial.
- $\min(p, r + q)$ if the nullspace of $P_h$ is non-trivial.

Proof. We begin with the case with no nullspace modes. Laplace transforming (49) yields,

$$\tilde{s}^k \tilde{e}_b = h^q P_h \tilde{e}_b + h^q \tilde{T}_b$$

where $\tilde{s} = sh^{q/k}$ and the entries of $h^q P_h \sim O(1)$. \{theo:main\}
This is solved as a difference equation in the interior yielding solutions of the form \( \kappa_j \), where \( j \) is the grid index. Hence, the internal solution is

\[
\{eb\_sol\} \quad (\dot{e}_b)_j = \sum_{i=1}^{z-1} \sigma_i \kappa_i^j,
\]

where \( z \) is the width of the internal stencil. The \( \sigma_i \) are unknowns constants that will be determined by the boundary stencils. The ones associated with \( |\kappa_i| < 1 \) (i.e., half of the \( \kappa \)'s) are determined by the boundary stencils at the left boundary and the other half by the right boundary. Since the scheme is assumed to be (energy) stable, there has to be at least as many boundary stencils as \( \sigma \); else the solution would be undetermined rendering the scheme unstable. However, there may be more boundary stencils that require extra unknowns \( \tau \) at the boundary.

To determine the unknowns, the solution (52) is inserted in the boundary stencils (possibly along with the extra unknowns). We denote the vector of unknowns as \( \sigma = (\sigma_1...\sigma_{z-1}, \tau_1...\tau) \) which results in

\[
C(\hat{s})\sigma = h^q \hat{T}_b',
\]

where \( \hat{T}_b' \) is the reduced (shortened) vector that only contains the non-zero elements. By Theorem 3.15, \( C(\hat{s}) \) is invertible and \( \sigma \sim h^q \hat{T}_b' = \mathcal{O}(h^{r+q}) \). We can transform the solution back to the \( x-t \) plane and obtain \( ||e_b|| < \mathcal{O}(h^{r+q}) \).

Turning to the case with a non-trivial nullspace, we make an observation. Although the nullspace modes are unaffected by \( P_b \), they might still be “fed with energy” since \( T_b \) generally does not lack the nullspace mode. Taking a few steps back, the error equation for the boundary error is:

\[
(53) \quad (e_b)_kt = P_b e_b + T_b.
\]

We split the error into the nullspace part \( e_N \) and the orthogonal complement \( e_\perp \), such that \( e_b = e_N + e_\perp \). Furthermore, we let \( w_j, \ j = 1...n \), denote an orthonormal basis of \( \mathcal{N}_b \) and split the truncation error as \( T_b = T_b^N + T_b^\perp \), where \( T_b^N = \sum_{j=1}^n w_j (w_j, T_b) \). In analogy with (36) to (38), we obtain

\[
(54) \quad (e_N + e_\perp)_kt = P_b (e_N + e_\perp) + (T_b^N + T_b^\perp).
\]

leading to,

\[
(55) \quad (e_N)_kt = P_b e_N + T_b^N.
\]

\[
(56) \quad (e_\perp)_kt = P_b e_\perp + T_b^\perp.
\]

By the definition of the spatial nullspace, \( P_b \) preserves this structure in the sense that \( P_b e_\perp \in \mathcal{N}^\perp \) and \( P_b e_N = 0 \). (See Prop. 2.23.) This reduces (55) to

\[
(57) \quad (e_N)_kt = T_b^N.
\]

The next step is to find the \( L^2 \) evolution of \( e_N \). To this end, we take the inner product of one of the orthonormal functions \( w_j \in \mathcal{N}_b \) and (57), which results in

\[
(58) \quad (w_j, e_N)_kt = (w_j, (e_N)_kt) = (w_j, T_b^N) = (w_j, T_b).
\]

where in the last step we have added \( (w_j, T_b^\perp) = 0 \). By using Lemma 4.8, we can now calculate \( w_j^T H T_b \) in (58). We have \( \mathcal{O}(h^q) = w_j^T H T = w_j^T H T_i + w_j^T H T_b \). Since the elements of \( T_i = \mathcal{O}(h^p) \), we have \( w_j^T H T_i = \mathcal{O}(h^p) \) and consequently also \( w_j^T H T_b = \mathcal{O}(h^p) \).
Next, we write $e_N = \sum_{j=1}^{n} \alpha_j(t)w_j$ and use the short-hand $(\alpha_j)_{(k-1)t} = \beta_j(t)$. Then

$$\frac{1}{2}(\|e_N\|^2)_{kt} = ((e_N)_{(k-1)t},(e_N)_{(k-1)t})_t = \sum_{j=1}^{n} (\beta_j,w_j) = \sum_{j=1}^{n} \beta_j(w_j,T_b) \sim \sum_{j=1}^{n} \beta_jO(h^p).$$

where we have used that the $w_j$'s are orthonormal. For the same reason, we have

$$\beta_j(\beta_j)_t = \beta_jO(h^p), \quad j = 1...n,$$

leading to

$$\int_0^T (\beta_j)_t dt = \beta_j(t) - \beta_j(0) = \int_0^T O(h^p) dt, \quad j = 1...n.$$

Clearly $\beta_j \sim O(h^p)$ since $\beta_j(0) = 0$. Since the number of nullspace modes, $n$, is finite, we obtain that $\|(e_N)_{(k-1)t}\| \sim O(h^p)$. Repeating the argument in the proof of Lemma 4.3 leads to $\|e_N\| \sim O(h^p)$. We conclude that the nullspace modes converge to order $p$.

Turning to the nullspace orthogonal part, $e_\perp$, it satisfies,

$$(e_\perp)_t = P_b e_\perp + T_b^\perp$$

Since $e_\perp$ does not contain energy in the nullspace modes, we can reduce the system by removing the nullspace modes. (See derivation leading to (39).) Then, the remaining part of the proof is the same as for $e_n$, when the nullspace is trivial. That is, we Laplace transform the reduced system (for which the determinant condition is satisfied) and obtain a convergence rate of $r + q$.

Corollary 4.10. If Assumption 4.2 holds, the scheme (2) is convergent if $p \geq 1$ and $r + q \geq 1$. The convergence rate is $\min(r + q,p)$. \hfill $\square$

Proof. Follows immediately from Theorem 4.9 and Lemma 4.3. \hfill $\square$

We end this section with a few remarks.

- Nullspace consistency is of utmost importance. If a scheme is not nullspace consistent, we could not have isolated the nullspace growth since the modes would mix. Without a separation of the modes, the system (53) remains uninvertible.
- As long as the scheme is nullspace consistent it need not be consistent in the boundary stencils to converge. It suffices that $r + q \geq 1$.
- When the nullspace is non-trivial, the (nullspace orthogonal part, i.e., $e_\perp$, of the) boundary errors converge with the rate $r + q$ but the accuracy of the nullspace modes ($e_N$) also limits the rate to $p$ by the internal accuracy. (Hence, it is only when the nullspace contribution is negligible or $e_\perp$ very large, that one can record the $r + q$ rate.)

4.1. Incompletely parabolic system. Once again, we focus on the incompletely parabolic case since for hyperbolic or parabolic systems the accuracy results from the scalar counterparts generalize trivially.

Hence, we consider (41). We assume that the internal accuracy of both $D_1$ and $D_2$ is $p$. (All other constructions would be strange.) The boundary accuracy of $D_1$ is $r_1$ and $D_2$ is $r_2$. Assuming that both schemes are nullspace consistent and energy stable, we know from Section 3.2 that the problem satisfies the (generalized) determinant condition. We divide the error into internal and boundary parts as
before. The internal error converges to zero at the rate \( p \) (by the energy estimate.) The boundary part satisfies an equation,

\[
\begin{pmatrix}
\frac{\partial}{\partial t} \tilde{e}_b^u \\
\frac{\partial}{\partial t} \tilde{e}_b^v
\end{pmatrix}
= \begin{pmatrix}
a_{11} & a_{12} \\
-a_{12} & a_{22}
\end{pmatrix}
\begin{pmatrix}
D_1^b \tilde{e}_b^u \\
D_1^b \tilde{e}_b^v
\end{pmatrix}
+ \begin{pmatrix}
0 \\
\epsilon D_2 \tilde{e}_b^v
\end{pmatrix}
+ \begin{pmatrix}
(T_b)_1 \\
(T_b)_2
\end{pmatrix},
\]

where \((T_b)_1 = O(h')\) and \((T_b)_2 = O(h^2, h'^2)\) are the boundary truncation errors. Laplace transforming and rescaling by \(h, h^2\) results in

\[
\begin{pmatrix}
sI + a_{11} D_1^b \\
ha_{12} D_1^b
\end{pmatrix}
\begin{pmatrix}
a_{12} D_1^b \\
sI + ha_{22} D_1^b - \epsilon D_2
\end{pmatrix}
\begin{pmatrix}
\hat{e}_b^u \\
\hat{e}_b^v
\end{pmatrix}
= \begin{pmatrix}
h(\tilde{T}_b)_1 \\
h^2(\tilde{T}_b)_2
\end{pmatrix},
\]

Equation (60) can be solved as a difference equation since \((\hat{e}_b^u, \hat{e}_b^v)\) is invertible and the errors, \(e_b^u\) and \(e_b^v\) will be of the order of the right-hand side. That is the convergence rate will be: \(\min(r_1 + 1, r_2 + 2, p)\).

\section{Examples and discussion}

According to Corollary 4.10, \( r \) can be zero, which corresponds to an inconsistent boundary discretization, as long as the discretization is nullspace consistent. This may seem counterintuitive but it is nevertheless the case.

\subsection{First-derivative in time}

In the theory presented above, we use convergence in \(L^2\) obtained from the energy estimate to rule out eigenvalues with \(Re \tilde{s} \geq 0\) \((\tilde{s} \neq 0)\). This requires \( r \geq 0 \). If the (generalized) determinant condition can be inferred by some other means there is nothing that prevents that \( r \) is actually negative. Computational examples where \( r \) is negative and still obey the convergence rules outlined in this paper are, for example, found in [Mat14, MS15, MW16].

\begin{example}
In [SN06], the biharmonic equation,

\[ u_{11} = -u_{xxxx}, \quad 0 < x < 1, \]

subject to the boundary conditions

\[ u(0, t) = g_1(t), \quad u_{xx}(0, t) = g_2(t), \]
\[ u(1, t) = g_3(t), \quad u_{xx}(1, t) = g_4(t), \]

was used as an example. It was discretized using the first-derivative SBP operator, \( D_1 \). That is, the fourth-derivative was discretized as \( D_4 u = D_1 D_1 D_1 D_1 u \). (See [SN06] for implementation of the boundary conditions.) The scheme is energy stable. Furthermore, \( D_1 \) is sixth-order in the interior and third-order near the boundary. This implies that \( D_4 \) has accuracy \( p = 6 \) and \( r = 0 \). The boundary schemes are inconsistent.

The nullspace of the initial-boundary value problem is \( i, x \) and it is easy to deduce that the semi-discrete scheme is nullspace consistent. Hence, the convergence rate should be \( \min(p, r + 4) = 4 \) which indeed was the case.
\end{example}

\begin{remark}
To deduce nullspace consistency in practice, one first tests if the nullspace modes of the PDE are nullspace modes of the scheme. That is, project the modes on the grid and apply the discrete scheme. If the result is zero, they are in the nullspace of the spatial operator. Furthermore, one needs to test if there are more discrete nullspace modes, e.g., by numerically calculating the eigenvalues of the spatial scheme. There should be as many zero eigenvalues as the number of nullspace modes in the continuous problem. If there are more, the scheme is not nullspace consistent. The eigenvalues can be calculated for a fixed grid size and the
number of zeros counted. No further zero eigenvalues can appear since the stencils added to an enlarged grid are by construction linearly independent to all others.

In the literature, there are numerous examples of energy stable schemes, including advection-diffusion and incompletely parabolic system, obeying the convergence rates derived in this article. We direct the reader to e.g. [SN06, MSN04, vdWGS15, KMRS09, KMRS12, NGvdWS09].

5.2. The second-order wave equation. The canonical example of an equation with a second-derivative in time is the wave equation. According to the present theory, an energy stable and nullspace consistent scheme should gain two orders in the boundary errors (excluding the nullspace modes). For instance, an energy stable and nullspace consistent SBP-SAT scheme for the wave equation with Neumann boundary conditions was shown in [SN06] to converge according to the present theory.

Next, we will discuss some recent numerical results for the wave equation with SBP-SAT schemes presented in [WK16]. To be directly comparable, we use the same SBP operator (found in [MN04]) of order $(p, r) = (6, 3)$.

5.2.1. Dirichlet boundary conditions. The first example is the wave equation with Dirichlet boundary conditions, given by

\[
\begin{align*}
  u_{tt} &= u_{xx}, \quad x \in (0, \infty), \quad 0 < t \leq T \\
  u(x, 0) &= f_1(x) \\
  u_t(x, 0) &= f_2(x) \\
  u(0, t) &= 0, \quad (u(y, t) \to 0, y \to 0, \text{ sufficiently fast})
\end{align*}
\]

(61)

(We have dropped the forcing function which is of no concern here. Furthermore, we ignore the right boundary to reduce notation.) This problem is semi-discretely approximated by

\[
v_{tt} = D_2 v - H^{-1}(S^T E_0 - \tau h^{-1} E_0) v.
\]

(The problem and the scheme are the same as discussed in Example 2.20.) To prove that this problem is stable a so-called “borrowing trick” must be used. (See [CNG99].) In the standard definition, we have $D_2 = H^{-1}(-\bar{A} + BS)$ where $\bar{A}$ is positive semi-definite. In order to bound the derivatives that appear on the boundary in the energy estimate, one has to borrow a derivative approximation from $\bar{A}$. This is done by rephrasing $D_2$ as

\[
D_2 = H^{-1}(-h\alpha_{2p}(BS)^T BS - \bar{A} + BS)
\]

where $\bar{A} = h\alpha_{2p}(BS)^T BS + \bar{A}$ and $\alpha_{2p}$ is the largest possible scalar that maintains positive semi-definiteness of $\bar{A}$. In [WK16], energy stability is proven for the scheme above under the condition that $\tau \geq 1/\alpha_{2p}$. In the marginal case, $\tau = 1/\alpha_{2p}$, the scheme is energy stable but not nullspace consistent. There is a nullspace mode introduced that is not present in the initial-boundary value problem. (An extra zero eigenvalue in the spatial operator.) Since the theory in this article requires nullspace consistency, we should not expect a gain of two orders. This conclusion is corroborated by the numerical simulations in [WK16], which show a drop in the convergence rate and demonstrates the importance of nullspace consistency.

Remark 5.3. Since there is a rank deficit in $\bar{A}(\tau = \alpha_{2p})$, $u^T \bar{A} u = \|u\|_{\bar{A}}^2$ is unable to bound $\|Du\|$ where $\|Du\|$ is the discrete $L^2$ norm of $Du = (u_{i+1} - u_i)/h$. Hence, the solution can not be bounded in $L^\infty$ and the drop of accuracy is consistent with the theory in [SN06] as well.
Turning to the case \( \tau > 1/\alpha^2 \), the theory of this paper predicts the rate 5 for the (6,3) scheme.

**Remark 5.4.** For \( \tau > 1/\alpha^2 \), all SBP schemes “gain” two orders in the boundary error. However, the global rate is of course limited by the interior accuracy. Hence, the global rate is 2 for the (2,1) scheme and 4 for the (4,2) scheme.

**Remark 5.5.** Also [SN06] predicts two orders gain since \( \tau > \alpha^2 \) implies that \( \tilde{A} \) has the same rank as \( A \) and a bound on \( \|Du\| \) is obtained, rendering the scheme pointwise stable.

To address the question whether the rates are sharp, we make the following observation: The size of \( \hat{e}_b \) is obtained by solving the difference equation. This is accomplished by the ansatz \( (\hat{e}_b)_i = \sigma r^i \), which is inserted in the interior stencil and yields the roots \( r \). The \( \sigma \)'s are determined from the boundary stencils via a system, \( C(\tilde{\sigma})\sigma = h^{q_T} b \).

In this case, there are no nullspace modes in the problem and as we have shown, the energy estimate implies that the boundary matrix \( C(\hat{\sigma}) \) is non-singular for \( \text{Re}\hat{\sigma} \geq 0 \). Furthermore, there is no explicit \( h \) dependence of \( C \) and the components are \( O(1) \). This implies that \( C^{-1} \) is \( O(1) \). Hence, \( \sigma \) has to be \( O(h^{q_T} b) = O(h^{q+\tau}) \). Furthermore, the boundary modes are global and their influence do not vanish as \( h \to 0 \). (If the influence vanished it would mean that the boundary conditions do not influence the solution.) Hence, the size of \( \sigma \) reflects the \( L^2 \) size of the corresponding mode. It would appear that this rate is indeed sharp. To test this, we ran simulations using the (6,3) scheme on ever finer grids. We use the standard fourth-order Runge-Kutta scheme to march in time and stopped at \( T = 0.2 \) where the errors were measured.

The first test uses the exact solution \( u_{ex} = \sin(2\pi x) \cos(2\pi t) \). The results are tabulated in Table 1. Beginning with the \( L^2 \)-errors we see that the errors level out at a resolution of 500 points. The convergence rate in \( L^2 \) appears to be around 5.5. The column denoted “mean rate” lists the averages of the rates in the previous column up to and including the current grid. (Mean rate = \( 1/k \sum_{j=1}^{k} \text{rate}_j \) for the \( k \) first refinements.) Also this appears to be 5.5. The remaining grids are polluted by round off and this is clearly seen in the erratic convergence rates. Had we only studied the \( L^2 \)-errors, it would appear that our theoretical prediction was not sharp and that the true rate is 5.5.

However, the \( L^2 \)-errors are a mix of boundary errors (that according to our theory converge to 5th order) and internal errors (6th order). At the available accuracy levels, the internal errors may be dominating and the \( L^2 \)-rate appear higher than 5. Turning to the \( L^\infty \)-errors we see a similar pattern that the errors level out at around the same grid resolution. Since it is a linear problem, and the solution is smooth, all modes are global and we should expect to see that the \( L^\infty \) convergence rate asymptotically approaches the \( L^2 \) rate. However, the \( L^\infty \) rate should give an earlier indication of what the global rate will be since it is more likely to pick up the lowest convergence rate. In Table 1, the \( L^\infty \) rate is more erratic than the \( L^2 \) but it is clearly lower. This is evident in the mean rate displayed in the last column, which is closer to 5th order.

However, this particular exact solution is not tailored for displaying boundary errors. The integer in the parenthesis in Table 1 is the grid index where the maximal error appear. Clearly, it is not at the boundary but at different locations in the interior. Hence, it is likely that the internal errors contribute significantly to the total error and the internal errors are known to converge at a higher rate.
Table 1. Convergence rate for $u_{ex} = \sin(2\pi x)\cos(2\pi t)$. (6,3)-scheme. Dirichlet boundary.

<table>
<thead>
<tr>
<th>N</th>
<th>$L^2$-error</th>
<th>$L^2$-rate</th>
<th>mean rate</th>
<th>$L^\infty$-error (index)</th>
<th>$L^\infty$-rate</th>
<th>mean rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.1627e-08</td>
<td>-</td>
<td>-</td>
<td>7.2714e-08 (98)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200</td>
<td>4.3022e-10</td>
<td>5.6516</td>
<td>5.6516</td>
<td>1.8335e-09 (36)</td>
<td>5.30955</td>
<td>5.30955</td>
</tr>
<tr>
<td>300</td>
<td>4.7190e-11</td>
<td>5.5512</td>
<td>5.5512</td>
<td>2.6522e-10 (56)</td>
<td>4.76845</td>
<td>5.0390</td>
</tr>
<tr>
<td>400</td>
<td>9.5298e-12</td>
<td>5.5544</td>
<td>6.0176e-11 (76)</td>
<td>5.15598</td>
<td>5.0780</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>2.8349e-12</td>
<td>5.5241</td>
<td>1.7635e-11 (96)</td>
<td>5.50048</td>
<td>5.1836</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>1.3450e-12</td>
<td>5.2372</td>
<td>9.9372e-12 (36)</td>
<td>3.14000</td>
<td>4.7761</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>4.5788e-13</td>
<td>5.5512</td>
<td>5.5512</td>
<td>2.6522e-10 (56)</td>
<td>4.76845</td>
<td>5.0390</td>
</tr>
<tr>
<td>800</td>
<td>9.9304e-13</td>
<td>5.5544</td>
<td>6.0176e-11 (76)</td>
<td>5.15598</td>
<td>5.0780</td>
<td></td>
</tr>
<tr>
<td>900</td>
<td>2.6507e-12</td>
<td>5.5512</td>
<td>5.5512</td>
<td>2.6522e-10 (56)</td>
<td>4.76845</td>
<td>5.0390</td>
</tr>
<tr>
<td>1000</td>
<td>4.8289e-12</td>
<td>5.5512</td>
<td>5.5512</td>
<td>2.6522e-10 (56)</td>
<td>4.76845</td>
<td>5.0390</td>
</tr>
</tbody>
</table>

Since the $L^\infty$ errors do not appear at the boundary, we next investigate another solution that highlights boundary errors better.

To this end, we follow the technique used in [SN17] and choose the solution $u(x, t) = \cos(t)(x^p - x)$. This function does not solve the wave equation and we augment the wave equation with an appropriate forcing function. (This does not affect the convergence theory.) The function has strong curvature close to the right boundary. We should expect to see the maximum error at that location. If so, it provides a good measure of the boundary error since they decay into the domain. Furthermore, if the global error is 5th order, we should expect the maximal error to decay at 5th order.

We compute the solution, $u(x, t)$, for $p = 15$. The results are given in Table 2.

The $L^2$ errors are quite erratic but the mean value is consistently beneath 5.5. Also the $L^\infty$ errors appear erratic but they consistently appear at the right boundary as expected. What is clear is that the mean rates of the $L^\infty$-errors are well below 5.5 and with one exception below 5. The mean rates increase as the grids
are refined and indicate a convergence rate of 5, in accordance with our theoretical estimates.

To further emphasize that the $L^\infty$ rate is 5, the problem was computed on grids with \(20, 40, 60, \ldots, 500\) points. The sequence of mean $L^\infty$ rates become, \(4.67, 4.76, 4.77, 4.70, 4.85, 5.00, 5.05, 5.04, 4.98, 4.81, 4.75, 4.70, 4.81, 5.03, 5.04, 5.05, 5.04, 4.91, 4.81, 4.72, 4.76, 5.04, 5.05\).

We end this section, with an observation. The two numerical examples demonstrate that it is difficult to capture the optimal rates since the boundary errors might not dominate at the available grid resolutions. (The problem is not what we can afford to run within a reasonable time but rather arithmetic limits. It is important to note that to investigate boundary errors numerically, one can not settle with one example and measure the $L^2$ rates and expect to have recorded the optimal rates.

In practice this of course means that for some problems one may enjoy a higher than expected rates, if interior errors dominate, but in other cases the boundary errors dominate and the rates fall to the ones predicted in this paper.

5.2.2. Neumann boundary conditions. Consider the problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= u_{xx}, \quad x \in (0, \infty), \quad 0 < t \leq T \\
u(x, 0) &= f_1(x), \\
u_t(x, 0) &= f_2(x) \\
u_t(0, t) &= g(t), \quad (u(y, t) \to 0, y \to 0, \text{ sufficiently fast})
\end{align*}
\]

(Once again, we focus on one boundary. For the computations, the scheme is appropriately mirrored.) The SBP-SAT scheme is

\[
\begin{align*}
\mathbf{v}_{tt} &= D_2 \mathbf{v} - H^{-1}(S \mathbf{v} - \mathbf{e}_0 g)
\end{align*}
\]

where \(\mathbf{e}_0 = (1, 0, 0, \ldots, 0)^T\). Choosing the (6,3) operators of [MN04], we have run simulations to study the practical convergence rates of this scheme. The current theory predicts that the boundary errors gain two orders and converges with a rate of five, while the internal errors and nullspace errors converge with a rate of six.

The convergence rates measured in practice are highly dependent on the choice of data. The exact solution used in [WK16] for this particular case is

\[
\begin{align*}
u &= \cos(10 \pi x + 1) \cos(10 \pi t + 2), \quad 0 \leq x \leq 1, 0 \leq t \leq 2.
\end{align*}
\]

With the solution (64), it is likely that the internal errors, which are 6th order, dominate the boundary errors making the convergence rate appear greater than 5th order (which is our theoretical result). Indeed, running this case suggests a convergence rate that is greater than 5.

Hence, and in order to emphasize the boundary errors, we use the monomial \(m(x) = x^n\). We run this problem with the 6th-order scheme and monitor both the $L^2$ errors and the $L^\infty$ errors on grids of size \(N = 100, 200, \ldots, 800\) at the time \(T = 0.2\). (We use the standard 4th-order Runge-Kutta scheme with a small time step to advance the solution and keep the temporal error negligible.) The results of the simulations for the monomial \(m(x) = x^{15}\) are listed in Table 3. We see that the $L^2$ convergence rate appear higher than 5. However, the $L^\infty$ errors tell a different story. The maximum errors appear on the right boundary, and as such it is a good measure of the boundary errors. We see that they are less than 5, and increasing towards 5. A strong indication that the boundary errors are converging to 5th order and that our theoretical result is sharp.

Remark 5.6. The theory in [WK16] predicted a convergence rate of 5.5, both for this case (with Neumann boundary conditions) and the previous one in Section
Table 3. Convergence rate for 15th-order monomial. (6,3)-scheme. Neumann boundary.

<table>
<thead>
<tr>
<th>N</th>
<th>$L^2$-error</th>
<th>$L^2$-rate</th>
<th>$L^\infty$-error</th>
<th>$L^\infty$-rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.5689e-07</td>
<td>-</td>
<td>1.4152e-06</td>
<td>-</td>
</tr>
<tr>
<td>200</td>
<td>6.3820e-09</td>
<td>5.3310</td>
<td>5.1113e-08</td>
<td>4.7912</td>
</tr>
<tr>
<td>300</td>
<td>7.2359e-10</td>
<td>5.3692</td>
<td>7.2535e-09</td>
<td>4.8156</td>
</tr>
<tr>
<td>400</td>
<td>1.5382e-10</td>
<td>5.3826</td>
<td>1.8042e-09</td>
<td>4.8364</td>
</tr>
<tr>
<td>500</td>
<td>4.6236e-11</td>
<td>5.3867</td>
<td>6.1091e-10</td>
<td>4.8531</td>
</tr>
<tr>
<td>600</td>
<td>1.7334e-11</td>
<td>5.3811</td>
<td>2.5266e-10</td>
<td>4.8427</td>
</tr>
<tr>
<td>700</td>
<td>7.5740e-12</td>
<td>5.3710</td>
<td>1.1925e-10</td>
<td>4.8705</td>
</tr>
<tr>
<td>800</td>
<td>3.7122e-12</td>
<td>5.3403</td>
<td>6.1503e-11</td>
<td>4.9588</td>
</tr>
</tbody>
</table>

5.2.1 (with Dirichlet boundary conditions), for the exact same scheme that we have used herein. This is neither supported by the current theory nor the numerical simulations. See also Appendix I and [SN17] for more details.

6. Summary

The main result in this paper is that the “gain of boundary accuracy” for energy stable finite difference schemes is equal to the order of the (spatial) principal part of the differential operator. It is only required that the scheme satisfies the standard consistency conditions (7) and (8), that it is nullspace consistent (see Def. 2.9), and is energy stable. Our theoretical results are consistent with convergence reported for energy stable schemes in the literature and the simulations reported in this article. In particular, this article supersedes [SN06] by relaxing the strong requirement of pointwise stability.

Of particular interest, is the case when the nullspace of (1) and (2) is non-trivial. We have shown the differential/difference operator does not affect the nullspace modes. If the problem is homogeneous, the nullspace modes are initially set by initial data and thereafter evolves independently as an ordinary differential equation (ODE). The non-nullspace part of the solution (the PDE-part) is, of course, also set by initial data but subsequently evolved by the differential operator.

If a forcing function is added, it means that part of it will feed the nullspace, which is governed by an ODE that is independent of the differential operator, and another part will drive the PDE-part of the problem in the usual way. To treat non-homogeneous boundary conditions, we transformed the problem to a homogeneous problem with a smooth forcing function, and it behaves as described above. Another interpretation is that the boundary conditions “feed” the nullspace modes and the PDE-part separately.

In all cases, the nullspace part of the solution does not affect the convergence rate of the semi-discrete scheme. The reason for this is that the non-nullspace error and nullspace error do not mix in the semi-discrete scheme, which, with the help of Lemma 4.8, allowed us to prove that the nullspace error remains of the same size as the error of the interior base scheme.

Although, the analysis of convergence rates was the main topic of this article, we proved a number of auxiliary results that are significant in their own right. For instance, in Section 3.3, we demonstrate that the determinant condition is valid with higher derivatives in time. (This is often assumed but we are not aware of any proof.) Furthermore, while it is well-known that $L^2$- (or energy) stability rules out eigenvalues with $Re \tilde{s} > 0$, it is proven in Section 3 that it also rules out $Re \tilde{s} \geq 0$, $\tilde{s} \neq 0$. By assuming nullspace consistency, we could also handle the $\tilde{s} = 0$ case.
and prove that the (generalized) determinant condition is satisfied. Hence, we have linked Laplace-transform stability and energy stability.

References


APPENDIX

I. ON THE “MODELLING” OF TRUNCATION ERRORS

We consider the (2,1) approximation of the problem given in Section 5.2.2. The current theory, as well as our previous theory in [SN06], predicts a convergence rate of 2 for this problem. However, in [WK16], it is claimed that their theoretical derivation of the convergence rate is in contradiction with [SN06]. A numerical example is provided that supposedly supports their conclusion. (We stress that the current theory is in agreement with [SN06].) These claims were rebutted in [SN17]. However, the current theory provides yet another perspective on this issue and hence, we will discuss it at some length. The (2,1)-version of the scheme (63) is,

$$\left\{\begin{array}{l}
(v_0)_t = \frac{2v_1 - 2v_0}{h^2} \\
(v_{i+1})_t = \frac{2v_i + v_{i-1}}{h^2}, \quad i = 1, 2, ..., N - 1 \\
(v_N)_t = -\frac{2v_N + 2v_{N-1}}{h^2}
\end{array}\right.$$  

The truncation errors appearing in the associated error equation are:

$$T = (O(h), O(h^2), ..., O(h^2), O(h))^T$$

and therefore

$$T_h = (O(h), 0, ..., 0)^T.$$  

This problem is energy stable and consistent (including nullspace consistent). However, there is one non-zero spatial nullspace mode, $w(x) = 1$. Hence, according to
the present theory the nullspace-orthogonal part of the boundary error converges with third-order accuracy. The nullspace error (if it is non-zero) converges with second-order and the internal error with second-order. The global order is thusly second-order. Hence, the present theoretical results are in accordance with [WK16] and [SN06] since, in all cases, a rate of 2 is predicted.

As a matter of fact, although there is no theoretical contradiction in the end result, it is maintained in [WK16], that only “1 order is gained” at the boundary and therefore a “numerical test” that allegedly shows a one order increase of accuracy was designed.

This test allegedly measures convergence towards the exact solution (64). This would of course give a second order convergence rate and to avoid this, a forcing function that was \(10\pi^3\) in the first grid point and zero elsewhere, was added. This function is termed “added truncation errors” and it is claimed that such a forcing function models a zeroth-order truncation error. Furthermore, they claim that [SN06], with the added forcing, predicts that the scheme should converge at a second-order rate. We emphasize that neither [SN06] nor the current theory, which states the same results as [SN06] under milder assumptions, predicts a second-order rate for this modified scheme.

There are many ways to demonstrate that this claim is false. It should be obvious that a non-smooth forcing does not satisfy the basic smoothness assumptions necessary for any convergence theory. (This was one of the observations used in [SN17] to rebut these claims.)

Another observation in view of the current theory, is that the error induced by the discontinuous forcing does not stem from the spatial discretization. The forcing feeds energy into the nullspace component such that it becomes proportional to the mean value of the forcing function. (See (58), where in this case \(w_1 = 1\) and \(T_b\) is the made-up forcing function. When \(w = 1\), the inner product becomes the mean value which, if \([T_b]_0 = \text{constant}\) and zero elsewhere, is \(O(h)\).) Incidentally, the convergence recorded is \(O(h)\), coinciding with the mean of the forcing. We stress that this has nothing to do with the boundary errors per se. The forcing function “feeds” errors into the nullspace mode (by equation (58)), which is not affected by the scheme. (This mechanism is described in the last paragraph of Section 6.)

**Remark I.1.** The mean of a real boundary truncation error is \(O(h^p)\). That is, of the same order as the internal accuracy. Hence, when a real truncation error “feeds” the nullspace, it does not affect the overall convergence rate. An arbitrary non-smooth forcing that is used to “model” a truncation error, does not have this property in general, unless it is very carefully designed.

As the next observation, let us not regard the “added truncation error” as a forcing but a direct modification of the scheme. This would imply that a constant \(c\) is added to the scheme in the first grid point. This alters the first equation of (65) to become

\[
(v_0)_{tt} = \frac{2v_1 - 2v_0 + ch^2}{h^2}.
\]

The nullspace of the IBVP is \(w(x) = 1\). The unmodified scheme (65) is nullspace consistent since \(v_1 = 1\) results in the spatial discretization being zero. However, in (66) \(v_1 = 1\) is not a nullspace mode. Hence, this scheme is not nullspace consistent according to Definition 2.9 and not consistent in the sense of (7) and (8). (The latter conditions are the standard consistency conditions for finite difference schemes and consistency is an assumption in [SN06] as well.) In fact, this is a prime example of a scheme where the truncation error is not a Taylor series, as discussed in Section 2.3. (Note that by choosing \(c \sim h^\alpha\), \(0 < \alpha < 1\) the scheme is consistent in the
sense (11) but still not nullspace consistent.) We conclude that such an error term is very illustrative of the current theory.

However, based on the above discussion it is obvious that modelling errors by artificially “adding truncation errors” lead to conclusions that differ from measuring the actual errors in a computation, and should therefore be avoided.

Finally, we remark that it is very difficult to single out the nullspace orthogonal part of the error in a numerical computation. In [SN17], a manufactured solution that only produced numerical errors from the boundary scheme was used in a simulation. Although the internal errors are zero, it is still not possible to isolate a 3rd order convergence rate of the orthogonal errors because the remaining errors contains nullspace and orthogonal errors from both boundaries, i.e., four components of the error. Nevertheless, in these simulations, as in all other simulations with this scheme, the convergence rate approaches two from above. A strong indication that there is a component converging at a faster rate than two. Since the rates are integer-valued, that rate has to be three.

Dept. of Mathematics, University of Bergen, P.O. Box 7803, 5020 Bergen, Norway
E-mail address: Magnus.Svard@uib.no

Computational Mathematics, Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden
E-mail address: jan.nordstrom@liu.no