Journal of Computational Physics 270 (2014) 86–104

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

The SBP-SAT technique for initial value problems

Tomas Lundquist*, Jan Nordström

Department of Mathematics, Computational Mathematics, Linköping University, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 5 November 2013
Received in revised form 7 March 2014
Accepted 25 March 2014
Available online 2 April 2014

Keywords:
Time integration
Initial value problems
High order accuracy
Initial boundary value problems
Boundary conditions
Global methods
Stability
Convergence
Summation-by-parts operators
Stiff problems

ABSTRACT

A detailed account of the stability and accuracy properties of the SBP-SAT technique for numerical time integration is presented. We show how the technique can be used to formulate both global and multi-stage methods with high order of accuracy for both stiff and non-stiff problems. Linear and non-linear stability results, including A-stability, L-stability and B-stability are proven using the energy method for general initial value problems. Numerical experiments corroborate the theoretical properties.

1. Introduction

The SBP-SAT technique for time integration offers a promising new way of solving initial value problems (IVP) with high accuracy and optimal stability properties. It is based on previous development for spatial discretizations using high order finite difference formulas on summation-by-parts (SBP) form [1,2] together with the simultaneous-approximation-term (SAT) technique for imposing boundary conditions weakly [3,4].

The initial development of this technique applied on IVPs was done in [5], where a global implicit formulation was discussed. The focus was mainly on problems resulting from the spatial discretization of linear initial boundary value problems (IBVP). Optimal fully discrete energy estimates were derived for energy stable linear problems, and high order rates of convergence for stiff and non-stiff problems were demonstrated numerically.

In a global method, the solution for the whole time interval of interest is computed at once through a coupled system. Other global methods have been proposed previously, e.g. using collocation and spectral approximations [6–9], but are often considered unpractical even though the accuracy and stability are hard to match with local methods.

The standard way to solve initial value problems is instead to integrate the solution successively over small time increments, either through implicit or explicit formulas. The most common ones include linear multi-step methods [10,11], the Runge–Kutta family of multi-stage methods [12,13] as well as various general linear methods [14,15].

In this paper we present a more complete theoretical description of the SBP-SAT technique as a method to solve general initial value problems. Stability results for both linear and non-linear problems are proven using the energy method, and theoretical convergence rates are derived for both stiff and non-stiff problems. We also extend the technique to a one-step

* Corresponding author.
E-mail addresses: tomas.lundquist@liu.se (T. Lundquist), jan.nordstrom@liu.se (J. Nordström).

http://dx.doi.org/10.1016/j.jcp.2014.03.048
0021-9991/© 2014 Elsevier Inc. All rights reserved.

multi-stage formulation as an alternative to the previous global approach. The stability results obtained with the energy method for the multi-stage version are related to the classical stability properties of implicit Runge–Kutta methods.

In Section 2 we formulate the global version of the SBP-SAT technique in time for both constant coefficient and general non-linear problems. In Section 3 we prove energy stability and contractivity for fully-discrete solutions. In Section 4 we prove accuracy and dual consistency for scalar constant coefficient problems. Sections 5 and 6 extend the technique to a multi-stage formulation. In Section 7 we study the classical stability properties for the multi-stage version. In Section 8 we present numerical examples that support the theoretical results. Finally in Section 9 we draw conclusions.

2. Global SBP-SAT approximations

We consider an initial value problem on the general form

$$\begin{aligned} u_t + F(t,u) &= g(t), \quad 0 < t \leqslant T \\ u(0) &= f, \end{aligned} \tag{1}$$

where $u = (u_0, u_1, \dots, u_M)^T \in \mathbf{C}^{M+1}$. The input data to the problem consists of the forcing function g as well as the initial data f. The Peano theorem (see e.g. [16]) guarantees the existence of solutions to (1) for continuous functions F and g. The non-linear function $F \in \mathbf{C}^{M+1}$ is often assumed to possess specific properties to ensure well-posedness. For example, if F is Lipschitz continuous in u, then the Picard–Lindelöf theorem guarantees the existence of a unique solution to (1) (see e.g. [16]). The time-dependent forcing term g on the other hand can often be disregarded in the analysis of well-posedness and stability, as we will see in Section 3.

Assume that we look for a numerical solution to (1) on a uniform grid $\vec{t}$ in time, with $N+1$ grid points:

$$\vec{t} = (0, \Delta t, \dots, T)^T,$$

where $\Delta t = T/N$. We then define the numerical solution to be

$$\vec{U} = \begin{pmatrix} U_0 \\ U_1 \\ \vdots \\ U_N \end{pmatrix}, \quad \text{where } U_i = \left(U_i^0, U_i^1, \dots, U_i^M\right)^T, \quad i = 0, 1, \dots, N.$$

In order to discretize (1), we need to define the numerical time derivative of $\vec{U}$. For this, we use discrete first derivative finite difference operators on summation-by-parts form, so-called SBP operators.

2.1. SBP operators

The standard L_2 inner product between two square integrable scalar functions ϕ and ψ on $0 \leqslant t \leqslant T$ is defined by

$$(\phi, \psi)_{L_2} = \int_0^T \bar{\phi}\psi \, dt, \qquad \|\phi\|_{L_2}^2 = (\phi, \phi)_{L_2},$$

where $\bar{\phi}$ is the complex conjugate of ϕ.

Assume further that ϕ is sufficiently smooth, and define the restriction of ϕ to $\vec{t}$ as

$$\vec{\phi} = \left(\phi(0), \phi(\Delta t), \dots, \phi(T)\right)^T$$

In order to approximate the time derivative of ϕ numerically, we use a discrete first derivative operator on the form $D = P^{-1}Q$, so that $D\vec{\phi} \approx \vec{\phi}_t$. D is said to be on summation-by-parts form if

$$P = P^T > 0, \quad Q + Q^T = E_N - E_0, \tag{2}$$

where $E_0 = \vec{e}_0\vec{e}_0^T$, $E_N = \vec{e}_N\vec{e}_N^T$, $\vec{e}_0 = (1\ 0\ \dots\ 0)^T$, $\vec{e}_N = (0\ \dots\ 0\ 1)^T$.

Moreover, P defines a discrete integration operator from which we can derive a discrete version of the L_2 inner product:

$$(\vec{\phi}, \vec{\psi})_P = \vec{\phi}^* P \vec{\psi}, \qquad \|\vec{\phi}\|_P^2 = (\vec{\phi}, \vec{\phi})_P, \tag{3}$$

where $\vec{\phi}^*$ is the conjugate transpose of $\vec{\phi}$. The discrete L_2 inner product defined in this way together with the discrete derivative operator now automatically satisfies the following summation-by-parts rule:

$$\begin{aligned} \left(P^{-1}Q\vec{\phi}, \vec{\psi}\right)_P &= \vec{\phi}^* Q^T \vec{\psi} = \vec{\phi}^*(-Q + E_N - E_0)\vec{\psi} \\ &= \bar{\phi}(T)\psi(T) - \bar{\phi}(0)\psi(0) - \left(\vec{\phi}, P^{-1}Q\vec{\psi}\right)_P. \end{aligned}$$

This imitates the integration-by-parts rule of the continuous L_2 inner product:

$$(\phi_t, \psi)_{L_2} = \bar{\phi}(T)\psi(T) - \bar{\phi}(0)\psi(0) - (\phi, \psi_t)_{L_2}.$$

For brevity, we will refer to P as the norm of the SBP operator. In the interior rows, an SBP operator typically consists of a central finite difference scheme, and here P is diagonal with positive entries. At the boundaries of the operator on the other hand, i.e. for a limited number of rows at the top and bottom, P can be either diagonal or have a pair of small symmetric positive definite blocks. Whether P is diagonal everywhere or not will have implications for both stability and accuracy, and we will refer to these cases as operators with diagonal norms and block norms respectively. In both cases the norm is scaled by the time step size, and can thus be written

$$P = \Delta t H, \tag{4}$$

where H has positive entries of order one in magnitude.

The accuracy conditions of an SBP operator can be expressed as the exact differentiation of monomials up to a specific order:

$$P^{-1} Q \vec{t}^{j} = j\vec{t}^{j-1}, \quad j = 0, 1, \ldots, \tag{5}$$

where we define $\vec{t}^{j} = (0, \Delta t^{j}, \ldots, T^{j})^T$.

It should be noted that the order of accuracy is in general higher for the central finite difference stencil in the interior of the SBP operator than for the boundary treatment. This means that (5) is satisfied for higher values of j in rows associated with the central difference scheme compared with a limited number of rows at the top and bottom. Operators have been constructed with internal order $2s$, for $s = 1, 2, 3, 4, 5$. When using diagonal norms the order at the boundaries is limited to s, while with block norms it can be increased to $2s - 1$ [2,17–19].

Together with the SAT technique for imposing the initial condition, the SBP operators described in the previous section can be used to discretize any initial value problem of the form (1).

Definition 1. Let $D = P^{-1}Q$ be an SBP operator with internal order $2s$, and boundary order p given by either $p = s$ (diagonal norm) or $p = 2s - 1$ (block norm). We then denote the method of solving the initial value problem (1) with the SBP-SAT technique by SBP($2s$, p).

2.2. The scalar constant coefficient case

Following [5], we first consider a scalar constant coefficient problem:

$$\begin{aligned} u_t + \lambda u &= g, \quad 0 < t \leqslant T \\ u(0) &= f, \end{aligned} \tag{6}$$

where λ is a complex constant. The SBP-SAT approximation of (6) is

$$P^{-1} Q \vec{U} + \lambda \vec{U} = \vec{g} + P^{-1}\sigma(U_0 - f)\vec{e}_0. \tag{7}$$

The last term in the right-hand side of (7) is the so-called SAT penalty term forcing the solution at $t = 0$ toward the initial condition weakly. The linear system to solve can also be written as

$$\big(P^{-1}\tilde{Q} + \lambda I\big)\vec{U} = \vec{g} - \sigma P^{-1} f \vec{e}_0, \tag{8}$$

where $\tilde{Q} = Q - \sigma E_0$. Extensive numerical evidence and a proof for the second order case supports the assumption that the spectrum of $P^{-1}\tilde{Q}$ lies strictly in the right half-plane for $\sigma < -\frac{1}{2}$ (see Assumption 1 in [5]). This guarantees a unique solution to (8) for $Re(\lambda) \geqslant 0$.

It is also reasonable to require that all elements in the matrix $(P^{-1}\tilde{Q} + \lambda I)^{-1}$ should be bounded as $\mathcal{O}(\Delta t)$. To motivate this, consider problem (6) with homogeneous initial data, i.e. $f = 0$. The exact solution is then $u = e^{-\lambda t}\int_0^t e^{\lambda\tau} g(\tau) d\tau$, and the corresponding discrete solution is $\vec{U} = (P^{-1}\tilde{Q} + \lambda I)^{-1}\vec{g}$. The value of g on a time interval of length Δt thus gives a contribution to the solution u that is also of order Δt. Analogously, the value of any single component in $\vec{g}$ should then give a contribution to $\vec{U}$ that is of the order of the grid size Δt.

Numerical experiments involving a variety of different SBP operators corroborate both these assumptions. For future reference, we formalize these results below.

Assumption 1. For $\sigma < -\frac{1}{2}$, all eigenvalues of $P^{-1}\tilde{Q}$ have strictly positive real parts.

Assumption 2. For $\sigma < -\frac{1}{2}$ and $Re(\lambda) \geqslant 0$, all elements of the matrix $(P^{-1}\tilde{Q} + \lambda I)^{-1}$ in (8) are at most of order Δt.

Corollary 1. *Assume that Assumption 2 holds. Then, for $\sigma < -\frac{1}{2}$ and $Re(\lambda) \geqslant 0$, we have $\|(P^{-1}\tilde{Q} + \lambda I)^{-1}\|_\infty \leqslant \mathcal{O}(1)$.*

Proof. Let a be the largest element in magnitude of $(P^{-1}\tilde{Q} + \lambda I)^{-1}$. Then, by definition we have

$$\begin{aligned}\big\|\big(P^{-1}\tilde{Q} + \lambda I\big)^{-1}\big\|_\infty &= \max_{\|x\|_\infty = 1} \big\|\big(P^{-1}\tilde{Q} + \lambda I\big)^{-1} x\big\|_\infty \\ &\leqslant (N+1)|a| \leqslant (N+1)\mathcal{O}(\Delta t) = \mathcal{O}(1). \qquad \square\end{aligned}$$

2.3. The constant coefficient system case

As a first extension to more general problems, we consider the case where (1) is a linear system with constant coefficients, i.e.

$$\begin{aligned} u_t + Au = g, &\quad 0 < t \leqslant T \\ u(0) = f, & \end{aligned} \tag{9}$$

and A is a constant matrix. The SBP-SAT approximation to this system is most easily constructed using a Kronecker product formulation. For two arbitrary matrices $\mathcal{A} \in R^{m\times n}$ and $\mathcal{B} \in R^{p\times q}$, the Kronecker product $\mathcal{A} \otimes \mathcal{B}$ is defined as

$$\mathcal{A} \otimes \mathcal{B} = \begin{bmatrix} a_{1,1}\mathcal{B} & \dots & a_{1,m}\mathcal{B} \\ \vdots & \ddots & \vdots \\ a_{n,1}\mathcal{B} & \dots & a_{m,n}\mathcal{B} \end{bmatrix}.$$

The Kronecker product is a bilinear and associative operation that obeys the following rules for conjugate transposition and mixed products:

$$\begin{aligned} (A \otimes B)^* &= A^* \otimes B^* \\ (\mathcal{A} \otimes \mathcal{B})(\mathcal{C} \otimes \mathcal{D}) &= (\mathcal{A}\mathcal{C} \otimes \mathcal{B}\mathcal{D}), \end{aligned}$$

if the matrix products $\mathcal{A}\mathcal{C}$ and $\mathcal{B}\mathcal{D}$ are defined. We can now formulate an SBP-SAT approximation to (9) as follows:

$$\big(P^{-1}Q \otimes I\big)\vec{U} + (I \otimes A)\vec{U} = \vec{g} + P^{-1}\sigma\vec{e}_0 \otimes (U_0 - f), \tag{10}$$

where I is a unit matrix of dimension $M+1$. If all eigenvalues of A have a real part that is larger than or equal to zero, then we can conclude, using Assumption 1, that all eigenvalues of the system (10) have strictly positive real parts. This follows from the fact that $(P^{-1}Q \otimes I)$ and $(I \otimes A)$ commute, and hence are simultaneously triangularizable (see [5] for more details). This guarantees that a unique solution to (10) exists.

2.4. The general non-linear case

In the general case, an SBP-SAT approximation of (1) can be formulated as:

$$\big(P^{-1}Q \otimes I\big)\vec{U} + \begin{pmatrix} F(t_0, U_0) \\ \vdots \\ F(t_N, U_N) \end{pmatrix} = \begin{pmatrix} g(t_0) \\ \vdots \\ g(t_N) \end{pmatrix} + P^{-1}\sigma\vec{e}_0 \otimes (U_0 - f), \tag{11}$$

where I is a unit matrix of dimension $M+1$. Note that only the linear constant coefficient terms can be expressed using Kronecker products.

3. The energy method

The stability properties in the discrete approximation (11) are related to those of the original equation. In general we say that an initial value problem (1) is *energy stable* if the solution is bounded by initial data, i.e. $\|u\| \leqslant \|f\|$ in some norm $\|\cdot\|$. A related concept is that of contractivity. We study how energy stability and contractivity are preserved in the fully discrete problem for the three different types of problems in Sections 2.2–2.4.

3.1. Energy stable scalar constant coefficient problems

We first consider the scalar case (6). For the purpose of well-posedness, it is sufficient to consider (6) with zero forcing function [20]. The energy method (multiplying with the complex conjugated solution and integrating over the domain) yields

$$\left|u(T)\right|^2 + 2Re(\lambda)\|u\|_{L_2}^2 = |f|^2. \tag{12}$$

If $Re(\lambda) \geqslant 0$, then the solution is bounded by initial data, so that (6) is energy stable. It is desirable that this property holds also for the discrete solution. The SBP-SAT approximation of (6) with $g = 0$ is

$$P^{-1}Q\vec{U} + \lambda\vec{U} = P^{-1}\big(\sigma(U_0 - f)\big)\vec{e_0}. \tag{13}$$

We are now ready to state the following proposition.

Proposition 1. *Let $\vec{U}$ be the solution to the SBP-SAT approximation (13) of (6) with $g = 0$. Then, $\sigma = -1$, $Re(\lambda) \geqslant 0$ implies $|U_N|^2 \leqslant |f|^2$.*

Proof. We follow the path set in [5]. The discrete energy method applied to (13) (multiplying from the left with $\vec{U}^*P$ and adding the conjugate transpose) leads to the energy estimate

$$|U_N|^2 + 2Re(\lambda)\|\vec{U}\|_P^2 = \sigma\big(\bar{U}_0(U_0 - f) + (\bar{U}_0 - \bar{f})U_0\big) + |U_0|^2. \tag{14}$$

By adding and subtracting $|f|^2$ we get

$$|U_N|^2 \leqslant |f|^2 + \begin{pmatrix} U_0 \\ f \end{pmatrix}^* \begin{pmatrix} 1+2\sigma & -\sigma \\ -\sigma & -1 \end{pmatrix} \begin{pmatrix} U_0 \\ f \end{pmatrix}.$$

With $\sigma = -1$, the matrix in the expression above is negative semi-definite, but with all other choices of penalty coefficient it is indefinite. □

Note that with the specific choice $\sigma = -1$ used in Proposition 1, the energy estimate (14) also becomes optimally sharp:

$$|U_N|^2 + 2Re(\lambda)\|\vec{U}\|_P^2 = |f|^2 - |U_0 - f|^2. \tag{15}$$

Compare (15) to the continuous estimate (12). The discrete bound perfectly mimics the continuous one, and has in addition the small damping term $-|U_0 - f|^2$. This kind of optimally strict estimate is to our knowledge never obtained using local time-stepping methods.

3.2. Energy stable constant coefficient systems

The stability theory for scalar problems can be easily extended to linear constant coefficient systems. Again it is sufficient for well-posedness to consider (9) with zero forcing function [20]. Let $\tilde{P}$ be a symmetric, positive definite matrix of dimension $M+1$, and define the inner product $(\cdot,\cdot)_{\tilde{P}}$ on $\mathbf{R}^{M+1}$ as

$$(u, v)_{\tilde{P}} = u^*\tilde{P}v, \qquad \|u\|_{\tilde{P}}^2 = (u, u)_{\tilde{P}}. \tag{16}$$

The energy method (multiplying from the left with $u^*\tilde{P}$, adding the conjugate transpose and integrating) then yields the following energy estimate:

$$\left\|u(T)\right\|_{\tilde{P}}^2 + \int_0^T u^*\big(\tilde{P}A + A^T\tilde{P}\big)u\,dt = \|f\|_{\tilde{P}}^2. \tag{17}$$

From this we see that (9) with $g = 0$ is energy stable if $\tilde{P}A + A^T\tilde{P} \geqslant 0$. The SBP-SAT approximation of (9) with $g = 0$ is

$$\big(P^{-1}Q \otimes I\big)\vec{U} + (I \otimes A)\vec{U} = P^{-1}\sigma\vec{e_0} \otimes (U_0 - f). \tag{18}$$

Proposition 2. *Let $\vec{U}$ be the solution to the SBP-SAT approximation (18) of (9) with $g = 0$. Then, $\sigma = -1$ and $\tilde{P}A + A^T\tilde{P} \geqslant 0$ implies $\|U_N\|_{\tilde{P}}^2 \leqslant \|f\|_{\tilde{P}}^2$.*

Proof. The discrete energy method (multiplying (18) from the left with $\vec{U}^*(P \otimes \tilde{P})$ and adding the conjugate transpose) yields the estimate

$$\|U_N\|^2_{\tilde{P}} + \vec{U}^*\big(P \otimes \big(\tilde{P}A + A^T\tilde{P}\big)\big)\vec{U} = \sigma\big(U_0^*\tilde{P}(U_0 - f) + (U_0 - f)^*\tilde{P}U_0\big) + \|U_0\|^2_{\tilde{P}}.$$

Note first that $\tilde{P}A + A^T\tilde{P} \geqslant 0$ implies that $P \otimes (\tilde{P}A + A^T\tilde{P}) \geqslant 0$. By adding and subtracting $\|f\|^2_{\tilde{P}}$ we then get

$$\|U_N\|^2_{\tilde{P}} \leqslant \|f\|^2_{\tilde{P}} + \begin{pmatrix} U_0 \\ f \end{pmatrix}^* \left(\begin{pmatrix} 1+2\sigma & -\sigma \\ -\sigma & -1 \end{pmatrix} \otimes \tilde{P}\right)\begin{pmatrix} U_0 \\ f \end{pmatrix}.$$

Again, the only choice of penalty parameter which gives a negative semi-definite matrix in the above expression is $\sigma = -1$. □

The discrete energy estimate with the choice $\sigma = -1$ can be written as

$$\|U_N\|^2_{\tilde{P}} + \vec{U}^*\big(P \otimes \big(\tilde{P}A + A^T\tilde{P}\big)\big)\vec{U} = \|f\|^2_{\tilde{P}} - \|U_0 - f\|^2_{\tilde{P}},$$

which perfectly mimics the continuous estimate (17).

3.3. Energy stable non-linear problems

The energy method applied to the general initial value problem (1) with zero forcing function yields the estimate

$$\big\|u(T)\big\|^2_{\tilde{P}} + \int_0^T 2Re\big(u, F(t,u)\big)_{\tilde{P}}\,dt = \|f\|^2_{\tilde{P}}. \tag{19}$$

Thus problem (1) is energy stable if the non-linear function F satisfies the condition:

$$Re\big(x, F(t,x)\big)_{\tilde{P}} \geqslant 0, \quad 0 \leqslant t \leqslant T,\ x \in \mathbf{R}^{M+1}. \tag{20}$$

Proposition 3. *Let $\vec{U}$ be the solution to the SBP-SAT approximation (11) of (1) with $g = 0$, using a diagonal norm P. Then, for $\sigma = -1$, energy stability (20) implies that $\|U_N\|^2_{\tilde{P}} \leqslant \|f\|^2_{\tilde{P}}$.*

Proof. Consider the discrete problem (11) with $g = 0$, and multiply from the left with $\vec{U}^*(P \otimes \tilde{P})$:

$$\vec{U}^*(Q \otimes \tilde{P})\vec{U} + \vec{U}^*(P \otimes \tilde{P})\begin{pmatrix} F(t_0, U_0) \\ \vdots \\ F(t_N, U_N) \end{pmatrix} = \sigma U_0^*\tilde{P}(U_0 - f).$$

Since P is diagonal, adding the complex conjugate now yields:

$$\|U_N\|^2_{\tilde{P}} + 2\sum_{i=0}^{N+1} P_{ii}Re\big(\big(U_i, F(t_i, U_i)\big)_{\tilde{P}}\big) = \sigma\big(U_0^*\tilde{P}(U_0 - f) + (U_0 - f)^*\tilde{P}U_0\big) + \|U_0\|^2_{\tilde{P}}.$$

Adding and subtracting $\|f\|^2_{\tilde{P}}$ and using energy stability (20) gives

$$\|U_N\|^2_{\tilde{P}} \leqslant \|f\|^2_{\tilde{P}} + \begin{pmatrix} U_0 \\ f \end{pmatrix}^* \left(\begin{pmatrix} 1+2\sigma & -\sigma \\ -\sigma & -1 \end{pmatrix} \otimes \tilde{P}\right)\begin{pmatrix} U_0 \\ f \end{pmatrix}.$$

As before the matrix above is negative semi-definite if and only if $\sigma = -1$. □

The discrete energy estimate with the choice $\sigma = -1$ can now be written as

$$\|U_N\|^2_{\tilde{P}} + 2\sum_{i=0}^{N+1} P_{ii}Re\big(\big(U_i, F(t_i, U_i)\big)_{\tilde{P}}\big) = \|f\|^2_{\tilde{P}} - \|U_0 - f\|^2_{\tilde{P}},$$

which perfectly mimics the continuous estimate (19).

Remark 1. Note that the proof only holds for diagonal norms P.

Remark 2. Note that Proposition 3 does not guarantee the existence of a unique discrete solution, but only that it is bounded by data if it exists. This is consistent with the observation that energy stability in itself is not a sufficient condition for well-posedness of the continuous problem.

3.4. Contractive problems

As we mentioned in the previous section, there is for general non-linear problems no direct link between energy stability and well-posedness. As an alternative approach to defining stability in the non-linear case, we can instead consider contractivity as the property we wish to preserve in the fully-discrete solution.

Assume that u and v are two different solutions to (1) with corresponding initial data f_1 and f_2 respectively. For the difference $u - v$ we then have

$$(u-v)_t + F(t,u) - F(t,v) = 0, \quad 0 < t \leqslant T$$
$$(u-v)(0) = f_1 - f_2$$

The energy method (multiplying from the left with $(u-v)^*\tilde{P}$, adding the complex conjugate and integrating) then yields:

$$\|u-v\|^2_{\tilde{P}} + \int_0^T 2Re\big(u-v, F(t,u) - F(t,v)\big)_{\tilde{P}} = \|f_1 - f_2\|^2_{\tilde{P}} \tag{21}$$

Now assume that the function F satisfies the following condition:

$$Re\big(x-y, F(t,x) - F(t,y)\big)_{\tilde{P}} \geqslant 0, \quad 0 \leqslant t \leqslant T, \ x, y \in \mathbf{R}^{M+1}. \tag{22}$$

Then, for the difference $u - v$, we get the bound

$$\|u-v\|_{\tilde{P}} \leqslant \|f_1 - f_2\|_{\tilde{P}}.$$

Note that this guarantees uniqueness of the solution to (1), given that one exists. If (22) holds, we say that the initial value problem (1) is *contractive*.

Proposition 4. *Let $\vec{U}$ and $\vec{V}$ be the solutions to the SBP-SAT approximation (11) of the IVP (1) with initial data f_1 and f_2 respectively, using a diagonal norm P. Then, $\sigma = -1$ and the contractivity condition (22) implies $\|U_N - V_N\|^2_{\tilde{P}} \leqslant \|f_1 - f_2\|^2_{\tilde{P}}$.*

Proof. The equation for $\vec{U} - \vec{V}$ can be written

$$\big(P^{-1}Q \otimes I\big)(\vec{U} - \vec{V}) + \begin{pmatrix} F(t_0, U_0) - F(t_0, V_0) \\ \vdots \\ F(t_N, U_N) - F(t_N, V_N) \end{pmatrix} = P^{-1}\sigma\vec{e}_0 \otimes \big((U-V)_0 - (f_1 - f_2)\big).$$

We multiply from the left with $(\vec{U} - \vec{V})^*(P \otimes \tilde{P})$:

$$(\vec{U} - \vec{V})^*(Q \otimes \tilde{P})(\vec{U} - \vec{V}) + (\vec{U} - \vec{V})^*(P \otimes \tilde{P}) \begin{pmatrix} F(t_0, U_0) - F(t_0, V_0) \\ \vdots \\ F(t_N, U_N) - F(t_N, V_N) \end{pmatrix}$$
$$= \sigma(U_0 - V_0)^*\tilde{P}\big((U-V)_0 - (f_1 - f_2)\big).$$

Since the norm P is diagonal, adding the conjugate transpose now yields:

$$\|U_N - V_N\|^2_{\tilde{P}} - \|U_0 - V_0\|^2_{\tilde{P}} + 2Re\sum_{i=0}^{N+1} P_{ii}\big(U_i - V_i, F(t_i, U_i) - F(t_i, V_i)\big)_{\tilde{P}}$$
$$= \sigma\Big((U_0 - V_0)^*\tilde{P}\big((U-V)_0 - (f_1 - f_2)\big)$$
$$+ \big((U-V)_0 - (f_1 - f_2)\big)^*\tilde{P}(U_0 - V_0)\Big).$$

Adding and subtracting $\|f\|^2_{\tilde{P}}$ and using the contractivity condition (22) gives

$$\|U_N - V_N\|^2_{\tilde{P}} \leqslant \|f_1 - f_2\|^2_{\tilde{P}} + \begin{pmatrix} U_0 - V_0 \\ f_1 - f_2 \end{pmatrix}^* \left(\begin{pmatrix} 1 + 2\sigma & -\sigma \\ -\sigma & -1 \end{pmatrix} \otimes \tilde{P}\right) \begin{pmatrix} U_0 - V_0 \\ f_1 - f_2 \end{pmatrix}.$$

As before the matrix above is negative semi-definite if and only if $\sigma = -1$. □

The discrete energy estimate with the choice $\sigma = -1$ can now be written as

$$\|U_N - V_N\|^2_{\tilde{P}} + 2Re\sum_{i=0}^{N+1} P_{ii}\big((U_i - V_i, F(t_i, U_i - V_i))_{\tilde{P}}\big) = \|f_1 - f_2\|^2_{\tilde{P}} - \big\|U_0 - V_0 - (f_1 - f_2)\big\|^2_{\tilde{P}},$$

which mimics the continuous one (21) perfectly.

4. Accuracy

Next we study the accuracy of the SBP-SAT technique, and limit the discussion to the constant coefficient test problem (6) with $u \in C^{2s}$ and $Re(\lambda) \geqslant 0$. The order of accuracy can be found by constructing an equation for the numerical error that involves the truncation error resulting from the discretization, see also [21].

The numerical error is the difference between the exact and the numerical solution vector:

$$\vec{e} = \vec{u} - \vec{U}. \tag{23}$$

By inserting (23) in (7) we get the error equation:

$$P^{-1}Q\vec{e} + \lambda\vec{e} = P^{-1}\sigma e_0\vec{e}_0 + \vec{T}_e, \tag{24}$$

where the truncation error $\vec{T}_e$ is given by

$$\vec{T}_e = P^{-1}Q\vec{u} - \vec{u}_t. \tag{25}$$

We can also write the numerical error explicitly from (24) as

$$\vec{e} = \left(P^{-1}\tilde{Q} + \lambda I\right)^{-1}\vec{T}_e, \tag{26}$$

where $\tilde{Q} = Q - \sigma E_0$ as usual.

4.1. Pointwise order of accuracy

We start with a general result on the pointwise order of accuracy for SBP$(2s, p)$.

Proposition 5. *If Assumption 2 holds, then the pointwise order of accuracy of* SBP$(2s, p)$ *in (7) when applied to the constant coefficient test problem (6) is* $p + 1$.

Proof. Since the order of accuracy of the SBP operator is p at the boundaries and $2s$ in the interior, we split the truncation error vector into the two corresponding parts:

$$\vec{T} = \vec{T}_e^b + \vec{T}_e^i,$$

where $\vec{T}_e^i = \mathcal{O}(\Delta t^{2s})$, $\vec{T}_e^b = \mathcal{O}(\Delta t^p)$. Moreover, $\vec{T}_e^b$ has only a finite number of non-zero components. Consider the explicit expression for the numerical error (26). By Corollary 1 we have $\|(P^{-1}\tilde{Q} + \lambda I)^{-1}\|_\infty \leqslant \mathcal{O}(1)$. This gives immediately the following estimate for the contribution from $\vec{T}_e^i$ to $\vec{e}$ in (26):

$$\left\|\left(P^{-1}\tilde{Q} + \lambda I\right)^{-1}\vec{T}_e^i\right\|_\infty \leqslant \left\|\left(P^{-1}\tilde{Q} + \lambda I\right)^{-1}\right\|_\infty \left\|\vec{T}_e^i\right\|_\infty = \mathcal{O}\left(\Delta t^{2s}\right).$$

A similar estimate can be made for the boundary part: Since $\vec{T}_b$ is non-zero only at a finite number of positions, we get from Assumption 2 that

$$\left\|\left(P^{-1}\tilde{Q} + \lambda I\right)^{-1}\vec{T}_e^b\right\|_\infty \leqslant \mathcal{O}(\Delta t)\left\|\vec{T}_e^b\right\|_\infty = \mathcal{O}\left(\Delta t^{p+1}\right).$$

This gives the result

$$\|e\|_\infty \leqslant \mathcal{O}\left(\Delta t^{2s}\right) + \mathcal{O}\left(\Delta t^{p+1}\right) = \mathcal{O}\left(\Delta t^{p+1}\right). \qquad \square$$

Even though Proposition 5 gives a bound on the error that is valid for any value of $\lambda \geqslant 0$ in the limit $\Delta t \to 0$, it is sometimes possible to obtain more strict ones. This happens for example when SBP$(2s, p)$ is applied on problems that are stiff, i.e. on problems with step sizes Δt much larger than $|\lambda|^{-1}$. A practical example of such a problem can be found by setting $g = \psi_t + \lambda\psi$ in (6), where ψ is the prescribed exact solution, and λ is very large. The time step size required to resolve this problem can then indeed be much larger than $|\lambda|^{-1}$. As was demonstrated in [22] for implicit Runge–Kutta methods, the order of accuracy with respect to Δt can decrease for stiff problems of this type. We shall see that this order reduction phenomenon also occurs with the SBP-SAT technique.

Definition 2. For the SBP-SAT approximation (7) of the scalar constant coefficient problem (6), the stiff limit is: $|\lambda\Delta t| \to \infty$ together with $\Delta t \to 0$.

This definition leads to the following result for stiff convergence rate of SBP$(2s, p)$.

Proposition 6. *In the stiff limit, the accuracy of the solution obtained with* SBP$(2s, p)$ *when applied to the scalar constant coefficient problem (6) is* $\mathcal{O}(\frac{1}{\lambda}\Delta t^{2s})$ *for interior points, and* $\mathcal{O}(\frac{1}{\lambda}\Delta t^{p})$ *for boundary points.*

Proof. Consider the expression for the numerical error (26). Since by definition $|\lambda \Delta t| \to \infty$, for any given value of Δt the following estimate holds:

$$\vec{e} = \frac{1}{\lambda}\left(I + \frac{1}{\Delta t \lambda} H^{-1}\tilde{Q}\right)^{-1} \vec{T}_e = \frac{1}{\lambda}\vec{T}_e + \mathcal{O}\left(\frac{1}{\Delta t \lambda^2} H^{-1}\tilde{Q}\vec{T}_e\right) \to \frac{1}{\lambda}\vec{T}_e, \quad \text{as } \lambda \to \infty.$$

By definition $\vec{T}_e$ is of order $\mathcal{O}(\Delta t^{2s})$ for interior points, and $\mathcal{O}(\Delta t^{p})$ for boundary points. □

4.2. Superconvergence

From Proposition 5 we know that the order of accuracy is in general only $s + 1$ when using diagonal norms, while for block norms we get back the order of the interior stencil, or $2s$. However, for certain components of the solution, the order can be increased to $2s$ even for diagonal norms. This phenomenon is called superconvergence, and makes the operators with diagonal norms, which have superior stability properties, more competitive in terms of accuracy.

In order to formulate the superconvergence results, we need first to introduce the concept of dual consistency. Dual consistency has previously been used to prove superconvergence of the SBP-SAT technique for steady boundary value problems [23] as well as for unsteady initial boundary value problems [24,25].

Consider the constant coefficient initial value problem (6) with homogeneous initial data:

$$\begin{aligned} u_t + \lambda u &= g, \quad 0 < t \leqslant T \\ u(0) &= 0. \end{aligned} \tag{27}$$

We define a linear functional J of u as $J(u) = (h, u)_{L_2}$, where $h \in C^{2s}$. The dual problem now consists of finding a function ϕ such that the functional $J(u)$ equals the inner product between ϕ and the forcing function g: $J(u) = (\phi, g)_{L_2}$.

We find the dual problem by expanding the expression for the functional:

$$\begin{aligned} J(u) = (h, u)_{L_2} &= (h, u)_{L_2} - (\phi, u_t + \lambda u - g)_{L_2} \\ &= (h, u)_{L_2} + (\phi, g)_{L_2} - (\bar{\lambda}\phi, u)_{L_2} - (\phi, u_t)_{L_2} \\ &= (\phi, g)_{L_2} + (h - \bar{\lambda}\phi, u)_{L_2} - \bar{\phi}(T)u(T) + \bar{\phi}(0)u(0) + (\phi_t, u) \\ &= (\phi, g)_{L_2} + (h - \bar{\lambda}\phi + \phi_t, u)_{L_2} - \bar{\phi}(T)u(T). \end{aligned}$$

Thus the dual problem is:

$$\begin{aligned} -\phi_t + \bar{\lambda}\phi &= h, \quad 0 < t \leqslant T \\ \phi(T) &= 0. \end{aligned} \tag{28}$$

In a similar way we can define the discrete dual problem. The SBP-SAT discretization of (27) is

$$\left(P^{-1}\tilde{Q} + \lambda I\right)\vec{U} = \vec{g}. \tag{29}$$

As discrete functional we define

$$J_p(\vec{U}) = (\vec{h}, \vec{U})_P.$$

The discrete dual problem then consists of finding a vector $\vec{\Phi}$ such that

$$J_p(\vec{U}) = (\vec{\Phi}, \vec{g}).$$

Analogous manipulations as in the continuous case now yields

$$\begin{aligned} J_p(\vec{U}) &= (\vec{h}, \vec{U})_P - \left(\vec{\Phi}, \left(P^{-1}\tilde{Q} + \lambda I\right)\vec{U} - \vec{g}\right)_P \\ &= (\vec{h}, \vec{U})_P + (\vec{\Phi}, \vec{g})_P - (\bar{\lambda}\vec{\Phi}, \vec{U})_P - \left(\vec{\Phi}, P^{-1}\tilde{Q}\vec{U}\right)_P \\ &= (\vec{\Phi}, \vec{g})_P + (\vec{h} - \bar{\lambda}\vec{\Phi}, \vec{U})_P - \left(P^{-1}\tilde{Q}^T\vec{\Phi}, \vec{U}\right)_P \\ &= (\vec{\Phi}, \vec{g})_P + \left(\vec{h} - \bar{\lambda}I + \left(P^{-1}\left(Q + (1+\sigma)E_0 - E_N\right)\right)\vec{\Phi}, \vec{U}\right)_P. \end{aligned}$$

Thus the discrete dual problem is

$$-P^{-1}Q\vec{\Phi} + \bar{\lambda}I\vec{\Phi} = \vec{h} + P^{-1}\left((1+\sigma)\Phi_0\vec{e}_0 - \Phi_N\vec{e}_N\right).$$

With the choice $\sigma = -1$, this becomes a consistent approximation of the continuous dual problem (28):

$$-P^{-1}Q\vec{\Phi} + \bar{\lambda}I\vec{\Phi} = \vec{h} - P^{-1}\Phi_N\vec{e}_N. \tag{30}$$

The problem (30) is very similar to (7), and we can prove the order of accuracy in an analogous way.

Lemma 1. *If Assumption 2 holds then, with the choice $\sigma = -1$, the discrete dual problem (30) is a $p+1$ order accurate approximation of the continuous dual problem (28).*

Proof. With $\sigma = -1$, the system matrix $P^{-1}(-Q + E_N) + \bar{\lambda}I$ in (30) can be derived from the matrix $P^{-1}\tilde{Q} + \lambda I$ in (8) through the following transformation:

$$P^{-1}(-Q + E_N) + \bar{\lambda}I = P^{-1}\left(P^{-1}\tilde{Q} + \lambda I\right)^* P$$

Thus the inverse can be written

$$\left(P^{-1}(-Q + E_N) + \bar{\lambda}I\right)^{-1} = P^{-1}\left(\left(P^{-1}\tilde{Q} + \lambda I\right)^{-1}\right)^* P$$

Using Assumption 2, the proof is now analogous to that of Proposition 5. □

We will need one more lemma, stating that the norm of an SBP operator with diagonal norm is a high order integrator.

Lemma 2. *Let $P^{-1}Q$ be an SBP operator with diagonal norm of order $2s$ in the interior, and let $\phi, \psi \in C^{2s}$. Then $(\vec{\phi}, \vec{\psi})_P$ is a $2s$ order accurate approximation of $(\phi, \psi)_{L_2}$, and moreover $(\vec{\phi}, P^{-1}Q\vec{\psi})_P$ is a $2s$ order approximation of $(\phi, \psi_t)_{L_2}$.*

Proof. See [26]. □

We are now ready to state the superconvergence results.

Proposition 7. *Let $\vec{U}$ be the solution to the SBP-SAT approximation (7) of (6) with $\sigma = -1$ and $Re(\lambda) \geqslant 0$, and let $h \in C^{2s}$. Then $J_P(\vec{U}) = (\vec{h}, \vec{U})_P$ is a $2s$ order accurate approximation of $J(u) = (h, u)_{L_2}$.*

Proof. For block norms, the solution is of order $2s$ everywhere, so the result follows immediately in this case. Thus we assume henceforth that P is diagonal, and we follow the technique outlined in the proof of Theorem 4 in [23]. Let ϕ be the solution to the dual problem (28), and Φ the solution to the discrete dual problem (30). Let moreover $\vec{\phi}$ and $\vec{u}$ be the restrictions of ϕ and u to $\vec{t}$. We can expand the expression for the exact functional $J(u)$ as

$$\begin{aligned} J(u) &= J_P(\vec{u}) + \mathcal{O}\left(\Delta t^{2s}\right) \\ &= (\vec{h}, \vec{u})_P + (\vec{h}, \vec{U})_P - (\vec{h}, \vec{U})_P + \mathcal{O}\left(\Delta t^{2s}\right) \\ &= J_P(\vec{U}) + (\vec{h}, \vec{u} - \vec{U})_P + \mathcal{O}\left(\Delta t^{2s}\right). \end{aligned} \tag{31}$$

It now remains to show that $(\vec{h}, \vec{u} - \vec{U})_P = \mathcal{O}(\Delta t^{2s})$. We start by establishing the following identity by combining (6) and (8):

$$\begin{aligned} 0 &= P^{-1}\tilde{Q}\vec{U} + \lambda\vec{U} - \vec{g} + P^{-1}\sigma(f)\vec{e}_0 \\ &\quad - \left(\vec{u}_t + \lambda\vec{u} - \vec{g} - P^{-1}\sigma\left(u(0) - f\right)\vec{e}_0\right) \\ &= -\left(P^{-1}\tilde{Q} + \lambda I\right)(\vec{u} - \vec{U}) + P^{-1}Q\vec{u} - \vec{u}_t. \end{aligned} \tag{32}$$

Using (32) we can now expand $(\vec{h}, \vec{u} - \vec{U})_P$ as

$$\begin{aligned} (\vec{h}, \vec{u} - \vec{U})_P &= (\vec{h}, \vec{u} - \vec{U})_P - \left(\vec{\Phi}, \left(P^{-1}\tilde{Q} + \lambda I\right)(\vec{u} - \vec{U})\right)_P \\ &\quad + \left(\vec{\Phi}, P^{-1}Q\vec{u} - \vec{u}_t\right)_P \\ &= \left(\vec{h} - P^{-1}\left(P^{-1}\tilde{Q} + \bar{\lambda}I\right)^T P\vec{\Phi}, \vec{u} - \vec{U}\right)_P \\ &\quad + \left(\vec{\Phi}, P^{-1}Q\vec{u} - \vec{u}_t\right)_P. \end{aligned} \tag{33}$$

The first term in the right hand side of (33) now vanish due to (30):

$$\begin{aligned} \vec{h} - P^{-1}\left(P^{-1}\tilde{Q} + \bar{\lambda}I\right)^T P\vec{\Phi} &= \vec{h} - \left(P^{-1}\left(-Q - (1+\sigma)E_0 + E_N\right) + \bar{\lambda}I\right)\vec{\Phi} \\ &= \vec{h} - \Phi_N\vec{e}_N - \left(-P^{-1}Q + \bar{\lambda}I\right)\vec{\Phi} = 0. \end{aligned}$$

Moreover, the second term in the right hand side of (33) becomes, using Lemma 2 and Lemma 1,

$$\begin{aligned}\big(\vec{\Phi}, P^{-1}Q\vec{u}-\vec{u}_t\big)_P &= \big(\vec{\phi}, P^{-1}Q\vec{u}\big)_P - (\vec{\phi},\vec{u}_t) + \big(\vec{\Phi}-\vec{\phi}, P^{-1}Q\vec{u}-\vec{u}_t\big)_P \\ &= (\vec{\phi},\vec{u}_t)_P + \mathcal{O}\big(\Delta t^{2s}\big) - \big((\vec{\phi},\vec{u}_t)_P + \mathcal{O}\big(\Delta t^{2s}\big)\big) + (\vec{\Phi}-\vec{\phi},\vec{T}_e)_P \\ &= \mathcal{O}\big(\Delta t^{2s}\big) + \mathcal{O}\big(\Delta t^{2p+2}\big).\end{aligned}$$

Now, since either $p = s$ or $p = 2s-1$, (33) becomes

$$(\vec{h}, \vec{u}-\vec{U})_P = \mathcal{O}\big(\Delta t^{2s}\big),$$

which concludes the proof. □

Proposition 8. *Let $\vec{U}$ be the solution to the SBP-SAT approximation (7) of (6) with $\sigma = -1$ and $Re(\lambda) \geqslant 0$. Then U_N is a $2s$ order accurate approximation of $u(T)$.*

Proof. For block norms the result follows immediately from Proposition 5. Thus we assume henceforth that P is a diagonal norm. We now observe that the first accuracy condition ((5) with $j = 0$) together with the SBP property (2) leads to $\vec{1}^T Q = -\vec{e}_0^T + \vec{e}_N^T$. Multiplying (24) with $\vec{1}^T P$ then yields:

$$-e_0 + e_N + \lambda\vec{1}^T P\vec{e} = \sigma e_0 + \vec{1}^T P\vec{T}.$$

With $\sigma = -1$, this becomes:

$$e_N = \vec{1}^T P\vec{T} - \lambda\vec{1}^T P\vec{e}.$$

Lemma 2 now gives:

$$\begin{aligned}\vec{1}^T P\vec{T} &= (\vec{1},\vec{u}_t)_P - \big(\vec{1}, P^{-1}Q\vec{u}\big) \\ &= (1,u_t)_{L_2} + \mathcal{O}\big(\Delta t^{2s}\big) - \big((1,u_t)_{L_2} + \mathcal{O}\big(\Delta t^{2s}\big)\big) = \mathcal{O}\big(\Delta t^{2s}\big).\end{aligned}$$

Moreover, from Lemma 2 and Proposition 7 we get:

$$\begin{aligned}\vec{1}^T P\vec{e} &= (\vec{1},\vec{u})_P - (\vec{1},\vec{U})_P \\ &= (1,u)_{L_2} + \mathcal{O}\big(\Delta t^{2s}\big) - \big((1,u)_{L_2} + \mathcal{O}\big(\Delta t^{2s}\big)\big) = \mathcal{O}\big(\Delta t^{2s}\big).\end{aligned}$$

Thus

$$e_N = \mathcal{O}\big(\Delta t^{2s}\big). \qquad \square$$

Remark 3. Note that the general accuracy results presented in Propositions 5, 7 and 8 are also valid in the stiff limit. For example, when using diagonal norms in the stiff limit, Proposition 8 gives an error bound of order $\mathcal{O}(\Delta t^{2s})$ for the solution at the last time step, while Proposition 6 gives $\mathcal{O}(\frac{1}{\lambda}\Delta t^s)$. Both these estimates hold, and the latter only becomes more strict if $|\lambda\Delta t^s| > 1$.

5. Multi-block formulation

For computational considerations it is often advantageous to split the time interval of interest into several smaller blocks, for example when constructing adaptive methods. The SBP-SAT technique can be applied on each block individually, combined with an interface coupling between them. In this section we restrict the analysis to the case where only two blocks are used, but the extension to an arbitrary number of blocks is completely analogous.

Thus we assume that the time domain is split into two blocks with an interface at $t = a$, where $0 < a < T$, and we define a numerical approximation on this domain as

$$\begin{aligned}\vec{U} &= (U_0\, U_1\, \ldots\, U_N)^T \approx \big(u(0)\, u(\Delta t_1)\, \ldots\, u(a)\big)^T, \\ \vec{V} &= (V_0\, V_1\, \ldots\, V_M)^T \approx \big(u(a)\, u(a+\Delta t_2)\, \ldots\, u(T)\big)^T.\end{aligned}$$

The full solution vector is $\vec{W} = (U_0\ \ldots U_N\ V_0\ \ldots\ V_M)^T$, and we define the corresponding discrete L_2 norm as $\|\vec{W}\|^2_{\bar{P}} = \vec{W}^*\bar{P}\vec{W}$, where

$$\bar{P} = \begin{pmatrix} P_l & 0 \\ 0 & P_r \end{pmatrix}.$$

The two-domain implementation of the SBP-SAT technique for the scalar constant coefficient problem (6) can then be formulated as follows:

$$\begin{pmatrix} P_l^{-1}Q_l & 0 \\ 0 & P_r^{-1}Q_r \end{pmatrix}\vec{W} + \lambda\vec{W} = P_l^{-1}\big(\sigma(U_0 - f)\big)\vec{e}_0 + P_l^{-1}\big(\sigma_l(U_N - V_0)\big)\vec{e}_N + P_r^{-1}\big(\sigma_r(V_0 - U_N)\big)\vec{d}_0, \tag{34}$$

where σ_l and σ_r are SAT penalty parameters forcing the two solutions U_N and V_0 at $t = a$ toward each other. The subscripts l and r denote the left and right domain respectively. $\vec{e}_0$, $\vec{e}_N$ and $\vec{d}_0$ are unit vectors with zeros everywhere except at the position corresponding to U_0, U_N and V_0 respectively. Note that $\vec{e}_N$ and $\vec{d}_0$ point to the same time value.

The energy method (multiplying from the left with $\vec{W}^*\bar{P}$ and adding the conjugate transpose) then yields

$$|V_M|^2 + 2Re(\lambda)\|\vec{W}\|_{\bar{P}}^2 + \begin{pmatrix} U_N \\ V_0 \end{pmatrix}^T \begin{pmatrix} 1-2\sigma_l & \sigma_l+\sigma_r \\ \sigma_l+\sigma_r & -1-2\sigma_r \end{pmatrix} \begin{pmatrix} U_N \\ V_0 \end{pmatrix} = |f|^2 - |U_0 - f|^2. \tag{35}$$

As was originally shown in [4], the matrix in expression (35) above is positive semi-definite if and only if the following expressions hold:

$$\sigma_l = \sigma_r + 1, \quad \sigma_r \leqslant -\frac{1}{2}.$$

Consider the discrete problem (34) again. The choice $\sigma_r = -1$ clearly makes the solution to the first equation in (34) independent of the solution to the second. After the first equation is solved, the solution component U_N can then be used as initial condition to the second equation. With this choice the energy estimate (35) becomes:

$$|V_M|^2 + 2Re(\lambda)\|\vec{W}\|_{\bar{P}}^2 = |f|^2 - |U_0 - f|^2 - |V_0 - U_N|^2.$$

6. Multi-stage formulation

An alternative to solving (1) with the global SBP-SAT technique is to use a one-step multi-stage method, through the multi-block technique described above. The problem can thus be solved successively over small time increments involving only a small number of grid points, while using the numerical solution at the end of the most recent time increment as initial data to the next. As was shown in the previous section, this formulation is identical to the multi-block approach (34) with penalty coefficients $\sigma_r = -1$ and $\sigma_l = 0$.

We first discretize the time domain with a grid using *arbitrary* step sizes, and define the corresponding numerical solution to the one-step SBP-SAT method:

$$\vec{t} = (0\; t_1\; t_2\; \ldots\; t_N = T)^T,$$

$$\vec{U} = \begin{pmatrix} U_0 \\ U_1 \\ \vdots \\ U_N \end{pmatrix}, \quad \text{where } U_i = \big(U_i^0\; U_i^1\; \ldots\; U_i^M\big)^T, \quad i = 0, 1, \ldots, N.$$

The original problem (1) can now be partitioned into N corresponding subproblems that can be solved numerically one after the other:

$$\begin{aligned} u_t + F(t, u) &= g(t), \quad t_{i-1} < t \leqslant t_i,\; i = 1, \ldots, N \\ u(0) &= U_{i-1}, \end{aligned} \tag{36}$$

where $U_0 = f$. To solve each of these subproblems, each interval $[t_{i-1}, t_i]$ is further divided into $n_i + 1$ *equispaced* grid points:

$$\vec{t^i} = (t_{i-1}\; t_{i-1} + \Delta t_i\; \ldots\; t_i)^T,$$

where

$$\Delta t_i = \frac{t_i - t_{i-1}}{n_i}, \quad i = 1, \ldots, N.$$

Finally we define a discrete solution vector for each subproblem:

$$\vec{U}_i = \begin{pmatrix} U_{i,0} \\ U_{i,1} \\ \vdots \\ U_{i,n_i} \end{pmatrix} \approx \begin{pmatrix} u(t_{i-1}) \\ u(t_{i-1} + \Delta t_i) \\ \vdots \\ u(t_i) \end{pmatrix}, \quad i = 1, \ldots, N. \tag{37}$$

We refer to U_i^0 through $U_i^{n_i}$ as the $n_i + 1$ *stage values* at t_i, and define the numerical solution at t_i to be the last of these values, i.e.

$$U_i = U_i^{n_i}.$$

Each subproblem (36) can now be solved successively with the SBP-SAT technique:

$$\left(P^{-1}Q \otimes I\right)\vec{U}_i + \begin{pmatrix} F(t_0, U_{i,0}) \\ \vdots \\ F(t_{n_i}, U_{i,n_i}) \end{pmatrix} = \begin{pmatrix} g(t_0) \\ \vdots \\ g(t_{n_i}) \end{pmatrix} + P^{-1}\sigma\vec{e}_0 \otimes (U_{i,0} - U_{i-1}), \tag{38}$$

where $U_0 = f$, and $P = \Delta t_i H$.

This formulation allows us to reuse the various definitions of stability already existing for implicit Runge–Kutta methods. One major difference between these and the SBP-SAT technique is that, in the latter, also the "zeroth" stage value, i.e. U_0^i, has to be computed in each step, due to the weak coupling between $\vec{U}_i$ and $\vec{U}_{i-1}$.

Remark 4. One additional advantage with the multi-stage formulation is that it opens up for adaptive methods in time, since the length of each subinterval is arbitrary.

7. Classical stability properties of the SBP-SAT method in time

We conclude the theoretical part of this paper by relating the stability properties of the SBP-SAT method to various standard stability properties of Runge–Kutta methods, see e.g. [20]. As we shall see, many of these are linked to the energy estimates derived in Section 3.

The most widely used stability definition for Runge–Kutta methods is that of A-stability. It is based on the scalar constant coefficient problem (6) with zero forcing function.

Definition 3. The multi-stage method (38) is said to be A-stable if, when applied to the scalar constant coefficient test equation (6) with $g = 0$, $Re(\lambda) \geqslant 0$ implies that $|U_i| \leqslant |U_{i-1}|$, for $i = 1, \dots, N$.

In some cases, A-stability may not be enough since decaying solution components might be damped out too slowly. This motivates the definition of *L-stability* [20].

Definition 4. The multi-stage method (38) is said to be L-stable if it is A-stable and if in addition, when applied to (6) with $g = 0$, $Re(\lambda) \geqslant 0$ implies that $\frac{|U_i|}{|U_{i-1}|} \to 0$ as $\Delta t_i\, Re(\lambda) \to \infty$, for $i = 1, \dots, N$.

The most general extension to include non-linear problems is based on the concept of contractivity discussed in Section 3.4.

Definition 5. The multi-stage method (38) is said to be B-stable if the contractivity property (22) of F implies that $\|U_i - V_i\|_{\tilde{P}} \leqslant \|U_{i-1} - V_{i-1}\|_{\tilde{P}}$, for $i = 1, \dots, N$, where $\vec{U}$ and $\vec{V}$ are solutions associated with different initial data f_1 and f_2 respectively.

Finally, in order to take advantage of the energy estimates that we derived for energy stable linear and non-linear problems in Section 3.3, we introduce two more definitions.

Definition 6. The multi-stage method (38) is said to be linearly stable if, when applied to the constant coefficient test problem (9) with $g = 0$, $\tilde{P}A + A^T\tilde{P} \geqslant 0$ implies that $|U_i| \leqslant |U_{i-1}|$, for $i = 1, \dots, N$.

Definition 7. The multi-stage method (38) is said to be energy stable if (20) implies that $\|U_i\|_{\tilde{P}} \leqslant \|U_{i-1}\|_{\tilde{P}}$, for $i = 1, \dots, N$.

Using the energy estimates in Section 3 we can now prove the following results for the SBP-SAT methods.

Proposition 9. *The time integration method* SBP$(2s, p)$ *in the multi-stage setting (38) with $\sigma = -1$ is A-stable, L-stable and linearly stable.*

Proof. A-stability and linear stability follows directly from Proposition 1 and 2. Moreover, consider the energy estimate for step i in the multi-stage version:

$$|U_i|^2 + 2\,Re(\lambda)\|\vec{U}_i\|_P^2 = |U_{i-1}|^2 - |U_i - U_{i-1}|^2.$$

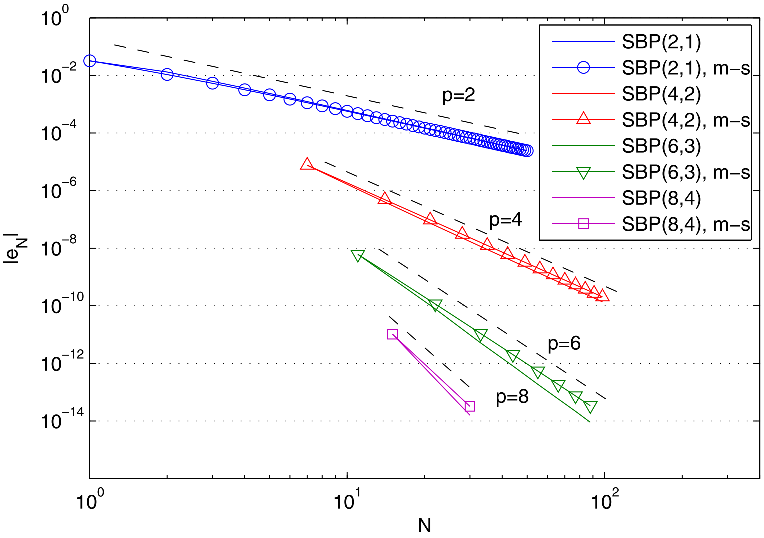

Fig. 1. Convergence at $t = 1$ in the non-stiff test case using diagonal norm operators.

Recall that $\|\vec{U}_i\|_P^2 = \vec{U}_i^* P \vec{U}_i$, and $P = \Delta t_i H$ where H has positive eigenvalues of order one. Let $\xi_{\min}$ be the smallest of those eigenvalues. Then the energy estimate above leads to

$$\big(1 + 2\xi_{\min}\Delta t_i \, Re(\lambda)\big)|U_i|^2 \leqslant |U_{i-1}|^2,$$

which can be rewritten as

$$\frac{|U_i|^2}{|U_{i-1}|^2} \leqslant \frac{1}{(1 + 2\xi_{\min}\Delta t_i \, Re(\lambda))} \to 0, \quad \text{as } \Delta t_i \, Re(\lambda) \to \infty. \qquad \square$$

The non-linear stability properties on the other hand are only attained for diagonal norms.

Proposition 10. *The time integration method* SBP$(2s, s)$ *in the multi-stage setting (38) with $\sigma = -1$ is B-stable and energy stable.*

Proof. The results follows directly from Propositions 3 and 4. □

For computational reasons it may be advantageous to use as few stages as possible to limit the size of problem (38). Since the number of stages is the same as the number of rows in the SBP operator $P^{-1}Q$, we note that the lower restriction on the number of stages is equivalent to the number of boundary rows in the SBP operator.

Proposition 11. *The number of stages for the SBP-SAT methods are limited by $n_i + 1 \geqslant 2$ for* SBP$(2, 1)$ *and $n_i + 1 \geqslant 4s$ for* SBP$(2s, p)$, *$s = 2, 3, 4$.*

Proof. See e.g. Lemma 2.9 and Theorem 2.3 in [2]. □

8. Numerical results

8.1. Accuracy

The stiff and non-stiff accuracy results given in Propositions 6 and 8 were demonstrated numerically in [5] for energy stable constant coefficient problems, and operators with internal order 2, 4, 6 and 8. We complete this picture by showing for a scalar example that the multi-stage formulation of SBP$(2s, p)$ produces errors of the same order as the global formulation given the same temporal resolution. For this purpose, we solve (6) with the exact solution $u = e^{-t}$ by setting the forcing function to $g = (\lambda - 1)e^{-t}$. We use the minimum number of stages, i.e. $n_i + 1 = 2, 8, 12$ and 16 for SBP$(2, p)$, SBP$(4, p)$, SBP$(6, p)$ and SBP$(8, p)$ respectively. We consider both a non-stiff case ($\lambda = 1$) and a stiff case ($\lambda = 1000$), and use both diagonal norms ($p = s$) and block norms ($p = 2s - 1$). Figs. 1 through 4 show the convergence of the global error at $t = 1$ for all these cases. For the stiff cases, the errors from the multi-stage versions are indistinguishable from those of the global versions. For the non-stiff case, note that the accuracy using diagonal norms and block norms is almost the same.

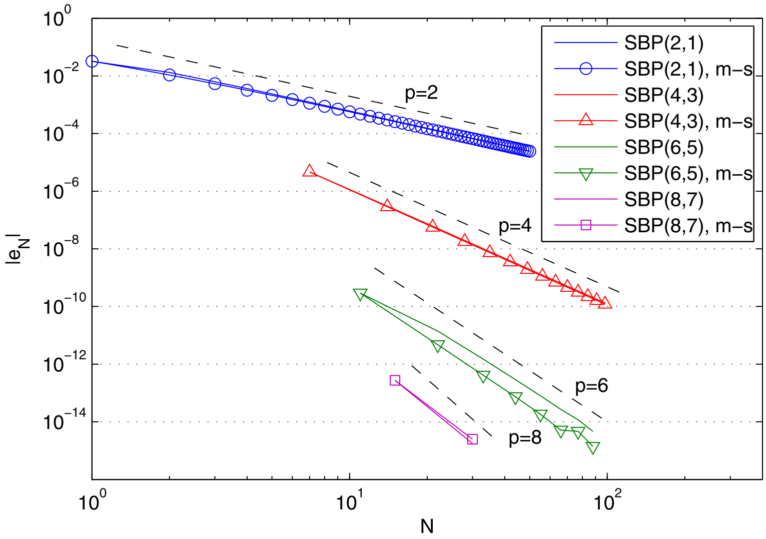

Fig. 2. Convergence at $t = 1$ in the non-stiff test case using block norm operators.

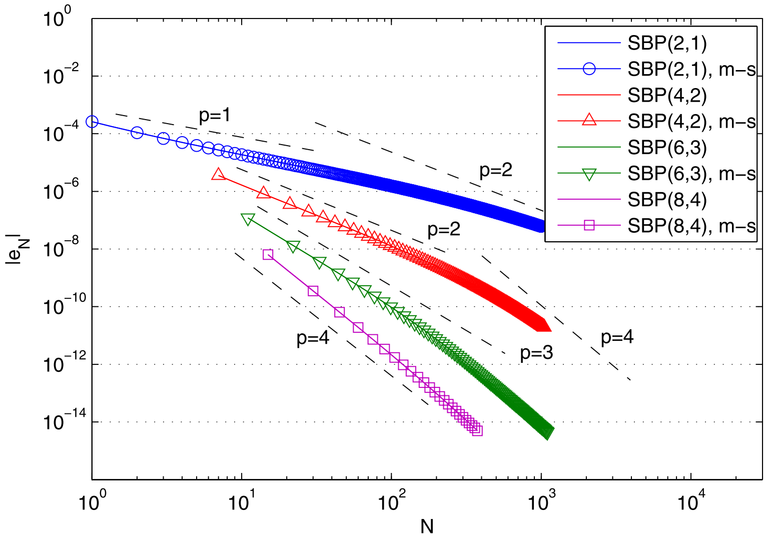

Fig. 3. Convergence at $t = 1$ in the stiff test case using diagonal norm operators. The multi-stage version is indistinguishable from the global version.

8.2. Stability

From Proposition 2 we know that energy stability of constant coefficient problems is preserved using SBP$(2s, p)$ for both diagonal norms and block norms. However, transforming the time coordinate introduces a time dependency in the coefficients, and in this case we know from Proposition 3 that stability can only be guaranteed for operators with diagonal norms. To test this, we consider a problem on the following form:

$$\begin{aligned} u_t + \tilde{P}^{-1}Au &= 0, \quad 0 < t \leqslant T \\ u(0) &= f, \end{aligned}$$

where $\tilde{P}$ is symmetric positive definite, and A is a skew-symmetric. The energy method then yields $\|u(T)\|_{\tilde{P}} = \|f\|_{\tilde{P}}$, i.e. the system is not only energy stable, but strictly energy conserving. Now we introduce a stretching of the time coordinate, and let $t = t(\tau)$. The system can then be rewritten as

$$\begin{aligned} u_\tau + t_\tau \tilde{P}^{-1}Au &= 0, \quad 0 < \tau \leqslant T \\ u(0) &= f. \end{aligned} \tag{39}$$

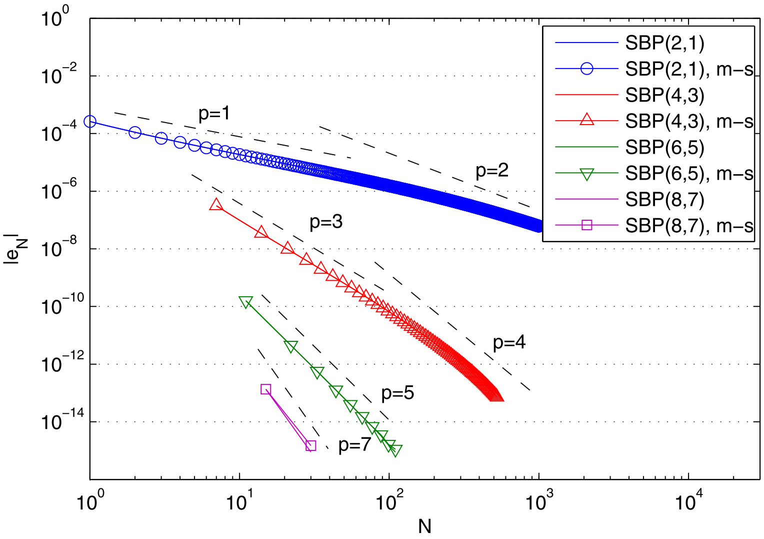

Fig. 4. Convergence at $t = 1$ in the stiff test case using block norm operators. The multi-stage version is indistinguishable from the global version.

Due to the time dependent factor t_τ in (39), we can only guarantee that energy stability is preserved using SBP($2s, p$) if a diagonal norm is used, see Proposition 3. To see what happens in detail, we consider the SBP-SAT approximation (11) of (39):

$$\left(P_\tau^{-1}Q \otimes I\right)\vec{U} + \left(\mathcal{T} \otimes \tilde{P}^{-1}A\right)\vec{U} = P_\tau^{-1}\sigma\vec{e}_0 \otimes (U_0 - f),$$

where $\mathcal{T} = Diag(\frac{d}{d\tau}(\vec{t}))$. The energy method (multiply with $u^*(P_\tau \otimes \tilde{P})$ from the left and adding the conjugate transpose) with $\sigma = -1$ yields:

$$\|U_N\|_{\tilde{P}}^2 + \vec{U}^*\left((P_\tau\mathcal{T} - \mathcal{T}P_\tau) \otimes A\right)\vec{U} = \|f\|_{\tilde{P}}^2 - \|U_0 - f\|_{\tilde{P}}^2. \tag{40}$$

If P_τ is diagonal, then $P\mathcal{T} - \mathcal{T}P = 0$, and stability follows. If P_τ is a block norm on the other hand, then $P_\tau\mathcal{T} - \mathcal{T}P_\tau$ is skew-symmetric. Since also A is skew symmetric, the eigenvalues of the matrix $(P_\tau\mathcal{T} - \mathcal{T}P_\tau) \otimes A$ are real and come in positive/negative pairs. This means that energy stability is not guaranteed, and the solution could thus potentially grow without bound.

As an example, we consider the following coupled hyperbolic system of partial differential equations:

$$\begin{aligned} u_t + u_x &= 0, \quad 0 \leqslant x \leqslant 1 \\ v_t - v_x &= 0, \quad 0 \leqslant x \leqslant 1 \\ u(0,t) &= v(0,t) \\ v(1,t) &= u(1,t) \\ u(x,0) &= f_1(x) \\ v(x,0) &= f_2(x). \end{aligned} \tag{41}$$

Note that this system is periodic in time with period 1, and that all energy is preserved. We introduce $t = t(\tau)$ as a stretching of the time coordinate, and define the solution vector as $w = (u, v)$. Then (41) can be written as:

$$\begin{aligned} w_t + t_\tau B w_x &= 0 \quad 0 \leqslant x \leqslant 1 \\ L_1 w(0) &= 0 \\ L_2 w(1) &= 0 \\ w(x,0) &= f(x), \end{aligned}$$

where

$$B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad f(x) = \begin{pmatrix} \sin(2\pi x) \\ -\sin(2\pi x) \end{pmatrix},$$
$$L_1 = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}, \qquad L_2 = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}.$$

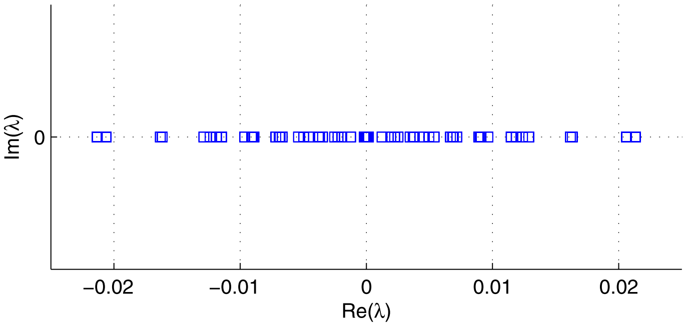

Fig. 5. Eigenvalue distribution of the matrix $(P_\tau T - T P_\tau) \otimes A$ in the case of block norm P_τ, for the resolution $N = M = 15$.

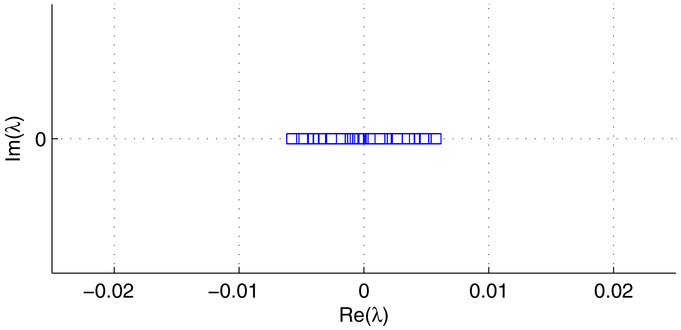

Fig. 6. Eigenvalue distribution of the matrix $(P_\tau T - T P_\tau) \otimes A$ in the case of block norm P_τ, for the resolution $N = M = 30$.

A semi-discrete approximation using the SBP-SAT technique can be formulated as follows:

$$\begin{aligned}\vec{w}_t + t_\tau B \otimes P_\xi^{-1} Q_\xi \vec{w} &= \left(I \otimes P_\xi^{-1}\right)\left(\sigma_1(u_0 - v_0)\vec{e}_0 + \sigma_2(u_M - v_M)\vec{e}_M\right.\\ &\quad \left. + \sigma_3(v_0 - u_0)\vec{d}_0 + \sigma_4(v_M - u_M)\vec{d}_M\right)\\ \vec{w}(0) &= f(\vec{x}),\end{aligned} \tag{42}$$

where $\vec{w} = (u(0), u(\Delta x), \ldots, u(1), v(0), v(\Delta x), \ldots, v(1))^T$. $\vec{e}_0$, $\vec{e}_M$, $\vec{d}_0$ and $\vec{d}_M$ are unit vectors with zeros everywhere except at the position corresponding to u_0, u_M, v_0 and v_M respectively. We can now rewrite (42) on the same form as (39) by setting

$$\tilde{P} = \left(I \otimes P_\xi^{-1}\right), \quad A = B \otimes Q_\xi - \Sigma,$$

where

$$\Sigma = (\sigma_1 L_1 - \sigma_2 L_2) \otimes E_0 + (\sigma_3 L_1 - \sigma_4 L_2) \otimes E_M.$$

With the choice $\sigma_1 = \sigma_4 = -1/2$ and $\sigma_2 = \sigma_3 = 1/2$, the matrix A becomes skew symmetric, leading to strict energy conservation in $\|\cdot\|_{\tilde{P}}$. For the numerical experiments, we used the following stretching of the time coordinate:

$$t = \tau\left(\frac{e^{-(\tau - \frac{4}{5})^2}}{e^{-\frac{1}{25}}}\right)^2, \quad 0 \leqslant \tau \leqslant 1.$$

The system was solved repeatedly for 25 000 periods with the same resolution in both space and time, and using the same stretching of the time coordinate in each step. The spatial part of the problem was discretized using a diagonal norm operator with global order 4, i.e. SBP(6, 3). The time integration was performed with SBP(4, 3) (block norm) as well as SBP(4, 2) (diagonal norm), both with global order 4.

Figs. 5 and 6 show the eigenvalue distribution of the matrix $(P_\tau T - T P_\tau) \otimes A$ in the case of block norms for resolution $N = M = 15$ and $N = M = 30$ in both space and time, while Fig. 7 shows the long term change of energy in the system. We can see that energy growth does indeed occur for the block norm operators. It is interesting to note that on the finer grid, the energy growth rate suddenly accelerates after approximately 15 000 periods, but is very slow before that point. The diagonal norm operators on the other hand produces monotonous energy decay due to the small damping term in the right hand side of (40).

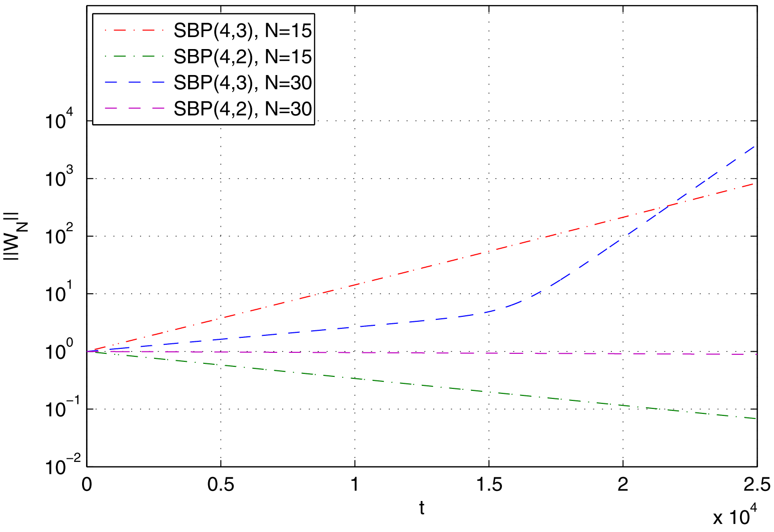

Fig. 7. The long-term change in energy of the fully discrete solution.

9. Summary and conclusions

The SBP-SAT technique applied to time integration of general initial value problem has been analyzed, with focus on the theoretical properties of accuracy and stability. High orders of convergence were proven for both stiff and non-stiff problems. The stability results were proven using the energy method.

It was shown how the SBP-SAT technique for time integration, originally formulated as a global method, can be used with flexibility as a one-step multi-stage method with a variable number of stages, without loss of accuracy compared to the global formulation. Classical stability results, including A-stability, L-stability and B-stability could also be proven using the energy method.

For SBP operators with diagonal norms, it was shown that half of the order of accuracy is lost for very stiff problems, while only one order is lost for block norms. However, non-linear stability could only be proven for diagonal norm operators. Numerical tests on an energy conserving linear problem with a stretched time coordinate showed that the block norm operators can lead to instability for long time integrations.

References

[1] H.-O. Kreiss, G. Scherer, Finite element and finite difference methods for hyperbolic partial differential equations, in: C. De Boor (Ed.), Mathematical Aspects of Finite Elements in Partial Differential Equation, Academic Press, New York, 1974.

[2] B. Strand, Summation by parts for finite difference approximation for d/dx, J. Comput. Phys. 110 (1994) 47–67.

[3] M.H. Carpenter, D. Gottlieb, S. Abarbanel, Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes, J. Comput. Phys. 111 (1994) 220–236.

[4] M. Carpenter, J. Nordström, D. Gottlieb, A stable and conservative interface treatment of arbitrary spatial accuracy, J. Comput. Phys. 148 (1999) 341–365.

[5] J. Nordström, T. Lundquist, Summation-by-parts in time, J. Comput. Phys. 251 (2013) 487–499.

[6] O. Axelsson, Global integration of differential equations through Lobatto quadrature, BIT 4 (1964) 69–86.

[7] F. Costabile, A. Napoli, A method for global approximation of the initial value problem, Numer. Algorithms 27 (2001) 119–130.

[8] B. Guo, Z. Wang, Legendre–Gauss collocation methods for ordinary differential equations, Adv. Comput. Math. 30 (2009) 249–280.

[9] Z. Wang, B. Guo, Legendre–Gauss–Radau collocation method for solving initial value problems of first order ordinary differential equations, J. Sci. Comput. 52 (2012) 226–255.

[10] W. Hundsdorfer, S.J. Ruuth, On monotonicity and boundedness properties of linear multistep methods, Math. Comput. 75 (2006) 655–672.

[11] W. Hundsdorfer, A. Mozartova, M.N. Spijker, Stepsize restrictions for boundedness and monotonicity of multistep methods, J. Sci. Comput. 50 (2012) 265–286.

[12] C.A. Kennedy, M.H. Carpenter, Additive Runge–Kutta schemes for convection–diffusion–reaction equations, Appl. Numer. Math. 44 (2003) 139–181.

[13] M.H. Carpenter, C.A. Kennedy, H. Bijl, S.A. Viken, V.N. Vatsa, Fourth-order Runge–Kutta schemes for fluid mechanics applications, J. Sci. Comput. 25 (2005) 157–194.

[14] J.C. Butcher, Initial value problems: numerical methods and mathematics, Comput. Math. Appl. 28 (1994) 1–16.

[15] J.C. Butcher, General linear methods for stiff differential equations, BIT Numer. Math. 41 (2001) 240–264.

[16] E. Hairer, S. Nørsett, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, Springer-Verlag, 1980.

[17] J. Nordström, Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation, J. Sci. Comput. 29 (2006) 375–404.

[18] J. Nordström, Error bounded schemes for time-dependent hyperbolic problems, SIAM J. Sci. Comput. 30 (2007) 46–59.

[19] K. Mattsson, M. Almquist, A solution to the stability issues with block norm summation by parts operators, J. Comput. Phys. 253 (2013) 418–442.

[20] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems, Springer-Verlag, 1980.

[21] S. Eriksson, J. Nordström, Analysis of the order of accuracy for node-centered finite volume schemes, Appl. Numer. Math. 59 (2009) 2659–2676.

[22] A. Prothero, A. Robinson, On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations, SIAM J. Sci. Comput. 28 (1974) 145–162.

[23] J.E. Hicken, D.W. Zingg, Superconvergent functional estimates from summation-by-parts finite-difference discretizations, SIAM J. Sci. Comput. 33 (2011) 893–922.

[24] J. Berg, J. Nordström, Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form, J. Comput. Phys. 231 (2012) 6846–6860.

[25] J. Berg, J. Nordström, On the impact of boundary conditions on dual consistent finite difference discretizations, J. Comput. Phys. 236 (2013) 41–55.

[26] J.E. Hicken, D.W. Zingg, Summation-by-parts operators and high order quadrature, J. Comput. Appl. Math. 237 (2013) 111–125.