Dual Time-Stepping Using Second Derivatives

Jan Nordström¹ · Andrea A. Ruggiu¹

Received: 18 June 2019 / Revised: 22 August 2019 / Accepted: 3 September 2019
© The Author(s) 2019

Abstract

We present a modified formulation of the dual time-stepping technique which makes use of two derivatives in pseudo-time. This new technique retains and improves the convergence properties to the stationary solution. When compared with the conventional dual time-stepping, the method with two derivatives reduces the stiffness of the problem and requires fewer iterations for full convergence to steady-state. In the current formulation, these positive effects require that an approximation of the square root of the spatial operator is available and inexpensive.

Keywords Dual time-stepping · Convergence acceleration · Steady state · Implicit time-integration · Pseudo time-integration

1 Introduction

The dual time-stepping (DTS) technique can be used for solving a large system of nonlinear equations. The DTS procedure consists in adding a pseudo time-derivative of the solution with respect to the so-called dual time and marching in dual time to steady-state. It was employed in [1] for solving the compressible Euler equations and in [2] for the incompressible Euler and Navier–Stokes equations. Other examples in which derivatives in pseudo time are used to solve systems of nonlinear equations include various engineering fields such as magnetohydrodynamics [3], simulations of launch environments [4] and electronics [5].

One drawback with the dual time-stepping technique is that the pseudo-time iterations must be fully converged in order to preserve time accuracy [6]. Moreover, if the dual time integration is carried out with an explicit scheme, the method may become unstable for dual time-steps exceeding the physical ones [7]. These two limitations may require a large number of iterations and hence to a computationally expensive method.

For these reasons, significant efforts have been made during the last decade to improve the performances of DTS. One strategy to accelerate the convergence is to introduce a preconditioner.

✉ Jan Nordström
jan.nordstrom@liu.se

Andrea A. Ruggiu
andrea.ruggiu@liu.se

¹ Department of Mathematics, Computational Mathematics, Linköping University, 581 83 Linköping, Sweden

tioner multiplying the pseudo-time derivative [2,8,9]. Other improvements can be achieved by developing hybrid discretizations involving the physical-time derivative. In [6,10] the alternating-direction implicit (ADI) scheme [11] is used in conjunction with the common second-order backward difference formula (BDF2). Another example is provided in [12], where the hybrid scheme is built with the lower-upper symmetric-Gauss–Seidel (LU-SGS) method [13]. A further improvement of the DTS is proposed in [14], where it is combined with a local time-stepping approach.

The goal of this paper is to explore if we can accelerate the convergence of DTS by adding a second order pseudo-time derivative. The article is organized as follows: in Sect. 2, the DTS technique is presented and its convergence properties are shown. Section 3 describes a new class of dual time-marching procedures and introduces the second-derivative DTS. In Sect. 4, numerical simulations that corroborate the theoretical results are presented, while in Sect. 5 the drawbacks of the scheme, and alternative formulations are discussed. In Sect. 6 conclusions are drawn.

2 The Dual Time-Stepping Technique

We start by illustrating how the conventional DTS is used.

2.1 A Hyperbolic Model Problem

Consider the one-dimensional advection equation

$$u_t + au_x = 0, \quad x \in \Omega, \quad t > 0, \quad (1)$$

where a is a positive constant and Ω the spatial domain. Let $u_x \approx D\mathbf{u}$ be a general discretization of the spatial derivative, where $\mathbf{u}$ is the vector approximating the solution on a spatial grid. By applying the Euler-backward scheme in time to (1) and indicating the time-step by Δt , we get

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + aD\mathbf{u}^{n+1} = 0. \quad (2)$$

Here, $\mathbf{u}^{n+1}$ and $\mathbf{u}^n$ represent the approximated solution at the different times $t^{n+1} = (n+1)\Delta t$ and $t^n = n\Delta t$, respectively. The calculation of $\mathbf{u}^{n+1}$ by directly inverting the matrix $(I/\Delta t + aD)$ in (2), where I is the identity matrix, may be excessively expensive.

Instead, we apply the DTS technique by replacing $\mathbf{u}^{n+1}$ with $\mathbf{w}$ and adding the dual-time derivative $\mathbf{w}_\tau$ on the left hand-side of (2) to obtain

$$\mathbf{w}_\tau + \frac{\mathbf{w} - \mathbf{u}^n}{\Delta t} + aD\mathbf{w} = 0, \quad \tau > 0. \quad (3)$$

If the solution $\mathbf{w}$ in (3) reaches steady-state, it will converge to $\mathbf{u}^{n+1}$ in (2). The scheme (3) can be rewritten in the following compact form

$$\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}, \quad \tau > 0, \quad (4)$$

where $F = I/\Delta t + aD$ and $\mathbf{R} = \mathbf{u}^n/\Delta t$ is given data.

2.2 Nonlinear Problems

Under mild restrictions, nonlinear differential problems can be related to linear formulations. As an example, consider a fully discretized problem using Euler-backward in time,

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + L(\mathbf{u}^{n+1}) = \mathbf{0}. \quad (5)$$

In (5), $L(\mathbf{u})$ is a nonlinear operator, typically coming from a nonlinear space approximation. Assuming small variations of the solution in time, a linearization of L can be performed:

$$L(\mathbf{u}^{n+1}) = L(\mathbf{u}^n) + \frac{\partial L}{\partial \mathbf{u}} \Delta \mathbf{u} + O(\|\Delta \mathbf{u}\|^2), \quad (6)$$

where $\partial L / \partial \mathbf{u}$ is the Jacobian matrix of L and $\Delta \mathbf{u} := \mathbf{u}^{n+1} - \mathbf{u}^n$. By substituting (6) into (5) and neglecting higher order terms, we obtain

$$\left(\frac{I}{\Delta t} + \frac{\partial L}{\partial \mathbf{u}} \right) \Delta \mathbf{u} \simeq -L(\mathbf{u}^n). \quad (7)$$

The linear problem (7) can be solved using the DTS technique (4), with $F = I / \Delta t + \partial L / \partial \mathbf{u}$ and $\mathbf{R} = -L(\mathbf{u}^n)$. Hence, one can relate nonlinear problems to the linear setting, as long as the Jacobian matrix $\partial L / \partial \mathbf{u}$ is well-defined. The assumption of small variations in time can always be fulfilled by considering sufficiently small time steps Δt .

2.3 Convergence

A general linear time-space discretization of a differential problem has the form

$$F\mathbf{u} = \mathbf{R}, \quad (8)$$

where F is a nonsingular matrix and $\mathbf{R}$ is given and independent of $\mathbf{u}$. To simplify the upcoming analysis, we assume that F is diagonalizable, i.e. $F = X\Lambda X^{-1}$ where X, Λ are the matrices containing the eigenvectors and eigenvalues of F , respectively.

By adding a dual-time derivative to the left hand-side of (8) we obtain (4), which converges in dual time to (8) if the following proposition holds.

Proposition 1 *Let all the eigenvalues of the diagonalizable F have positive real parts. Then the solution of the dual-time dependent problem (4) converges to the solution of (8).*

Proof Applying the eigendecomposition of F to (4) yields

$$\mathbf{v}_\tau + \Lambda \mathbf{v} = X^{-1} \mathbf{R}, \quad \tau > 0, \quad (9)$$

where $\mathbf{v} = X^{-1} \mathbf{w}$. By multiplying (9) with $e^{\Lambda \tau}$ from the left and integrating we find

$$\mathbf{v}(\tau) = e^{-\Lambda \tau} (\mathbf{v}(0) - (X\Lambda)^{-1} \mathbf{R}) + (X\Lambda)^{-1} \mathbf{R},$$

which converges as $\tau \rightarrow +\infty$ if all the eigenvalues of F have positive real parts. The steady-state solution $\mathbf{w} = F^{-1} \mathbf{R}$ is recovered by multiplying $\mathbf{v}$ with X . $\square$

Remark 1 The eigenvalue with the minimum real part determines the convergence rate in (4).

2.4 A Note on Preconditioning

To increase the convergence rate we may introduce a preconditioner Π which multiplies the first-derivative term in (4), yielding

$$\Pi^{-1}\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}. \quad (10)$$

The optimal choice of Π in (10) depends on the specific problem, and will not be discussed in detail in this paper. We simply observe that the choice $\Pi = cF^{-1}$, with $c > 0$, leads to a problem whose convergence does not depend on the eigenvalues of F , since (10) becomes

$$\mathbf{w}_\tau + c\mathbf{w} = cF^{-1}\mathbf{R}. \quad (11)$$

Note that, according to Proposition 1, this formulation is always convergent. On the other hand, even though the magnitude of c can be chosen in order to get a fast convergence of (4), the formulation (11) requires the inverse of F .

2.5 Model Problem

The proof of Proposition 1 indicates that rather than considering the matrix-vector problem (8) at once, one may instead study the scalar model problem

$$w_\tau + \lambda w = r, \quad \tau > 0. \quad (12)$$

The problem (12) is defined by considering each row in (9) separately, with the corresponding steady-state solution

$$u = r/\lambda, \quad \lambda \in \mathbb{C} \setminus \{0\}. \quad (13)$$

3 The Second-Derivative DTS Technique

To possibly get an even faster decay to steady-state, we add two pseudo-time derivatives to the fully-discretized problem (8),

$$\mathbf{w}_{\tau\tau} + 2G\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}, \quad \tau > 0, \quad (14)$$

where G is a matrix to be chosen in order to improve the convergence.

Remark 2 A matrix multiplying the second derivative term in (14) would play the same role as Π^{-1} in (10) for the classical DTS formulation. Hence we consider (14) to be the general second derivative DTS formulation.

We choose a diagonalizable matrix $G = X\Gamma X^{-1}$ in (14) with the same eigenvectors as F . This allows us to rewrite (14) as a system of independent ODEs of the form

$$w_{\tau\tau} + 2\gamma w_\tau + \lambda w = r, \quad \tau > 0, \quad (15)$$

where γ, λ are eigenvalues of G and F , respectively. Note that the steady-state solution of (15) is given by (13). Thus, the convergence properties of the classical and second-derivative DTS can be compared by studying the scalar equations (12) and (15).

The second-order ordinary differential equation (15) can be written as a system of first-order equations

$$\mathbf{z}_\tau + A\mathbf{z} = \mathbf{b}, \quad \text{where } \mathbf{z} = \begin{bmatrix} w \\ w_\tau \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -1 \\ \lambda & 2\gamma \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ r \end{bmatrix}. \quad (16)$$

By using the matrix exponential notation the solution to the system (16)

$$\mathbf{z}(\tau) = e^{-A\tau} (\mathbf{z}(0) - A^{-1}\mathbf{b}) + A^{-1}\mathbf{b} \quad (17)$$

converges to $A^{-1}\mathbf{b} = [u, 0]^T$, i.e. $w(\tau) \rightarrow u$, for any $\mathbf{z}(0)$ as $\tau \rightarrow +\infty$, if the eigenvalues of A have positive real parts.

Remark 3 The matrix exponential $e^{-A\tau}$ can be obtained from the Jordan form of $A = VJV^{-1}$, where V is invertible and J is a triangular matrix composed by Jordan blocks. In particular, $e^{-A\tau} = Ve^{-J\tau}V^{-1}$ and the eigenvalues of A characterize the convergence of (16). For distinct eigenvalues, J and V are the matrices containing the eigenvalues and eigenvectors of A , respectively.

3.1 Initial Convergence Analysis

An interpretation of (15) is given by the damped harmonic oscillator [15], if the coefficients γ and λ are real. This system converges to steady-state if both γ and λ in (15) are positive. Furthermore, the system approaches steady-state as quickly as possible, without oscillating, when it is critically damped, i.e. when $\gamma = \sqrt{\lambda}$. In this section we will prove these results and use them as guidelines for the case with complex coefficients. However, we start by considering $\gamma \in \mathbb{R}$, $\lambda \in \mathbb{R} \setminus \{0\}$.

From Proposition 1, the classical DTS in (12) converges to the steady-state solution as $\tau \rightarrow +\infty$ if $\lambda > 0$. For the second-derivative DTS (15) we prove

Proposition 2 *Let γ and λ be real coefficients. The solution to the problem (15) converges to its steady-state solution as $\tau \rightarrow +\infty$ if γ and λ are positive.*

Proof The solution (17) converges to the steady-state solution if the eigenvalues of A , given by

$$\mu_{1,2} = \gamma \pm \sqrt{\gamma^2 - \lambda}, \quad (18)$$

have positive real parts. If γ and λ are positive, then both real parts of $\mu_{1,2}$ in (18) are positive and convergence follows. $\square$

Next, our aim is to find conditions on γ that lead to faster convergence than the classical DTS technique in (12), i.e. we need

$$\operatorname{Re}(\mu_{1,2}) \geq \operatorname{Re}(\lambda), \quad (19)$$

where μ_1 and μ_2 are given by (18). Condition (19) gives rise to

Proposition 3 *The solution to the unsteady problem (15) converges to steady-state faster than the solution to (12) as $\tau \rightarrow +\infty$ if*

$$(\lambda, \gamma) \in S := \left\{ (\lambda, \gamma) \in \mathbb{R}^2 \mid 0 < \lambda \leq 1, \lambda \leq \gamma \leq \frac{1+\lambda}{2} \right\}. \quad (20)$$

Proof We prove each constraint in S , starting from $\lambda \leq \gamma$. By substituting (18) into (19) and observing that $\gamma, \lambda \in \mathbb{R}$, we find that

$$\gamma \pm \operatorname{Re}\left(\sqrt{\gamma^2 - \lambda}\right) \geq \lambda$$

must hold. Since the real part of the square root is nonnegative for real arguments, it suffices to satisfy the inequality with the minus sign, which is the most restrictive case. Moreover, if $\gamma < \lambda$ the condition

$$\operatorname{Re} \left(\sqrt{\gamma^2 - \lambda} \right) \leq \gamma - \lambda \quad (21)$$

can not be fulfilled and hence $\gamma \geq \lambda$ is required.

Due to the constraint $\gamma \geq \lambda$, we can show that the DTS with two derivatives (15) can be faster than the classical DTS (12) only if $0 < \lambda \leq 1$. We prove this by contradiction: if $\lambda > 1$, then $\gamma^2 \geq \lambda^2 > \lambda$ and (21) becomes

$$\sqrt{\gamma^2 - \lambda} = \operatorname{Re} \left(\sqrt{\gamma^2 - \lambda} \right) \leq \gamma - \lambda.$$

On the other hand, $\gamma \geq \lambda$ implies $2\gamma - \lambda \geq \lambda > 1$ and

$$\gamma - \lambda = \sqrt{\gamma^2 - \lambda} (2\gamma - \lambda) < \sqrt{\gamma^2 - \lambda} \leq \gamma - \lambda \Rightarrow \gamma - \lambda < \gamma - \lambda$$

which proves that $0 < \lambda \leq 1$ is required.

Finally, the remaining inequality in (20) is obtained by considering (21) again. This relation is trivially fulfilled for $\gamma \geq \lambda$. Now, let $\gamma \geq \sqrt{\lambda}$. By squaring (21) we get

$$\lambda [\lambda + (1 - 2\gamma)] \geq 0 \Rightarrow \gamma \leq \frac{1 + \lambda}{2}.$$

□

Proposition 3 provides conditions on the coefficient γ that leads to faster decay of (15) with respect to (12) for any fixed $\lambda \in (0, 1]$. It is legitimate to ask if there exists an optimal choice of γ .

Proposition 4 *The choice $\gamma = \sqrt{\lambda}$ provides the fastest decay for the second-derivative DTS formulation (15).*

Proof The eigenvalue of the matrix A in (16) with the smallest real part determines the decay to the steady-state solution. According to (18), this eigenvalue has a real part given by

$$\operatorname{Re}(\mu_1) = \begin{cases} \gamma, & \text{if } \gamma < \sqrt{\lambda}, \\ \gamma - \sqrt{\gamma^2 - \lambda}, & \text{if } \gamma \geq \sqrt{\lambda}. \end{cases} \quad (22)$$

Since the real part of μ_1 increases for γ less than $\sqrt{\lambda}$ and decreases for γ greater than $\sqrt{\lambda}$, we conclude that $\gamma = \sqrt{\lambda}$ maximizes the real part of μ_1 . □

From (18), the optimal value of γ implies that the eigenvalues of A in (16) are $\mu_1 = \mu_2 = \sqrt{\lambda}$. The optimal DTS formulation (15) hence becomes

$$w_{\tau\tau} + 2\sqrt{\lambda}w_{\tau} + \lambda w = r. \quad (23)$$

This formulation leads to convergence if $\lambda > 0$. Moreover, faster decay with respect to (12) is achieved if $0 < \lambda \leq 1$, since in this case $\sqrt{\lambda} \geq \lambda$. Note that also small perturbations of γ from the optimal value $\sqrt{\lambda}$, i.e. $\gamma \approx \sqrt{\lambda}$, allow for faster convergence, see Fig. 1. In particular, if $\gamma = \sqrt{\lambda} + \delta$, (20) leads to

$$0 < \lambda \leq 1, \quad \sqrt{\lambda}(\sqrt{\lambda} - 1) \leq \delta \leq \frac{(\sqrt{\lambda} - 1)^2}{2}. \quad (24)$$

Fig. 1 Region $\mathcal{S}$ of the values $(\lambda, \gamma) \in \mathbb{R}^2$ in (15) which lead to faster convergence than the conventional DTS (12) according to Proposition 3. The optimal choice $\gamma = \sqrt{\lambda}$ is contained in $\mathcal{S}$ for $0 < \lambda \leq 1$. Note that also $\gamma \approx \sqrt{\lambda}$ may lead to faster convergence

A detailed convergence analysis for (15) with $\gamma, \lambda \in \mathbb{C}$ is beyond the scope of this study. We restrict ourselves to the formulation (23) with $\lambda \in \mathbb{C} \setminus \{0\}$ and prove

Proposition 5 *The solution to the problem (23) converges to its steady-state solution as $\tau \rightarrow +\infty$ if, and only if, λ is not a negative real number.*

Proof The problem (23) can be written as the system of first-order equations (16) with $\gamma = \sqrt{\lambda}$. The eigenvalues of A are $\mu_1 = \mu_2 = \sqrt{\lambda}$ and lead to convergence if $\text{Re}(\mu_{1,2}) > 0$. The number $\sqrt{\lambda}$, interpreted as the principal square root of λ , has always a non-negative real part. If λ is a negative real number, then $\text{Re}(\sqrt{\lambda}) = 0$ which implies no convergence. $\square$

In conclusion, we will consider $G = F^{\frac{1}{2}} = X\Lambda^{\frac{1}{2}}X^{-1}$ as the optimal choice in (14). In $\Lambda^{\frac{1}{2}}$ only the principal square roots are considered, i.e. the square roots with non-negative real parts.

3.2 The New DTS Technique

Consider the new DTS technique applied to the original problem (8)

$$\mathbf{w}_{\tau\tau} + 2F^{\frac{1}{2}}\mathbf{w}_{\tau} + F\mathbf{w} = \mathbf{R}. \quad (25)$$

This formulation generalizes the critically damped harmonic oscillator [15] and converges to steady-state with optimal rate, see Remark 3 and Proposition 4. We can now prove

Proposition 6 *The decay to steady-state for the new DTS formulation (25) is determined by the square roots of the eigenvalues of F .*

Proof The pseudo-time differential problem (25) can be written as a system of first-order equations

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{w}_{\tau} \end{bmatrix}_{\tau} + \begin{bmatrix} I & 0 \\ F^{\frac{1}{2}} & I \end{bmatrix}^{-1} \begin{bmatrix} F^{\frac{1}{2}} & -I \\ 0 & F^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} I & 0 \\ F^{\frac{1}{2}} & I \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_{\tau} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}. \quad (26)$$

Clearly, the convergence of the system is determined by the eigenvalues of $F^{\frac{1}{2}}$. Note that (26) can be rewritten using the auxiliary variable $\mathbf{v} = \mathbf{w}_{\tau} + F^{\frac{1}{2}}\mathbf{w}$, which leads to

Fig. 2 Complex numbers with nonnegative real parts (left figure) and their square root. Note that the distribution of points near the origin of the complex plane tends to rarefy

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix}_\tau + \begin{bmatrix} F^{\frac{1}{2}} & -I \\ 0 & F^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}.$$

□

The main consequence of Proposition 6 is that the new DTS formulation (25) converges to steady-state if the eigenvalues of F are non-zero and do not lie on the negative real axis. If the eigenvalues of F have positive real parts, then both the DTS formulations (4) and (25) are time convergent. In particular, the decay rates are determined by the eigenvalue with the smallest real part of F and $F^{\frac{1}{2}}$, respectively.

Remark 4 The new DTS technique (25) can drive the solution to steady-state, when the classical one (4) fails to do that according to Proposition 5.

Remark 5 The square root of a number close to the imaginary axis has an output which is more distant from it. Similarly, if it is applied to a number with a large magnitude, the square root returns a number less distant from the origin. These two effects are illustrated in Fig. 2.

As pointed out in Remark 5, if F has eigenvalues close to the imaginary axis, the second-derivative DTS decays faster than the classical formulation. Another important effect of the square root is that it narrows the spectrum of F . Pushing the eigenvalues away from the imaginary axis increases the decay rate, while contracting the spectrum enables the use of larger dual time-steps for an explicit time-integrator. Both these characteristics (important for fast convergence) could possibly be obtained by using another matrix G than $F^{\frac{1}{2}}$ in (14).

Remark 6 The real and scalar analysis made in Sect. 3.1 only highlights the first of these two effects for (25). The effect of modifying the spectrum is discussed further below.

4 Numerical Experiments

In this Section we perform numerical tests for both the classical (4) and the new DTS technique (25). In all experiments, the discrete operators are sixth order accurate in the interior and the matrix square root is computed with the MATLAB function `sqrtn` [19].

4.1 First-Order Ordinary Differential Equations

Consider the steady problem

$$\begin{aligned} u_x &= f, \quad 0 < x < 1, \\ u(0) &= g, \end{aligned} \quad (27)$$

where $f(x) = 10\pi \cos(10\pi x)$ and $g = 1$. The analytical solution to (27) is $u(x) = \sin(10\pi x) + 1$.

To discretize (27), we use an $(N+1)$ -point uniform grid over $[0, 1]$, where $x_j = jh$, $j = 0, \dots, N$ and $h = 1/N$. Let $\mathbf{f}$ be a grid function such that $f_j = f(x_j)$ and $\mathbf{u}$ the approximate solution to (27). By applying a Summation-by-Parts (SBP) discretization to (27) for the derivative and a Simultaneous-Approximation-Term (SAT) to impose the boundary condition (see “Appendix A and B” for details and [16] for references), we get

$$P^{-1} Q \mathbf{u} = \mathbf{f} + \sigma P^{-1} \mathbf{e}_0 (u_0 - g), \quad (28)$$

where σ is a penalty parameter and $\mathbf{e}_0 = [1, 0, \dots, 0]^T \in \mathbb{R}^{(N+1)}$. Note that (28) has the form (8) with

$$F = P^{-1} (Q - \sigma E_0), \quad \mathbf{R} = \mathbf{f} - \sigma P^{-1} \mathbf{e}_0 g, \quad E_0 = \text{diag}(1, 0, \dots, 0). \quad (29)$$

The penalty term in (28) makes the classical DTS technique (4) stable in the P -norm $\|\mathbf{w}\|_P = \sqrt{\mathbf{w}^T P \mathbf{w}}$ if $\sigma < -1/2$ (see “Appendix B”). Also, for these values of the penalty parameter the new DTS (25) applied to (28) gives rise to a stable scheme since Proposition 6 holds.

Remark 7 For $\sigma < -1/2$ the classical pseudo-time marching technique (4) is convergent since all the eigenvalues of F have positive real parts. The new DTS formulation also converges since $F^{\frac{1}{2}}$ has only eigenvalues with positive real part [17].

Consider $\sigma = -1$. We use a spatial increment $h = 0.01$ to represent the solution on $[0, 1]$ and the fourth order Runge–Kutta scheme as time-integrator, unless otherwise specified. For both the schemes (4) and (25), we have used $\mathbf{w} = \mathbf{1} := [1, \dots, 1]^T$ as the initial guess, while we have set $\mathbf{w}_\tau$ initially equal to $\mathbf{0}$ for (25). Let $\mathbf{w}^n$ be the solution to either (4) or (25) at the time $\tau^n = n\Delta\tau$. We consider the solution to be converged if $\|\mathbf{w}^n - \mathbf{u}\|_P < 10^{-6}$, where $\mathbf{u}$ is the solution to (28).

The improved convergence can be seen directly by comparing the spectra of F and $F^{\frac{1}{2}}$. From Fig. 3, it is clear that the second-derivatives DTS has better convergence properties since the eigenvalue with minimum real part is further away from the imaginary axis. The minimum number of iterations to convergence for the classical DTS is 177, corresponding to $\Delta\tau = 0.01775$. Figure 4 shows that the new DTS (25) allows for the use of larger dual time-steps, since this formulation is less stiff than the classical one. The minimum number of iterations for the new DTS formulation is 36, which is reached for $\Delta\tau = 0.198$. We conclude that the new DTS formulation is approximately five times more efficient than the old one.

In Fig. 5, we have also tested the classical DTS with two fourth-order accurate Kinnmark–Gray integration methods (with $K = 6$ and $K = 8$) which maximize the allowable dual-time step for hyperbolic problems [18]. For $K = 6$, the optimal dual-time step is $\Delta\tau = 0.0229$ and the convergence is achieved in 135 iterations. Furthermore, the Kinnmark–Gray method with $K = 8$ converges in 152 iterations for $\Delta\tau = 0.0235$. The results show that the new DTS formulation (25) is almost 4 times faster than the classical DTS (4), despite having optimized it with a time-integrator specifically tailored for the problem (27).

Fig. 3 The spectrum of F and of its square root for the first order problem with $\sigma = -1$

Fig. 4 Number of iterations for the classical and new dual-time marching techniques for $\sigma = -1$. The new DTS is less stiff than the former formulation and a larger dual-time-step can be chosen

Fig. 5 Number of iterations for the classical dual-time marching technique for $\sigma = -1$ and two Kinnmark-Gray methods as time-integrators [with $K = 6$ (left) and $K = 8$ (right)]

The convergence results in Fig. 5 are shown in a neighborhood of the optimal dual-time step. It is important to mention that these methods are stable and convergent for $\Delta\tau \leq 0.0309$ and $\Delta\tau \leq 0.0437$, see Fig. 6. In other words, the number of iterations for convergence is not monotonically decreasing with increasing $\Delta\tau$. Hence, the optimal choice of the dual-time step is not necessarily the largest $\Delta\tau$ which leads to a stable numerical method. A similar

Fig. 6 Number of iterations for the classical dual-time marching technique for $\sigma = -1$ and two Kinnmark–Gray methods as time-integrator [with $K = 6$ (left) and $K = 8$ (right)]. All the dual-time steps which lead to convergence are shown

Fig. 7 Number of iterations for the classical and new dual-time marching techniques for $\sigma = -1$. All the dual-time steps which lead to convergence are shown

consideration (although less dramatic) holds for the fourth order Runge–Kutta scheme applied to the classical and the new DTS formulation, see Fig. 7.

Remark 8 The convergence results in Figs. 4 and 5 show that the choice of the time-integration scheme and its fit to the spectrum have a significant influence on the convergence rate of DTS formulations.

Now consider $\sigma = -1/2$. The DTS technique (4) applied to (28) is not provably stable, but all the eigenvalues to the matrix F have positive real parts, as shown in Fig. 8. As a result, both (4) and (25) converge to steady-state. In Fig. 9, the number of iterations to convergence is shown as a function of $\Delta\tau$ for both procedures. The optimal dual time-step for the classical DTS is $\Delta\tau = 0.01778$, leading to 284 iterations. The minimum number of iterations for the new DTS formulation is 36, corresponding to $\Delta\tau = 0.1964$. The new DTS formulation is approximately eight times more efficient than the old one.

For $\sigma = -1/4$ the classical DTS is not energy-stable and F in (29) has eigenvalues with negative real parts. Nonetheless, the second derivative DTS drives the solution to steady-state since no eigenvalue of F lies on the negative real axis (see Proposition 5). This particular situation, predicted by Remark 4, is illustrated in Fig. 10, where both the spectra of F and $F^{\frac{1}{2}}$ are presented. As a result, the classical DTS fails to converge for

Fig. 8 The spectrum of F and of its square root for the first order problem with $\sigma = -1/2$

Fig. 9 Number of iterations for the classical and new dual-time marching techniques for $\sigma = -1/2$. The new DTS is less stiff than the former formulation and a larger dual time-step can be chosen

Fig. 10 The spectrum of F and of its square root for the first order problem with $\sigma = -1/4$

any dual time-step (for example, $\Delta\tau = 0.01$ in Fig. 11). Vice versa, the new DTS technique is convergent and the minimum number of iterations to convergence is 35, reached for $\Delta\tau = 0.1996$.

Fig. 11 In the left figure, the classical DTS fails to drive the solution to steady-state with $\Delta\tau = 0.01$ and $\sigma = -1/4$. In the right figure, the number of iterations needed for convergence of the new dual-time marching technique is shown for $\sigma = -1/4$

4.2 A Model of the Time-Dependent Compressible Navier–Stokes Equations

Next, we study both DTS approaches applied to the following system

$$\begin{aligned} \mathbf{u}_t + \mathbf{A}\mathbf{u}_x &= \varepsilon \mathbf{B}\mathbf{u}_{xx} + \mathbf{F}(x, t), & 0 < x < 1, & t > 0, \\ \mathbf{u}(x, 0) &= \mathbf{f}(x), & 0 < x < 1, & \\ \left(u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x} \right)(0, t) &= g_0(t), & t > 0, & \\ \left(u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x} \right)(1, t) &= g_1(t), & t > 0, & \end{aligned} \quad (30)$$

where $\mathbf{u}(x, t) = [u_1(x, t), u_2(x, t)]^T$, $\varepsilon = 10^{-2}$. The matrices A and B are real and given by

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

while $\mathbf{F}(x, t)$, $\mathbf{f}(x)$, $g_0(t)$, $g_1(t)$ are given data.

The specific boundary conditions $\left(u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x} \right)(0, t) = g_0(t)$ and $\left(u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x} \right)(1, t) = g_1(t)$ applied to the linear Navier–Stokes like system (30) makes the problem strongly well-posed, i.e. a unique solution to (30) exists and its norm is bounded by the boundary and initial data. Moreover, the corresponding semi-discrete problem is strongly stable, if the SBP-SAT approach is used. These theoretical results are shown in “Appendix C and D”.

Here we limit ourselves to the study of the fully-discrete problem

$$\frac{3\mathbf{v}^{n+1} - 4\mathbf{v}^n + \mathbf{v}^{n-1}}{2\Delta t} + D \otimes A\mathbf{v}^{n+1} = \varepsilon D_2 \otimes B\mathbf{v}^{n+1} + \tilde{\mathbf{F}}^{n+1} + \mathbf{SAT}^{n+1}, \quad (31)$$

with $\mathbf{v}^0 = \tilde{\mathbf{f}}$. The formulation (31) is obtained from (30) by discretizing in space with SBP-SAT and using BDF2 in time. This two-step method requires also $\mathbf{v}^1$ as initial data, which is computed using the same space discretization and Euler backward in time.

We consider a grid with $x_j = jh$, $j = 0, \dots, N$ where $h = 1/N$ is the grid spacing, and the grid functions $\tilde{\mathbf{f}}, \tilde{\mathbf{F}}^n \in \mathbb{R}^{2(N+1)}$ which approximate $\mathbf{f}, \mathbf{F}(t^n)$ in the continuous problem (30). With each grid point we associate the approximate solution $\mathbf{v} \in \mathbb{R}^{2(N+1)}$, such that

$$v_{2j}^n \cong u_1(x_j, t^n), \quad v_{2j+1}^n \cong u_2(x_j, t^n), \quad j = 0, \dots, N.$$

In the fully-discrete problem (31), the symbol $\otimes$ denotes the Kronecker product defined by

$$A = \{a_{ij}\} \in \mathbb{R}^{m \times n}, \quad B \in \mathbb{R}^{n \times p}, \quad A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{m \times p}.$$

Moreover, D and D_2 are SBP operators for the first and second derivatives and the vector **SAT** collects the penalty terms for the boundary conditions. The $\mathbf{SAT}^{n+1}$ term in (31) can be written as

$$\begin{aligned} \mathbf{SAT}^{n+1} = & - (P^{-1}E_0 \otimes \Sigma) [(I_{N+1} \otimes H_0) \mathbf{v} - \varepsilon (D \otimes H_D) \mathbf{v} - \tilde{\mathbf{g}}_0]^{n+1} \\ & + (P^{-1}E_N \otimes \Sigma) [(I_{N+1} \otimes H_N) \mathbf{v} - \varepsilon (D \otimes H_D) \mathbf{v} - \tilde{\mathbf{g}}_N]^{n+1}, \end{aligned} \quad (32)$$

where $E_0 = \text{diag}(1, 0, \dots, 0)$, $E_N = \text{diag}(0, \dots, 0, 1)$ and I_M indicates the $M \times M$ identity matrix. Furthermore, we have used

$$\Sigma = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad H_0 = \begin{bmatrix} 1 & \sqrt{2} \\ 1 & \sqrt{2} \end{bmatrix}, \quad H_N = \begin{bmatrix} 1 & -\sqrt{2} \\ 1 & -\sqrt{2} \end{bmatrix}, \quad H_D = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \quad (33)$$

and $\tilde{\mathbf{g}}_0^{n+1} = g_0(t^{n+1}) \mathbf{1}$, $\tilde{\mathbf{g}}_N^{n+1} = g_1(t^{n+1}) \mathbf{1}$ are $2(N+1)$ vectors.

To solve the discrete problem (31) we can write the classical (4) and the new DTS formulation (25) by defining

$$\begin{aligned} F = & \frac{3}{2\Delta t} I_{2(N+1)} + D \otimes A - \varepsilon D_2 \otimes B \\ & + (P^{-1}E_0 \otimes \Sigma) [(I_{N+1} \otimes H_0) - \varepsilon (D \otimes H_D)] \\ & - (P^{-1}E_N \otimes \Sigma) [(I_{N+1} \otimes H_N) - \varepsilon (D \otimes H_D)] \end{aligned} \quad (34)$$

and

$$\mathbf{R} = \frac{2\mathbf{v}^n}{\Delta t} - \frac{\mathbf{v}^{n-1}}{2\Delta t} + (P^{-1}E_0 \otimes \Sigma) \tilde{\mathbf{g}}_0^{n+1} - (P^{-1}E_N \otimes \Sigma) \tilde{\mathbf{g}}_N^{n+1} + \tilde{\mathbf{F}}^{n+1}. \quad (35)$$

To obtain the computational results we have used the following manufactured solutions

$$u_1(x, t) = \cos(10\pi x - t), \quad u_2(x, t) = \sin(10\pi x - t),$$

with a spatial increment $h = 0.005$ and a physical time-step $\Delta t = 0.1$. The new DTS (25) is less stiff than the classical time-marching technique (4) also for this problem, see Fig. 12.

Figure 13 shows the number of iterations needed for convergence with the fourth-order Runge–Kutta scheme as pseudo time-integrator. The optimal dual time-step for the classical DTS (4) is $\Delta\tau = 1.119 \times 10^{-3}$. With the stopping criterion $\|\mathbf{w}^n - \mathbf{u}\|_P < 10^{-6}$ this formulation reaches steady-state in 542 iterations. The optimal choice for the two-derivatives DTS is $\Delta\tau = 0.052$ and it leads to convergence in 57 inner iterations. This implies that the new DTS is approximately ten times more efficient than the classical one.

5 Main Drawbacks and Open Questions

The previous numerical tests show that the new DTS formulation (25) has better convergence properties compared to the conventional time-marching technique (4). However, when we rewrite (25) in first-order form as in (16) we obtain a system which has twice the dimensions of the one in (4). Moreover, the computation of the principal square root of F may be expensive

Fig. 12 The spectrum of F and of its square root for the linearized Navier–Stokes equations

Fig. 13 Number of iterations to convergence using the classical and the new DTS

Table 1 Execution times of the DTS schemes with optimal smoothing step for the steady problem (27)

Nodes (N)	Classical DTS (s)	Second-derivative DTS (s)	Square root (s)
100	0.045337	0.107452	0.044975
1000	2.119714	1.820615	1.767909
2000	12.703043	8.927953	15.609013

The elapsed time for the second-derivative DTS is indicated without the computation of the square root of F . The matrix $F^{\frac{1}{2}}$ is computed with the optimized routine **sqrtm** presented in [19]

if the dimension of the system (8) is large. In Table 1, the computational times of both DTS techniques (4) and (25) are shown for the numerical experiment in Sect. 4.1. The last column provides the elapsed time for computing $F^{\frac{1}{2}}$ with the routine **sqrtm** presented in [19].

If the square root of F is given, then Table 1 shows that when the number of nodes increases, the second-derivative DTS (25) provides a better result with respect to the classical technique (4). However, the computation of the square root becomes expensive. Therefore, we are interested in suboptimal formulations of (14) which do not involve fractional or negative powers of F .

5.1 Alternative Formulations

Our goal is to provide provably convergent DTS schemes of the form (14), but avoid having to compute $F^{\frac{1}{2}}$. This system of second order differential equations can be written as a first-derivative formulation

$$\mathbf{z}_\tau + A\mathbf{z} = \mathbf{b}, \quad \text{where } \mathbf{z} = \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -I \\ F & 2G \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}.$$

Let $K = K(F)$ be a function of F . By choosing $G = (K^{-1}F + K)/2$, we can rotate the system into

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix}_\tau + \begin{bmatrix} I & 0 \\ K^{-1}F & I \end{bmatrix}^{-1} \begin{bmatrix} K^{-1}F & -I \\ 0 & K \end{bmatrix} \begin{bmatrix} I & 0 \\ K^{-1}F & I \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}. \quad (36)$$

The optimal formulation with $G = K = F^{\frac{1}{2}}$ is included in (36) and leads to (26).

There are two straightforward alternatives for K . The first one is $K = \kappa I$ with $\kappa > 0$. This choice gives rise to a convergent formulation with a decay determined by κ and λ/κ , where λ is any eigenvalue of F . If κ is big, the damping of the system is reduced by the scaled eigenvalues λ/κ . However, we would have the same behavior as that of the preconditioned classical DTS (10), with $\Pi = \kappa I$. For small values of κ , every mode of the solution to (36) converges to steady-state uniformly, but slow. The second choice is $K = F$, which leads to the same damping as the classical DTS (4) if all the eigenvalues have real part less than one. Otherwise, the convergence is dominated by the spurious eigenvalues 1. Therefore, these two choices do not lead to an improved formulation compared to the classical DTS (4).

All other alternatives for K that we have investigated lead to a matrix G which involves inverse matrices or fractional powers of F . For this reason, we conclude that the choice $G = (K^{-1}F + K)/2$ in (14) leads to either inefficient or expensive DTS schemes. The existence of alternative formulations not affected by these two effects is still a matter of research.

5.2 Approximations of the Matrix Square Root

Table 1 shows that the direct computation of the square root with the MATLAB function `sqrtn` is costly. A possible alternative is to approximate $F^{\frac{1}{2}}$ through an iterative method, as in [20,21], and to consider an approximation $G = F^{\frac{1}{2}} + \Delta$ such that Δ is small in some norm. In this case, the DTS with two pseudo-time derivative (15) can be rewritten as

$$\mathbf{w}_{\tau\tau} + 2F^{\frac{1}{2}}\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R} - 2\Delta\mathbf{w}_\tau$$

or, equivalently,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix}_\tau + \underbrace{\begin{bmatrix} F^{\frac{1}{2}} & -I \\ 2\Delta F^{\frac{1}{2}} & F^{\frac{1}{2}} - 2\Delta \end{bmatrix}}_M \begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}. \quad (37)$$

If $G = F^{\frac{1}{2}}$ and $\Delta = 0$, this formulation is equivalent to (25), whose convergence is determined by the eigenvalues of $F^{\frac{1}{2}}$. If $\|\Delta\| < \varepsilon$ we may have faster convergence with respect to the classical DTS (4) for small values of ε , as suggested by (24) and Fig. 1.

Fig. 14 The spectra of M in (37) for $\Delta = \xi R$, where R is a random matrix whose entries $r_{ij} \in [-1/2, 1/2]$. The eigenvalues of M are shown for $\xi = 10^{-3}$ (left) and $\xi = 10^{-1}$ (right)

Fig. 15 Number of iterations for convergence using the second derivative DTS with an approximated matrix square root $G = F^{\frac{1}{2}} + \Delta$ in (14) and $\Delta = \xi R$, where R is a random matrix whose entries $r_{ij} \in [-1/2, 1/2]$ and $\xi \in \{10^{-3}, 10^{-1}\}$

As an example, consider the matrix F in (29) with $\sigma = -1$. In Fig. 14 the eigenvalues of M in (37) are shown for $\Delta = \xi R$, where $\xi \in \{10^{-3}, 10^{-1}\}$ and R is a random matrix whose entries $r_{ij} \in [-1/2, 1/2]$. In both cases, the dual time-stepping with two derivatives converges faster than the classical DTS, which required at least 177 iterations. For $\xi = 10^{-3}$ the convergence is achieved in 37 iterations ($\Delta\tau = 0.193$), while the perturbation with $\xi = 10^{-1}$ requires 48 iterations for $\Delta\tau = 0.1778$ (see Fig. 15).

However, the iterative methods that could be used to approximate the matrix square root of F usually require matrix inversion, which is as costly as solving the problem (8). Therefore, also the possibility of approximating the matrix square root requires more research or new ideas.

6 Conclusions and Future Work

A new second-derivative dual-time stepping technique has been proposed. The new DTS technique has been analyzed and optimized theoretically. The formulation involves a matrix

in front of the first derivative in dual time which can be chosen to obtain the highest possible decay rate.

We have compared the performances of the new formulation with the ones of the classical DTS. Our technique improves the decay rate compared to the classical time-marching technique if the eigenvalues of the operator representing the system are near the imaginary axis. Furthermore, if the spectrum is not contained within the unitary circle, the new second-derivatives technique provides a system of equations which is less stiff than the classical DTS formulation.

Numerical computations for a first-order ordinary differential equation and a system modeling the Navier–Stokes equations corroborate the theoretical results. The simulations reveal that the new formulation is more efficient than the standard one as the size of the problem increase, provided that the required matrix $F^{\frac{1}{2}}$ is available.

However, if the computation of $F^{\frac{1}{2}}$ is required, the new DTS formulation is less efficient than the classical dual time-stepping technique. Nonetheless, our findings show that it is theoretically possible to achieve faster convergence to steady-state by employing second derivatives in pseudo-time. Our findings also point to the possibility of using other preconditioners than $F^{\frac{1}{2}}$ in order to produce a suitable spectrum for the time-integration method. This is an interesting avenue for future work.

Acknowledgements Open access funding provided by Linköping University. This work was supported by VINNOVA, the Swedish Governmental Agency for Innovation Systems, under contract number 2013-01209. We thank both reviewers for valuable new information that improved the paper, and gave ideas for future work.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

A SBP-SAT Space Discretization

For the discretization in space of the differential problems we have used the Summation-By-Parts (SBP) operators in conjunction with the Simultaneous-Approximation-Terms (SAT) for the boundary treatment. The main feature of the first is to mimic the property of integration by parts, whereas the second add penalty-like terms that enforces the boundary conditions weakly.

Definition 1 $D = P^{-1}Q$ is a first-derivative SBP operator if P is a symmetric positive definite matrix and $Q + Q^T = H = \text{diag}(-1, 0, \dots, 0, 1)$.

These operators can be built also for the second derivative [22].

Definition 2 $D_2 = P^{-1}(-S^T M + H)S$ is a second derivative SBP operator if M is positive semidefinite and S approximates the first derivative operator at the boundaries.

As an example, choosing $S = P^{-1}Q$ in Definition 2 leads to the so called wide version of D_2 , i.e. $D_2 = D^2$. Both first- and second-derivative SBP operators can be built for even orders $2p$ at the interior, while at the boundary closure their accuracy is p . For further details on the construction of the SBP operators for the first derivative with $p \leq 4$, see [23].

The SBP finite difference operators with a strong treatment of the boundary conditions only admits stability proofs for very simple problems. This result has been shown in [24], where SAT were proposed to enhance the SBP schemes. Discretizing a well-posed initial-boundary-value-problem (IBVP) in space with both SBP operators and the SAT penalty terms (SBP-SAT approach), it is possible to prove that the corresponding semidiscrete problem is stable. Further theoretical details on the SBP-SAT discretization, well-posedness of an IBVP or the stability of its discretization are given in [16].

B Stability of the First Numerical Experiment

In this section we verify the stability of the classical DTS (4) applied to the discretized problem (28) with $\mathbf{f} = \mathbf{0}$ and $\sigma < -1/2$.

For each fixed $\tau > 0$ the dual-time marching technique can be rewritten as

$$\mathbf{w}_\tau = \sigma P^{-1} (w_0 - g) \mathbf{e}_0 - P^{-1} Q \mathbf{w}.$$

The P -norm of the solution $\mathbf{w}$ is $\|\mathbf{w}\|_P = \sqrt{\mathbf{w}^T P \mathbf{w}}$. Thus

$$\begin{aligned} \frac{d}{d\tau} \|\mathbf{w}\|_P^2 &= \mathbf{w}_\tau^T P \mathbf{w} + \mathbf{w}^T P \mathbf{w}_\tau = -\mathbf{w}^T (Q + Q^T) \mathbf{w} + 2\sigma (w_0 - g) w_0 \\ &= -w_N^2 + (1 + 2\sigma) w_0^2 - 2\sigma g w_0 \\ &= -w_N^2 + \frac{1}{1 + 2\sigma} [(1 + 2\sigma) w_0 - \sigma g^2]^2 - \frac{\sigma^2}{1 + 2\sigma} g^2 \leq -\frac{\sigma^2}{1 + 2\sigma} g^2, \end{aligned}$$

where w_0 and w_N are approximations for the solution at the boundaries and the last inequality holds for $\sigma < -1/2$. Since g is a given data, the P -norm of the solution $\mathbf{w}$ is bounded in time. This implies stability of the classical DTS applied to the discretization (28). Equivalently, we have proven that F in (29) has only eigenvalues with non-negative real part, since the energy of the solution to (25) is bounded for any $\tau > 0$.

C Well-Posedness of the Navier–Stokes Model Problem

Consider the model of the compressible Navier–Stokes equations (30). In Sect. 4.2 we claimed that the characteristic boundary conditions make the problem strongly well-posed in the Hadamard sense. To prove this statement we show that (30) admits a unique solution and that the norm of this solution is bounded by the given data $\mathbf{F}(x, t)$, $\mathbf{f}(x)$, $g_0(t)$ and $g_1(t)$.

We start by deriving the characteristic boundary conditions in (30). By premultiplying with $\mathbf{u}^T$ and integrating over $[0, 1]$ we find

$$\begin{aligned} \frac{d}{dt} \|\mathbf{u}(\cdot, t)\|^2 + 2\varepsilon \int_0^1 \mathbf{u}_x^T B \mathbf{u} dx &= \mathbf{u}^T (A \mathbf{u} - 2\varepsilon B \mathbf{u}_x)(0, t) \\ &\quad - \mathbf{u}^T (A \mathbf{u} - 2\varepsilon B \mathbf{u}_x)(1, t) + \int_0^1 \mathbf{u}^T \mathbf{F} dx. \end{aligned} \quad (38)$$

Furthermore, the boundary terms in (38) can be written as

$$\mathbf{u}^T (A \mathbf{u} - 2\varepsilon B \mathbf{u}_x) = \frac{1}{2\sqrt{2}} \left(u_1 + \sqrt{2} u_2 - \varepsilon u_{2,x} \right)^2 - \frac{1}{2\sqrt{2}} \left(u_1 - \sqrt{2} u_2 - \varepsilon u_{2,x} \right)^2.$$

and therefore the boundary conditions

$$\begin{cases} (u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x})(0, t) = g_0(t), & t > 0, \\ (u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x})(1, t) = g_1(t), & t > 0 \end{cases} \quad (39)$$

bound the boundary terms in (38). To prove the boundedness of the solution to (30) we use the Cauchy–Schwarz and Young inequalities with a constant $\eta > 0$ for the integral with the forcing term $\mathbf{F}$ in (38). Moreover, since the matrix B is positive semidefinite,

$$\|\mathbf{u}(\cdot, T)\|^2 \leq e^{\eta T} \left\{ \|\mathbf{f}\|^2 + \int_0^T e^{-\eta t} \left[\frac{1}{2\sqrt{2}} (g_0(t)^2 + g_1(t)^2) + \frac{1}{\eta} \|\mathbf{F}(\cdot, t)\|^2 \right] dt \right\} \quad (40)$$

which proves that the solution to (30) is bounded. The estimate (40) together with the fact that we use the minimal number of boundary conditions implies both existence and uniqueness, and consequently well-posedness.

D Stability of the Semi-Discrete Navier–Stokes Model Problem

In this section we will prove that the discrete energy of the solution to

$$\mathbf{w}_t + D \otimes A \mathbf{w} = \varepsilon D_2 \otimes B \mathbf{w} + \tilde{\mathbf{F}}(\tau) + \mathbf{SAT}, \quad t > 0, \quad (41)$$

is bounded. Without loss of generality we consider the homogeneous problem, i.e. $\tilde{\mathbf{F}} = \tilde{\mathbf{g}}_0 = \tilde{\mathbf{g}}_N = 0$ and $D_2 = D^2$. Let $\mathbb{P} = P \otimes I_2$ and $\|\mathbf{w}\|_{\mathbb{P}} = \sqrt{\mathbf{w}^T \mathbb{P} \mathbf{w}}$. Thus

$$\begin{aligned} \frac{d}{dt} \|\mathbf{w}\|_{\mathbb{P}}^2 &= 2\mathbf{w}^T \mathbb{P} \mathbf{SAT} - \mathbf{w}^T \left((Q + Q^T) \otimes A \right) \mathbf{w} \\ &\quad + \varepsilon (D\mathbf{w})^T (Q^T \otimes B) \mathbf{w} + \varepsilon \mathbf{w}^T (Q \otimes B) D\mathbf{w}. \end{aligned} \quad (42)$$

Making use of the facts that $Q + Q^T = E_N - E_0$ and B is symmetric, we may write

$$\mathbf{w}^T (Q \otimes B) D\mathbf{w} = \mathbf{w}_N^T B (D\mathbf{w})_N - \mathbf{w}_0^T B (D\mathbf{w})_0 - (D\mathbf{w})^T (P \otimes B) D\mathbf{w}, \quad (43)$$

$$(D\mathbf{w})^T (Q^T \otimes B) \mathbf{w} = \mathbf{w}_N^T B (D\mathbf{w})_N - \varepsilon \mathbf{w}_0^T B (D\mathbf{w})_0 - (D\mathbf{w})^T (P \otimes B) D\mathbf{w}, \quad (44)$$

where $\mathbf{w}_0, \mathbf{w}_N, (D\mathbf{w})_0, (D\mathbf{w})_N \in \mathbb{R}^2$ are numerical approximations of the solution and its derivative to the continuous problem (30) at the boundaries, respectively. By combining (42), (43), (44) and the expression of the vector $\mathbf{SAT}$ in (32), we obtain

$$\begin{aligned} \frac{d}{dt} \|\mathbf{w}\|_{\mathbb{P}}^2 &= -2\varepsilon (D\mathbf{w})^T (P \otimes B) D\mathbf{w} \\ &\quad + \mathbf{w}_0^T [(A - 2\Sigma H_0) \mathbf{w}_0 - 2\varepsilon (B - \Sigma H_D) (D\mathbf{w})_0] \\ &\quad - \mathbf{w}_N^T [(A - 2\Sigma H_N) \mathbf{w}_N - 2\varepsilon (B - \Sigma H_D) (D\mathbf{w})_N]. \end{aligned} \quad (45)$$

The right hand-side of (45) can be seen as summation of three terms: the first one is non-positive, since $P \otimes B$ is positive semidefinite. The last two contributions are boundary terms, which can be expressed as

$$\begin{bmatrix} \mathbf{w}_i \\ \varepsilon (D\mathbf{w})_i \end{bmatrix}^T \begin{bmatrix} A - 2\Sigma H_i & - (B - \Sigma H_D) \\ - (B - \Sigma H_D) & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \varepsilon (D\mathbf{w})_i \end{bmatrix} = \mathbf{y}_i^T C_i \mathbf{y}_i, \quad i = \{0, N\}. \quad (46)$$

Inserting the expression of H_0 , H_N , H_D and Σ in (33) into (46) leads to

$$C_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & -2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad C_N = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since C_0 is negative semidefinite and C_N positive semidefinite, the estimate (45) implies that the energy of the solution decreases and proves the stability of the problem (41).

References

1. Jameson, A.: Time dependent calculations using multigrid, with applications to unsteady flows past airfoils and wings. In: AIAA paper 91-1596, AIAA 10th Computational Fluid Dynamics Conference. Honolulu, Hawaii (1991)
2. Belov, A., Martinelli, L., Jameson, A.: A new implicit algorithm with multigrid for unsteady incompressible flow calculations. In: AIAA Paper 95-0049, AIAA 33rd Aerospace Sciences Meeting. Reno, Nevada (1995)
3. Knoll, D.A., McHugh, P.R.: Enhanced nonlinear iterative techniques applied to a nonequilibrium plasma flow. *SIAM J. Sci. Comput.* **19**(1), 291–301 (1998)
4. Housman, J.A., Barad, M.F., Kiris, C.C.: Space-time accuracy assessment of CFD simulations for the launch environment. In: 29th AIAA Applied Aerodynamics Conference (2011)
5. Grasser, T., Selberherr, S.: Mixed-mode device simulation. *Microelectron. J.* **31**, 873–881 (2000)
6. Hsu, J.M., Jameson, A.: An implicit–explicit hybrid scheme for calculating complex unsteady flows. In: 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, 2002-0714 (2002)
7. Arnone, A., Liou, M.-S., Povinelli, L.A.: Multigrid time-accurate integration of Navier–Stokes equations. In: AIAA-93-3361 (1993)
8. Pandya, S.A., Venkateswaran, S., Pulliam, T.H.: Implementation of preconditioned dual-time procedures in overflow. In: AIAA Paper 2003-0072 (2003)
9. Helenbrook, B.T., Cowles, G.W.: Preconditioning for dual-time-stepping simulations of the shallow water equations including Coriolis and bed friction effects. *J. Comput. Phys.* **227**(9), 4425–4440 (2008)
10. Marongiu, C., et al.: An Improvement of the dual time stepping technique for unsteady RANS computations. In: Conference Paper, European Conference for Aerospace Sciences (EUCASS), Moscow (2005)
11. Peaceman, D.W., Rachford Jr., H.H.: The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.* **3**(1), 28–41 (1955)
12. Dwight, R.P.: Time-accurate Navier–Stokes calculations with approximately factored implicit schemes. In: Computational Fluid Dynamics 2004, Proceedings of the 3rd International Conference on Computational Fluid Dynamics, ICCFD3, Toronto, 12–16 July 2004, pp. 211–217 (2006)
13. Yoon, S., Jameson, A.: An LU-SSOR scheme for the Euler and Navier–Stokes equations. *AIAA J.* **26**, 1025–1026 (1988)
14. Bevan, R.L., Nithiarasu, P.: Accelerating incompressible flow calculations using a quasi-implicit scheme: local and dual time stepping approaches. *Comput. Mech.* **50**(6), 687 (2012)
15. Halliday, D., Resnick, R., Walker, J.: Principles of Physics, 10th edn. Wiley, Hoboken (2014)
16. Svård, M., Nordström, J.: Review of summation-by-parts schemes for initial-boundary-value problems. *J. Comput. Phys.* **268**, 17–38 (2014)
17. Nordström, J., Lundquist, T.: Summation-by-parts in time. *J. Comput. Phys.* **251**, 487–499 (2013)
18. Kinnmark, I.P.E., Gray, W.G.: One step integration methods of third-fourth order accuracy with large hyperbolic stability limits. *Math. Comput. Simul.* **26**(3), 181–188 (1984)
19. Higham, N.J.: A New *sqrtrm* for MATLAB. Numerical Analysis Report No. 336, Manchester Centre for Computational Mathematics, Manchester, England (1999)
20. Amat, S., Ezquerro, J.A., Hernández-Verón, M.A.: Iterative methods for computing the matrix square root. *SeMA J.* **70**(1), 11–21 (2015)
21. Sadeghi, A.: Approximating the principal matrix square root using some novel third-order iterative methods. *Ain Shams Eng. J.* **9**(4), 993–999 (2018)
22. Mattson, K., Nordström, J.: Summation by parts operators for finite difference approximations of second derivatives. *J. Comput. Phys.* **199**, 503–540 (2004)

23. Strand, B.: Summation by parts for finite difference approximations for d/dx . J. Comput. Phys. **110**, 47–67 (1994)
24. Carpenter, M.H., Gottlieb, D., Abarbanel, S.: Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes. J. Comput. Phys. **111**(2), 220–236 (1994)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.