BIT Numerical Mathematics (2024) 64:14
https://doi.org/10.1007/s10543-024-01007-y

BIT

The variable two-step BDF method for parabolic equations

Georgios Akrivis[1,2] · **Minghua Chen**[3] · **Jianxing Han**[3] · **Fan Yu**[3] · **Zhimin Zhang**[4]

Received: 31 July 2023 / Accepted: 5 January 2024

Abstract
The two-step backward difference formula (BDF) method on variable grids for parabolic equations with self-adjoint elliptic part is considered. Standard stability estimates for adjacent time-step ratios $r_j := k_j/k_{j-1} \le 1.8685$ and 1.9104, respectively, have been proved by Becker (BIT 38:644–662, 1998) and Emmrich (J Appl Math Comput 19:33–55, 2005) by the energy technique with a single multiplier. Even slightly improving the ratio is cumbersome. In this paper, we present a novel technique to examine the positive definiteness of banded matrices that are neither Toeplitz nor weakly diagonally dominant; this result can be viewed as a variant of the Grenander–Szegő theorem. Then, utilizing the energy technique with two multipliers, we establish stability for adjacent time-step ratios up to 1.9398.

Keywords Two-step BDF method · Variable step-size · Stability estimate · Parabolic equations

Mathematics Subject Classification 65L06 · 65M12

1 Introduction

Let $T > 0, u^0 \in H$, and consider the initial value problem of seeking $u \in C((0,T];\mathscr{D}(A)) \cap C([0,T];H)$ satisfying

Communicated by Axel Målqvist.

The work of the second author was supported by the Science Fund for Distinguished Young Scholars of Gansu Province under Grant No. 23JRRA1020 and the Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2023-06.

✉ Georgios Akrivis
akrivis@cse.uoi.gr

Extended author information available on the last page of the article

Published online: 01 March 2024

$$\begin{cases} u'(t) + Au(t) = f(t), \quad 0 < t < T, \\ u(0) = u^0, \end{cases} \tag{1.1}$$

with A a positive definite, selfadjoint, linear operator on a Hilbert space $(H, (\cdot, \cdot))$ with domain $\mathscr{D}(A)$ dense in H and $f : [0, T] \to H$ a given forcing term.

The backward difference formula (BDF) methods are popular for stiff differential equations, in particular, for parabolic equations. They are frequently implemented on nonuniform partitions for numerical efficiency.

For an integer $N \geq 2$, consider a partition $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $[0, T]$, with time steps $k_n := t_n - t_{n-1}$, $n = 1, \ldots, N$. We recursively define a sequence of approximations u^n to the nodal values $u(t_n)$ of the exact solution by the variable two-step BDF method,

$$D_2 u^n + Au^n = f^n, \quad n = 2, \ldots, N, \tag{1.2}$$

with $f^n := f(t_n)$, assuming that arbitrary starting approximations u^0 and u^1 are given. Here,

$$D_2 v^n := \left(1 + \tfrac{k_n}{k_{n-1}}\right) \frac{v^n - v^{n-1}}{k_n} - \frac{k_n}{k_{n-1}} \frac{v^n - v^{n-2}}{k_n + k_{n-1}}.$$

Let $|\cdot|$ denote the norm on H induced by the inner product $(\cdot, \cdot)$, and introduce on V, $V := \mathscr{D}(A^{1/2})$, the norm $\|\cdot\|$ by $\|v\| := |A^{1/2} v|$. We identify H with its dual, and denote by V' the dual of V, and by $\|\cdot\|_\star$ the dual norm on V', $\|v\|_\star = |A^{-1/2} v|$. We shall use the notation $(\cdot, \cdot)$ also for the antiduality pairing between V' and V. For simplicity, we denote by $\langle \cdot, \cdot \rangle$ the inner product on V, $\langle v, w \rangle := (A^{1/2} v, A^{1/2} w)$.

1.1 Main result

We establish the following stability result.

Theorem 1.1 (Stability estimate) *Let u^n satisfy (1.2), with $u^0, u^1 \in V$, and assume that*

$$r_n := \frac{k_n}{k_{n-1}} \leq r^\star \approx 1.9398, \quad n = 2, \ldots, N; \tag{1.3}$$

the bound $r^\star$ is expressed in terms of the multipliers $\delta = 0.9672$ and $\eta = -0.1793$ in (3.1); see also (4.17) for more precise values of the bound $r^\star$ as well as of the multipliers. Then, the variable two-step BDF method (1.2) is stable in the sense that

$$\begin{aligned} &|u^n|^2 + \sum_{j=2}^{n} k_j \|u^j\|^2 \\ &\quad \leq C \mathrm{e}^{c\Gamma_n} \left(|u^0|^2 + |u^1|^2 + k_2 \|u^0\|^2 + k_2 \|u^1\|^2 + \sum_{j=2}^{n} k_j \|f^j\|_\star^2 \right), \end{aligned} \tag{1.4}$$

$n = 2, \dots, N$. *Here,* Γ_n *is a mesh-dependent quantity,*

$$\Gamma_n := \sum_{j=2}^{n-2} [r_j - r_{j+2}]_+ \quad \textit{with} \quad [x]_+ := \max(x, 0), \tag{1.5}$$

and C, c *denote generic constants, independent of* T *and the operator* A *as well as of* f *and of the partition of the time interval.*

Let us recall some partitions for which Γ_n is finite; see [12, p. 175]. If the sequence of the ratios (r_n) is monotone (and bounded), then Γ_n is bounded; more precisely, $\Gamma_n = 0$ if (r_n) is nondecreasing, and $\Gamma_n = r_2 + r_3 - r_{n-1} - r_n$ if (r_n) is decreasing. More generally, Γ_n is bounded in the practically reasonable case that the number of changes in monotonicity of the sequence (r_n) is bounded, uniformly with respect to the number N of time steps. For partitions of the form $t_i = (i/N)^\alpha$, with $\alpha > 1$, the time steps k_i increase and the ratios r_i decrease to 1, whence, in particular, $r_i \le r^\star$ except for a finite number of i.

1.2 Main ingredients of the proof

We shall use the energy technique. Let $r_n = k_n/k_{n-1}$, $n = 2, \dots, N$, be the adjacent time step ratios. With the notation

$$\delta_k v^n := v^n - v^{n-k}, \quad \omega_n := \frac{1}{1 + r_n}, \quad \psi_n := \left(\frac{r_n}{1 + r_n}\right)^2,$$

the backward difference quotient $D_2 v^n$ can be written in the form (cf. [2])

$$D_2 v^n = \frac{1}{\omega_n k_n}\big(\delta_1 v^n - \psi_n \delta_2 v^n\big). \tag{1.6}$$

Testing the BDF method (1.2) by $2\omega_n k_n \left(u^n - \delta u^{n-1} - \eta u^{n-2}\right)$, with $0 < \delta < 1$ and $-1 < \eta < 0$ two multipliers to be suitably chosen below, we obtain

$$\mathscr{D}_n + \mathscr{A}_n = \mathscr{F}_n, \quad n = 2, \dots, N, \tag{1.7}$$

with

$$\begin{cases} \mathscr{D}_n := 2\omega_n k_n \left(D_2 u^n, u^n - \delta u^{n-1} - \eta u^{n-2}\right), \\ \mathscr{A}_n := 2\omega_n k_n \langle u^n, u^n - \delta u^{n-1} - \eta u^{n-2}\rangle, \\ \mathscr{F}_n := 2\omega_n k_n \big(f^n, u^n - \delta u^{n-1} - \eta u^{n-2}\big). \end{cases} \tag{1.8}$$

The terms $\mathscr{F}_n$ on the right-hand side of (1.7), accounting for the forcing term f, can be easily estimated from above by the generalized Cauchy–Schwarz and the weighted arithmetic–geometric mean inequalities. We shall estimate $\mathscr{D}_n$ from below,

and subsequently the sum over all $\mathscr{D}_n$, in Sect. 3.1, while in Sect. 3.2 we shall directly estimate the sum over all $\mathscr{A}_n$ from below rather than each term $\mathscr{A}_n$ separately. The key point in the estimate of the sum over all $\mathscr{A}_n$ is the positive definiteness of families of certain banded matrices; this property is described and established in Sect. 2.

1.3 Previous work

Stability of the A-stable two-step BDF method for parabolic equations for equidistant partitions can be easily established by the energy technique. The zero-stability property, and thus the stability for o.d.e's satisfying the Lipschitz condition, of the variable two-step BDF method is also well-understood; a sufficient condition is $r^\star < 1 + \sqrt{2} \approx 2.414$ in (1.3) and the bound is sharp; see [4, 7] as well as [10, p. 405]. In contrast, the analysis of the variable two-step BDF method for parabolic equations is cumbersome and still incomplete.

Grigorieff proved stability for linear parabolic equations, with bounds independent of Γ_n, for $r^\star \le (1 + \sqrt{3})/2 \approx 1.366$ in [8, 9]. In [2], Becker established stability of the form (1.4) and derived error estimates for linear parabolic equations for $r^\star \le (2 + \sqrt{13})/3 \approx 1.8685$; see also [12, pp. 174–180]. Emmrich [5] further relaxed the bound to 1.9104 for semilinear parabolic equations. For stability estimates for the three-step BDF method, with a mesh-dependent quantity similar to Γ_n, we refer to [3].

In [2, 12] and [5] the method is tested by linear combinations of two terms, u^n and u^{n-1}; here, to relax the condition on the ratios, as we mentioned, we test by linear combinations of all three terms that enter into the method, namely, of u^n, u^{n-1}, and u^{n-2}. Furthermore, we directly estimate the sum of the terms accounting for the elliptic operator from below; this is in sharp contrast to Becker [2], Thomée [12], Emmrich [5], where each one of these terms is estimated separately; see Sect. 3.2.

Several stability estimates of a different kind, in which the difference quotient $(u^1 - u^0)/k_1$ enters on the right-hand side, have been recently established both for linear and nonlinear parabolic equations, for bounds $r^\star$ significantly larger than the optimal bound $1 + \sqrt{2}$ for zero-stability; see [11] and references therein. Notice that $(u^1 - u^0)/k_1$ may enter implicitly, if, for instance, the starting value u^1 is computed by employing one step of the trapezoidal method.

We establish key auxiliary results in Sect. 2 and provide the proof of Theorem 1.1 in Sect. 3. We motivate the choice of the multipliers δ and η in Sect. 4.

2 Auxiliary results

Our main tool in the proof of the stability result in Theorem 1.1 will be the positive definiteness of families of certain banded matrices. This property will allow us to suitably estimate from below the sum over all terms $\mathscr{A}_n$ entering into (1.7).

For given real numbers δ and $\eta \le 0$, we are interested in properties of families of banded lower triangular $(n-1)\times(n-1)$ real matrices of the form

$$\mathbb{L}(r_2,\dots,r_n) := \begin{pmatrix} \frac{1}{1+r_2} & & & \\ -\delta\frac{\sqrt{r_3}}{1+r_3} & \frac{1}{1+r_3} & & \\ -\eta\frac{\sqrt{r_3 r_4}}{1+r_4} & -\delta\frac{\sqrt{r_4}}{1+r_4} & \frac{1}{1+r_4} & \\ & \ddots & \ddots & \ddots \\ & & -\eta\frac{\sqrt{r_{n-1}r_n}}{1+r_n} & -\delta\frac{\sqrt{r_n}}{1+r_n} \quad \frac{1}{1+r_n} \end{pmatrix} \tag{2.1}$$

with positive $r_2,\dots,r_n \le r$, with a uniform positive upper bound r, for all $n \ge 4$.

Lemma 2.1 (Property of matrices of the form (2.1)) *Let $(\cdot,\cdot)_2$ and $\|\cdot\|_2$ denote the Euclidean inner product and norm, respectively, on $\mathbb{R}^{n-1}$, and let c be a real constant. Then,*

$$(\mathbb{L}(r_2,\dots,r_n)x, x)_2 \ge c\|x\|_2^2 \quad \forall x \in \mathbb{R}^{n-1}, \tag{2.2}$$

for all matrices of the form (2.1) and for all $n \ge 4$, if and only if

$$p(y) = \frac{1}{1+r}\big[1+\eta r - \delta\sqrt{r}\,y - 2\eta r y^2\big] \ge c \quad \forall y \in [-1,1]. \tag{2.3}$$

As we shall see later on, the necessity of (2.3) is an easy consequence of well-known properties of the spectrum of symmetric, banded Toeplitz matrices.

Proof First, we shall prove that condition (2.3) implies the estimate (2.2). With

$$J := \begin{pmatrix} \frac{1}{1+r_2} & & & \\ & \frac{1}{1+r_3} & & \\ & & \ddots & \\ & & & \frac{1}{1+r_n} \end{pmatrix} \quad \text{and} \quad G := \begin{pmatrix} 0 & & & \\ \sqrt{r_3} & 0 & & \\ & \ddots & \ddots & \\ & & \sqrt{r_n} & 0 \end{pmatrix}$$

the matrix $\mathbb{L} := \mathbb{L}(r_2,\dots,r_n)$ in (2.1) can be rewritten as

$$\mathbb{L} = J - \delta JG - \eta JG^2.$$

It suffices to consider the symmetric part $\mathbb{L}_s$ of the matrix $\mathbb{L}$,

$$\mathbb{L}_s = \frac{1}{2}(\mathbb{L}+\mathbb{L}^\top) = J - \frac{\delta}{2}(JG + G^\top J) - \frac{\eta}{2}\big(JG^2 + (G^\top)^2 J\big),$$

since $(\mathbb{L}x,x)_2 = (\mathbb{L}_s x, x)_2$. With $K := J^{1/2}$, we have

$$2K^{-1}\mathbb{L}_s K^{-1} = 2I - \delta\big(KGK^{-1} + K^{-1}G^\top K\big) - \eta\big(KG^2K^{-1} + K^{-1}(G^\top)^2 K\big).$$

Letting

$$P := KGK^{-1} = \begin{pmatrix} 0 & & & \\ \sqrt{\frac{1+r_2}{1+r_3}}\, r_3 & 0 & & \\ & \ddots & \ddots & \\ & & \sqrt{\frac{1+r_{n-1}}{1+r_n}}\, r_n & 0 \end{pmatrix},$$

we can rewrite $2K^{-1}\mathbb{L}_s K^{-1}$ in the form

$$2K^{-1}\mathbb{L}_s K^{-1} = 2I - \delta(P + P^\top) - \eta\big(P^2 + (P^\top)^2\big),$$

i.e.,

$$2K^{-1}\mathbb{L}_s K^{-1} = 2I - \delta\sqrt{r}\,\frac{P + P^\top}{\sqrt{r}} - \eta r\,\frac{P^2 + (P^\top)^2}{r}.$$

Therefore, with

$$Z := \frac{P}{\sqrt{r}} = \begin{pmatrix} 0 & & & \\ z_3 & 0 & & \\ & \ddots & \ddots & \\ & & z_n & 0 \end{pmatrix},$$

we have

$$2K^{-1}\mathbb{L}_s K^{-1} = 2I - \delta\sqrt{r}(Z + Z^\top) - \eta r\big(Z^2 + (Z^\top)^2\big).$$

Using here the identity $Z^2 + (Z^\top)^2 = (Z + Z^\top)^2 - ZZ^\top - Z^\top Z$, we see that

$$2K^{-1}\mathbb{L}_s K^{-1} = 2M - \eta r(2I - ZZ^\top - Z^\top Z) \tag{2.4}$$

with the symmetric matrix M,

$$M := (1 + \eta r)I - \delta\sqrt{r}Z_s - 2\eta r Z_s^2, \tag{2.5}$$

where $Z_s := (Z + Z^\top)/2$ is the symmetric part of the matrix Z.

Since $\frac{r_i}{1+r_i} \le \frac{r}{1+r}$ and $1 + r_{i-1} \le 1 + r$, we have $z_i = \sqrt{\frac{r_i}{1+r_i}\,\frac{1+r_{i-1}}{r}} \le 1$,

$$0 < z_i \le 1, \quad i = 3, \ldots, n. \tag{2.6}$$

To prove (2.2), we shall proceed in two steps: first we shall show that (2.6) implies that the diagonal matrix $2I - ZZ^\top - Z^\top Z$ is positive semidefinite, and subsequently,

using the Rayleigh quotient criterion, that the eigenvalues of the matrix M are bounded from below by $c(1+r)$.

Now,

$$ZZ^\top = \begin{pmatrix} 0 & & & \\ & z_3^2 & & \\ & & \ddots & \\ & & & z_n^2 \end{pmatrix} \quad \text{and} \quad Z^\top Z = \begin{pmatrix} z_3^2 & & & \\ & \ddots & & \\ & & z_n^2 & \\ & & & 0 \end{pmatrix},$$

and, thus, the matrix $2I - ZZ^\top - Z^\top Z$ is diagonal. In view of (2.6), its diagonal entries are nonnegative; consequently, this matrix is indeed positive semidefinite. Notice also that $\eta \le 0$.

To complete the proof of (2.2), it remains to show that the eigenvalues of the symmetric matrix M are bounded from below by $c(1+r)$. Now, the eigenvalues μ_i and λ_i of the symmetric matrices M and Z_s, respectively, are related by

$$\mu_i = 1 + \eta r - \delta\sqrt{r}\lambda_i - 2\eta r\lambda_i^2 = (1+r)\,p(\lambda_i); \tag{2.7}$$

see (2.5) and (2.3).

Let us first show that $\lambda_i \in [-1, 1]$ via the Rayleigh quotient criterion. Indeed, for $y = (y_2, y_3, \dots, y_n)^\top \in \mathbb{R}^{n-1}$, we have

$$(Z_s y, y)_2 = \sum_{i=3}^{n} z_i y_i y_{i-1},$$

whence, in view of (2.6),

$$|(Z_s y, y)_2| \le \frac{1}{2}\sum_{i=3}^{n}\big((y_{i-1})^2 + (y_i)^2\big) = \|y\|_2^2 - \frac{1}{2}\big[(y_2)^2 + (y_n)^2\big].$$

Therefore,

$$|\lambda_i| \le \sup_{\substack{y\in\mathbb{R}^{n-1}\\ y\neq 0}} \frac{|(Z_s y, y)_2|}{\|y\|_2^2} \le 1.$$

Now, it follows immediately from (2.3) and (2.7) that the eigenvalues μ_i of the symmetric matrix M are bounded from below by $c(1+r)$. Thus, for $x \in \mathbb{R}^{n-1}$,

$$\big(K^{-1}\mathbb{L}_s K^{-1}x, x\big)_2 \ge (Mx, x)_2 \ge c(1+r)\|x\|_2^2,$$

which, in combination with $\|K^{-1}x\|_2^2 \le (1+r)\|x\|_2^2$, yields the asserted estimate (2.2).

Next, we prove that condition (2.3) is necessary for (2.2).

It suffices to show that condition (2.3) is necessary for (2.2) for all matrices of the form (2.1) with $r_2 = \cdots = r_n = r$. The symmetric part $\mathbb{L}_s(r, \ldots, r) := \big(\mathbb{L}(r, \ldots, r) + \mathbb{L}(r, \ldots, r)^\top\big)/2$ of the $(n-1) \times (n-1)$ matrix $\mathbb{L}(r, \ldots, r)$ is a symmetric pentadiagonal Toeplitz matrix with generating function g (see [1, 6]),

$$g(x) := \frac{1}{1+r}\big[1 - \delta\sqrt{r}\cos x - \eta r \cos(2x)\big], \quad x \in \mathbb{R}.$$

Now, with p the polynomial of (2.3) and the change of variables $y = \cos x$, we have

$$g_{\min} := \min_{x\in\mathbb{R}} g(x) = \min_{-1\le y\le 1} p(y).$$

Assume that (2.3) is not satisfied; then, we would have $g_{\min} < c$. From Theorem 2.1, a simplified version of more general results for symmetric banded Toeplitz matrices, we would then infer that the matrices $\mathbb{L}_s(r, \ldots, r)$ possess, for sufficiently large dimension, eigenvalues less than c, a contradiction to (2.2). □

Theorem 2.1 (Grenander–Szegő theorem, and asymptotic behavior of extreme eigenvalues of symmetric, banded Toeplitz matrices; cf. [6, Theorems 6.1 and 6.6]) *Let g be a nonconstant, real and even, 2π-periodic, trigonometric polynomial. Then, the eigenvalues of all symmetric, banded, $n \times n$ Toeplitz matrices T_n, with generating function g, belong to the open interval $(g_{\min}, g_{\max})$ with $g_{\min}$ and $g_{\max}$ the minimum and maximum of g, respectively.*

Let $\lambda_1(T_n) \ge \lambda_2(T_n) \ge \cdots \ge \lambda_n(T_n)$ be the eigenvalues of T_n sorted in nonincreasing order. Then, for each fixed integer $j \ge 1$, we have

$$\lim_{n\to\infty} \lambda_j(T_n) = g_{\max} \quad \textit{and} \quad \lim_{n\to\infty} \lambda_{n-j+1}(T_n) = g_{\min}.$$

Remark 2.1 The Grenander–Szegő theorem applies to Toeplitz matrices; see the first part of Theorem 2.1. Here, Lemma 2.1 can be viewed as a variant of the Grenander–Szegő theorem, applicable to a class of non-Toeplitz matrices.

3 Proof of Theorem 1.1

In this section, we prove Theorem 1.1.

Let us first recall a discrete version of Gronwall's lemma that we will need in the sequel.

Lemma 3.1 (Discrete Gronwall inequality; Emmrich, [5]) *Let $\alpha_n, \beta_n, \xi_n, \varphi_n$ be nonnegative numbers, with a monotonically increasing sequence $(\xi_n)_{n\ge 2}$, satisfying the inequalities*

$$\alpha_n + \beta_n \le \sum_{i=2}^{n-1} \varphi_i \alpha_i + \xi_n, \quad n = 2, 3, \ldots .$$

Then, the following estimate is valid

$$\alpha_n + \beta_n \le \xi_n \exp\left(\sum_{i=2}^{n-1} \varphi_i\right), \quad n = 2, 3, \dots .$$

3.1 Estimation of the terms accounting for the difference quotient

Let us first focus on the first term on the left-hand side of (1.7).

Lemma 3.2 (Estimation of $\mathscr{D}_n$) *Assume that* $0 < \delta < 1$, $-1 < \eta < 0$ *with* $2-\delta+2\eta \ge 0$, $1+\delta+3\eta \ge 0$, *and* $r_j \le r$, $j = 2, \dots, N$, *with* r *such that*

$$r \le \frac{\sqrt{1+\delta+3\eta}}{2\sqrt{1+\eta} - \sqrt{1+\delta+3\eta}} =: r^\star(\delta, \eta).^{1} \tag{3.1}$$

Then,

$$\sum_{j=2}^{n} \mathscr{D}_j \ge (1-\delta-\eta)\Big((1-\psi_n)|u^n|^2 - \psi_{n-1}|u^{n-1}|^2 - |u^1|^2 - \sum_{j=2}^{n-2}[\psi_j - \psi_{j+2}]_+ |u^j|^2\Big) - \big[-\eta + (2-\delta+2\eta)\,\psi_2\big]|\delta_1 u^1|^2, \tag{3.2}$$

$n = 3, \dots, N$. *For* $n = 2$, *(3.2) is also valid without the second and fourth terms on the right-hand side, i.e.,*

$$\mathscr{D}_2 \ge (1-\delta-\eta)\Big((1-\psi_2)|u^2|^2 - |u^1|^2\Big) - \big[-\eta + (2-\delta+2\eta)\,\psi_2\big]|\delta_1 u^1|^2.$$

Proof We shall estimate each term $\mathscr{D}_j$ from below separately and subsequently sum over j to obtain (3.2).

Using (1.6) and expanding $\mathscr{D}_n$ in (1.8), we have

$$\mathscr{D}_n = I_1^n + I_2^n + I_3^n + I_4^n + I_5^n + I_6^n \tag{3.3}$$

with

$$\begin{cases} I_1^n = 2(\delta_1 u^n, u^n), & I_2^n = -2\psi_n(\delta_2 u^n, u^n), & I_3^n = -2\delta(\delta_1 u^n, u^{n-1}), \\ I_4^n = 2\delta\psi_n(\delta_2 u^n, u^{n-1}), & I_5^n = -2\eta(\delta_1 u^n, u^{n-2}), & I_6^n = 2\eta\psi_n(\delta_2 u^n, u^{n-2}). \end{cases} \tag{3.4}$$

Using the identities

$$2(\delta_k u^n, u^n) = \delta_k|u^n|^2 + |\delta_k u^n|^2, \quad 2(\delta_k u^n, u^{n-k}) = \delta_k|u^n|^2 - |\delta_k u^n|^2,$$

[1] Note that $\sup\{r^\star(\delta, \eta)\} = 1 + \sqrt{2} = r^\star(1, 0)$, which agrees with the optimal bound for o.d.e's.

we see that

$$I_1^n = \delta_1|u^n|^2 + |\delta_1 u^n|^2, \quad I_2^n = -\psi_n\big(\delta_2|u^n|^2 + |\delta_2 u^n|^2\big),$$
$$I_3^n = -\delta\big(\delta_1|u^n|^2 - |\delta_1 u^n|^2\big), \quad I_6^n = \eta\psi_n\big(\delta_2|u^n|^2 - |\delta_2 u^n|^2\big).$$

Furthermore, since $\delta_2 u^n = \delta_1 u^n + \delta_1 u^{n-1}$, we have

$$\begin{aligned} I_4^n &= 2\delta\psi_n(\delta_1 u^n + \delta_1 u^{n-1}, u^{n-1}) \\ &= \delta\psi_n\big(\delta_2|u^n|^2 - |\delta_1 u^n|^2 + |\delta_1 u^{n-1}|^2\big), \\ I_5^n &= -2\eta(\delta_1 u^n, u^{n-2}) = -2\eta(\delta_2 u^n, u^{n-2}) + 2\eta(\delta_1 u^{n-1}, u^{n-2}) \\ &= -\eta\big(\delta_1|u^n|^2 - |\delta_2 u^n|^2 + |\delta_1 u^{n-1}|^2\big). \end{aligned}$$

Collecting terms, we therefore obtain from (3.3) and (3.4)

$$\begin{aligned} \mathscr{D}_n = J_1^n &+ (1 + \delta - \delta\psi_n)|\delta_1 u^n|^2 + (\delta\psi_n - \eta)|\delta_1 u^{n-1}|^2 \\ &+ (\eta - \eta\psi_n - \psi_n)|\delta_2 u^n|^2 \geq J_1^n + J_2^n \end{aligned} \tag{3.5}$$

with

$$J_1^n = (1 - \delta - \eta)\big(\delta_1|u^n|^2 - \psi_n\delta_2|u^n|^2\big), \quad J_2^n = A_n|\delta_1 u^n|^2 - B_n|\delta_1 u^{n-1}|^2,$$

where

$$A_n := 1 + \delta + 2\eta - (2 + \delta + 2\eta)\psi_n, \quad B_n := -\eta + (2 - \delta + 2\eta)\psi_n;$$

in the derivation of the inequality in (3.5), we used the obvious estimate $|\delta_2 u^n|^2 \leq 2|\delta_1 u^n|^2 + 2|\delta_1 u^{n-1}|^2$.

Now,

$$\begin{aligned} \sum_{j=2}^{n} J_1^j = (1 - \delta - \eta)\Bigg(&(1 - \psi_n)\,|u^n|^2 - \psi_{n-1}|u^{n-1}|^2 - |u^1|^2 - \sum_{j=2}^{n-2}\big(\psi_j - \psi_{j+2}\big)\,|u^j|^2\Bigg) \\ &+ (1 - \delta - \eta)\big(\psi_2|u^0|^2 + \psi_3|u^1|^2\big). \end{aligned}$$

Hence, noting that $\delta + \eta < 1$, we have

$$\sum_{j=2}^{n} J_1^j \geq (1 - \delta - \eta)\Bigg((1 - \psi_n)|u^n|^2 - \psi_{n-1}|u^{n-1}|^2 - |u^1|^2 - \sum_{j=2}^{n-2}[\psi_j - \psi_{j+2}]_+|u^j|^2\Bigg). \tag{3.6}$$

Moreover,

$$\begin{aligned} \sum_{j=2}^{n} J_2^j &= \sum_{j=2}^{n}\big(A_j|\delta_1 u^j|^2 - B_j|\delta_1 u^{j-1}|^2\big) \\ &= \sum_{j=2}^{n-1}(A_j - B_{j+1})|\delta_1 u^j|^2 + A_n|\delta_1 u^n|^2 - B_2|\delta_1 u^1|^2. \end{aligned} \tag{3.7}$$

We shall now show that if $r_j \le r^\star(\delta,\eta)$ for all j, then $A_j - B_{j+1} \ge 0$. Assume that $r_j \le r$ for all j. Since $2+\delta+2\eta > 0$ and $2-\delta+2\eta \ge 0$, A_j and B_j are decreasing and increasing functions of r_j, respectively; thus,

$$\begin{aligned} A_j - B_{j+1} &\ge 1+\delta+2\eta - (2+\delta+2\eta)\left(\frac{r}{1+r}\right)^2 + \eta - (2-\delta+2\eta)\left(\frac{r}{1+r}\right)^2 \\ &= 1+\delta+3\eta - 4\,(1+\eta)\left(\frac{r}{1+r}\right)^2. \end{aligned}$$

In view of (3.1), there holds $A_j - B_{j+1} \ge 0$, and (3.7) yields

$$\sum_{j=2}^{n} J_2^j \ge -B_2 |\delta_1 u^1|^2. \tag{3.8}$$

The asserted estimate (3.2) is an immediate consequence of (3.5), (3.6), and (3.8). □

3.2 Estimation of the terms $\mathscr{A}_n$ accounting for the elliptic operator

Here, we shall estimate the sum of the terms $\mathscr{A}_n$ from below. Lemma 2.1 plays a key role in the proof.

Lemma 3.3 (Estimation of $\mathscr{A}_n$) *Let $\delta = 0.9672$ and $\eta = -0.1793$, and assume that $r_j \le r^\star(0.9672, -0.1793) \approx 1.9398$, $j = 2,\dots,N$; see (3.1). Then,*

$$\frac{1}{2}\sum_{j=2}^{n} \mathscr{A}_j \ge c_1 \sum_{j=2}^{n} k_j \|u^j\|^2 - \delta\omega_2 k_2 \langle u^2, u^1\rangle - \eta\omega_2 k_2 \langle u^2, u^0\rangle - \eta\omega_3 k_3 \langle u^3, u^1\rangle, \tag{3.9}$$

$n = 3,\dots,N$, with $c_1 = 10^{-6}$; for $n = 2$, (3.9) is also valid without the last term on the right-hand side.

Proof We rewrite the sum on the left-hand side of (3.9) in the form

$$\frac{1}{2}\sum_{j=2}^{n} \mathscr{A}_j = \sum_{i,j=1}^{n-1} L_{ij}\langle u^{i+1}, u^{j+1}\rangle - \delta\omega_2 k_2 \langle u^2, u^1\rangle - \eta\omega_2 k_2 \langle u^2, u^0\rangle - \eta\omega_3 k_3 \langle u^3, u^1\rangle, \tag{3.10}$$

with L_{ij} the entries of the matrix $L \in \mathbb{R}^{n-1,n-1}$,

$$L := \begin{pmatrix} \omega_2 k_2 & & & & \\ -\delta\omega_3 k_3 & \omega_3 k_3 & & & \\ -\eta\omega_4 k_4 & -\delta\omega_4 k_4 & \omega_4 k_4 & & \\ & \ddots & \ddots & \ddots & \\ & & -\eta\omega_n k_n & -\delta\omega_n k_n & \omega_n k_n \end{pmatrix}. \tag{3.11}$$

With $\mathbb{L}(r_2, \ldots, r_n)$ the matrix in (2.1) and Λ the diagonal matrix

$$\Lambda := \operatorname{diag}\Big(\frac{1}{k_2}, \frac{1}{k_3}, \ldots, \frac{1}{k_n}\Big),$$

it is easily seen that $\mathbb{L}(r_2, \ldots, r_n) = \Lambda^{1/2} L \Lambda^{1/2}$.

It suffices to show that

$$(\mathbb{L}(r_2, \ldots, r_n)x, x)_2 \geq c_1 \|x\|_2^2 \quad \forall x \in \mathbb{R}^{n-1}. \tag{3.12}$$

Indeed, then, the first term on the right-hand side of (3.10) is larger than or equal to $c_1\big(k_2\|u^2\|^2 + \cdots + k_n\|u^n\|^2\big)$ and thus (3.12) leads to the asserted estimate (3.9).

To see that (3.12) is valid for $n \geq 4$, we note that the quadratic polynomial p in (2.3) with r replaced by $r^\star$, attains its minimum in $[-1, 1]$ at $y^\star = \frac{-\delta\sqrt{r^\star}}{4\eta r^\star}$. For $\delta = 0.9672$ and $\eta = -0.1793$, we have $r \leq r^\star(0.9672, -0.1793) \approx 1.9398$ by inequality (3.1); then, indeed, $0 < y^\star < 1$. Furthermore,

$$p(y^\star) \approx 7.3592 \cdot 10^{-6} > c_1.$$

Notice that (3.12) is valid also for $n = 2$ and $n = 3$. Indeed, for $n = 2$, $(\mathbb{L}(r_2)x, x)_2 = \frac{1}{1+r_2}\|x\|_2^2 \geq \frac{1}{1+r^\star}\|x\|_2^2 \geq c_1\|x\|_2^2$, and for $n = 3$, $(\mathbb{L}(r_2, r_3)x, x)_2 = \frac{1}{1+r_2}x_2^2 - \delta\frac{\sqrt{r_3}}{1+r_3}x_2x_3 + \frac{1}{1+r_3}x_3^2 \geq (\frac{1}{1+r^\star} - \frac{\delta}{2}\frac{\sqrt{r_3}}{1+r_3})\|x\|_2^2 \geq (\frac{1}{1+r^\star} - \frac{\delta}{4})\|x\|_2^2 \geq c_1\|x\|_2^2$.

Now, in view of (3.12), (3.10) and (2.2) lead to the asserted estimate (3.9).

For the motivation of the specific choice of the multipliers δ and η, see Sect. 4. □

3.3 Proof of Theorem 1.1

Here, we use Lemmata 3.2 and 3.3, the discrete Gronwall inequality in Lemma 3.1, and elementary inequalities, to prove Theorem 1.1.

Replacing n by j in (1.7), summing from $j = 2$ to $j = n$, and using (3.2) and (3.9), we obtain

$$\begin{aligned}
&(1-\delta-\eta)\,(1-\psi_n)\,|u^n|^2 + 2c_1\sum_{j=2}^{n} k_j\|u^j\|^2 \\
&\leq (1-\delta-\eta)\,\psi_{n-1}|u^{n-1}|^2 + C\left(|u^0|^2 + |u^1|^2\right) + (1-\delta-\eta)\sum_{j=2}^{n-2}[\psi_j - \psi_{j+2}]_+|u^j|^2 \\
&\quad + \sum_{j=2}^{n}\mathscr{F}_j + 2\delta\omega_2 k_2\langle u^2, u^1\rangle + 2\eta\omega_2 k_2\langle u^2, u^0\rangle + 2\eta\omega_3 k_3\langle u^3, u^1\rangle.
\end{aligned}$$

Now, the terms involving the forcing term or the starting approximations can be estimated by the Cauchy–Schwarz inequality and the elementary inequality

$$2ab \le \varepsilon a^2 + \varepsilon^{-1} b^2, \quad a, b \in \mathbb{R},$$

with $\varepsilon > 0$ small enough. We obtain

$$\mathscr{F}_j \le \omega_j k_j \varepsilon_1^{-1}(1+\delta-\eta)\|f^j\|_\star^2 + \omega_j k_j \varepsilon_1 \left(\|u^j\|^2 + \delta\|u^{j-1}\|^2 - \eta\|u^{j-2}\|^2\right)$$

and

$$2|\langle u^i, u^j\rangle| \le \varepsilon_2\|u^i\|^2 + \varepsilon_2^{-1}\|u^j\|^2, \quad i = 2, 3, \quad j = 0, 1,$$

with sufficiently small ε_1 and ε_2, and we are lead to the inequality

$$\begin{aligned}|u^n|^2 + c_1\sum_{j=2}^{n} k_j\|u^j\|^2 \le \frac{\psi_{n-1}}{1-\psi_n}\left|u^{n-1}\right|^2 + C\big(|u^0|^2 + |u^1|^2 + k_2\|u^0\|^2 + k_2\|u^1\|^2\big) \\ + C\sum_{j=2}^{n-2}[\psi_j - \psi_{j+2}]_+|u^j|^2 + C\sum_{j=2}^{n} k_j\|f^j\|_\star^2, \quad n \ge 2.\end{aligned}$$

Since $\frac{\psi_{n-1}}{1-\psi_n} \le \bar{c} < 1$, and $[\psi_j - \psi_{j+2}]_+ \le C[r_j - r_{j+2}]_+$ (see [12, p. 179]), we have

$$\begin{aligned}|u^n|^2 + c_1\sum_{j=2}^{n} k_j\|u^j\|^2 \le \bar{c}|u^{n-1}|^2 + C\left(|u^0|^2 + |u^1|^2 + k_2\|u^0\|^2 + k_2\|u^1\|^2\right) \\ + C\sum_{j=2}^{n-2}[r_j - r_{j+2}]_+|u^j|^2 + C\sum_{j=2}^{n} k_j\|f^j\|_\star^2, \quad n \ge 2.\end{aligned} \tag{3.13}$$

Hence, we have

$$|u^n|^2 \le \bar{c}|u^{n-1}|^2 + K_n, \quad n \ge 2,$$

where

$$K_n = C\left(|u^0|^2 + |u^1|^2 + k_2\|u^0\|^2 + k_2\|u^1\|^2 + \sum_{j=2}^{n-2}[r_j - r_{j+2}]_+|u^j|^2 + \sum_{j=2}^{n} k_j\|f^j\|_\star^2\right).$$

Let $2 \le n_\star \le n$, be such that $|u^{n_\star}| = \max_{1\le\ell\le n}|u^\ell|$. Setting $n := n_\star$ in the above inequality and using the fact that $K_{n_\star} \le K_n$, we get

$$|u^{n_\star}|^2 \le \bar{c}|u^{n_\star-1}|^2 + K_{n_\star} \le \bar{c}|u^{n_\star}|^2 + K_n,$$

which leads to

$$|u^{n-1}|^2 \le |u^{n_\star}|^2 \le \frac{1}{1-\bar{c}} K_n.$$

Thus, (3.13) yields

$$|u^n|^2 + c_1 \sum_{j=2}^{n} k_j \|u^j\|^2 \le \bar{c}|u^{n-1}|^2 + K_n \le \frac{1}{1-\bar{c}} K_n.$$

Applying here the discrete Gronwall Lemma 3.1, we obtain the asserted stability estimate (1.4). □

Remark 3.1 Proceeding as in the proof of Theorem 1.1, we can see that Emmrich's bound ($r^* \approx 1.9104$) is optimal for a single multiplier as far as the positive definiteness of suitable matrices is concerned, with $\delta = 0.72349$, $\eta = 0$.

4 On the choice of the multipliers δ and η

Here, we comment on the choice $\delta = 0.9672$ and $\eta = -0.1793$ of the multipliers; we also give more precise values of the multipliers and of the bound $r^\star$; see (4.17).

We recall that in our stability analysis we used two conditions on the bound r of the ratios, namely,

$$0 < r \le r^\star(\delta, \eta) = \frac{\sqrt{1+\delta+3\eta}}{2\sqrt{1+\eta} - \sqrt{1+\delta+3\eta}} \tag{4.1}$$

and the positivity condition

$$P(y) = (1+r)p(y) = 1 + \eta r - \delta\sqrt{r}y - 2\eta r y^2 > 0 \quad \forall y \in [-1,1] \tag{P}$$

to estimate the terms accounting for the difference quotient and for the elliptic operator, respectively; see (3.1) and (2.3). Our goal here is to choose the multipliers δ and η in such a way that both conditions, (4.1) and (P), are satisfied for values of r as large as possible.

Let us focus on the condition (P) and introduce the domain

$$D := \big\{(\delta,\eta) : 0 < \delta < 2, -1 < \eta < 0, 1+\delta+3\eta \ge 0\big\} = D_1 \cup D_2 \tag{4.2}$$

with

$$D_1 := \big\{(\delta,\eta) \in D : \frac{3}{16}\delta^2 \le -\eta\big\}, \quad D_2 := \big\{(\delta,\eta) \in D : \frac{3}{16}\delta^2 > -\eta\big\};$$

see Fig. 1. Notice that instead of trying to determine optimal multipliers in the set of admissible multipliers, i.e., multipliers satisfying all conditions needed in our stability

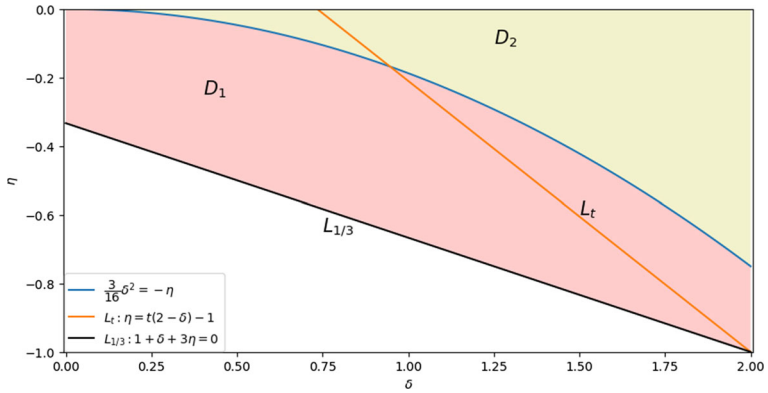

Fig. 1 The domains D (colored region), D_1, and D_2, as well as the segments L_t; see (4.2) and (4.7)

analysis, we find it more convenient to determine optimal multipliers in the larger domain D, in which only some of the conditions of our stability analysis are automatically satisfied, and, a posteriori, check that these multipliers are indeed admissible.

Claim. For $(\delta, \eta) \in D$, the positivity condition (P) is satisfied if and only if

$$r < h(\delta, \eta) := \begin{cases} -\dfrac{1}{\eta} - \dfrac{\delta^2}{8\eta^2}, & (\delta, \eta) \in D_1, \\ \left(\dfrac{2}{\delta + \sqrt{\delta^2 + 4\eta}} \right)^2, & (\delta, \eta) \in D_2. \end{cases} \tag{4.3}$$

To see this, we consider the cases that (δ, η) belongs to D_1 or to D_2 separately. We write P in the form

$$P(y) = -2\eta r \left(y + \frac{\delta}{4\eta\sqrt{r}} \right)^2 + 1 + \eta r + \frac{\delta^2}{8\eta}, \quad y \in [-1, 1]. \tag{4.4}$$

Since the first term on the right-hand side is nonnegative, P is positive in $[-1, 1]$ provided that $1 + \eta r + \frac{\delta^2}{8\eta}$ is positive, i.e.,

$$r < -\frac{1}{\eta} - \frac{\delta^2}{8\eta^2}. \tag{4.5}$$

Notice that (4.5) is also necessary if $-\frac{\delta}{4\eta\sqrt{r}} \le 1$.

For $(\delta, \eta) \in D_1$, in case $-\frac{\delta}{4\eta\sqrt{r}} > 1$, i.e., for $r < \frac{\delta^2}{16\eta^2}$, a seemingly milder condition for the positivity of P in $[-1, 1]$ suffices, namely, $P(1) > 0$. However, we have

$$\frac{\delta^2}{16\eta^2} \le -\frac{1}{\eta} - \frac{\delta^2}{8\eta^2} \iff \frac{3}{16}\delta^2 \le -\eta,$$

which is the motivation for the definition of D_1.

Summarizing, for $(\delta, \eta) \in D_1$, P is positive in $[-1, 1]$ if and only if (4.5) holds; this proves (4.3) for $(\delta, \eta) \in D_1$.

Next, we consider the case $(\delta, \eta) \in D_2$. For $0 < -\dfrac{\delta}{4\eta\sqrt{r}} \le 1$, i.e., for $r \ge \frac{\delta^2}{16\eta^2}$, we have

$$1 + \eta r + \frac{\delta^2}{8\eta} \le 1 + \frac{3\delta^2}{16\eta} < 0 \quad \text{since} \quad \frac{3}{16}\delta^2 > -\eta \quad \text{for} \quad (\delta, \eta) \in D_2,$$

and we easily infer from (4.4) that (P) is not satisfied.

For $-\dfrac{\delta}{4\eta\sqrt{r}} > 1$, i.e., for $0 < \sqrt{r} < -\dfrac{\delta}{4\eta}$, P is positive in $[-1, 1]$ if and only if

$$P(1) = -\eta\left(\sqrt{r} + \frac{\delta}{2\eta}\right)^2 + 1 + \frac{\delta^2}{4\eta} > 0.$$

The discriminant $\delta^2 + 4\eta$ is positive for $(\delta, \eta) \in D_2$, whence $P(1)$ has two real roots $\sqrt{r} = \dfrac{\delta \pm \sqrt{\delta^2 + 4\eta}}{-2\eta}$. In this case, we have

$$0 < \sqrt{r} < \frac{\delta - \sqrt{\delta^2 + 4\eta}}{-2\eta} < -\frac{\delta}{4\eta} \quad \text{since} \quad \frac{3}{16}\delta^2 > -\eta \quad \text{for} \quad (\delta, \eta) \in D_2.$$

Summarizing, for $(\delta, \eta) \in D_2$, P is positive in $[-1, 1]$ if and only if

$$r < \left(\frac{2}{\delta + \sqrt{\delta^2 + 4\eta}}\right)^2, \quad (\delta, \eta) \in D_2;$$

this proves (4.3) for $(\delta, \eta) \in D_2$.

Obviously, the mildest condition on r such that (4.1) and (4.3) are satisfied is

$$r < \max_{(\delta,\eta)\in D} \min\{r^\star(\delta, \eta), h(\delta, \eta)\}. \tag{4.6}$$

It will be convenient to rewrite the expression on the right-hand side of (4.6). Since $\frac{1+\eta}{2-\delta} \in [1/3, \infty)$ for $(\delta, \eta) \in D$, we let

$$t := \frac{1 + \eta}{2 - \delta} \quad \text{with} \quad t \in [1/3, \infty),$$

and, for fixed t, consider the secant segments $L_t \subset D$,

$$L_t : \quad \eta = t(2 - \delta) - 1 \quad \text{for} \quad \eta \in (-1, 0). \tag{4.7}$$

Notice that the secant segment $L_{1/3}$ is $1+\delta+3\eta=0$, and, as t increases from $1/3$ to ∞, the secant segment L_t rotates clockwise and approaches the right boundary of the domain D, whence L_t sweeps the whole domain D; see the colored part in Fig. 1. Consequently, (4.6) can be equivalently written in the form

$$r < \max_{t\in[1/3,\infty)} \max_{(\delta,t(2-\delta)-1)\in L_t} \min\{r^\star(\delta, t(2-\delta)-1), h(\delta, t(2-\delta)-1)\}. \quad (4.8)$$

From (4.7) and (4.1), we get

$$\begin{aligned} H(t) &:= r^\star(\delta, t(2-\delta)-1) = \frac{\sqrt{(3t-1)(2-\delta)}}{2\sqrt{t(2-\delta)}-\sqrt{(3t-1)(2-\delta)}} \\ &= \frac{\sqrt{3t-1}}{2\sqrt{t}-\sqrt{3t-1}}. \end{aligned} \quad (4.9)$$

Analogously, in view of (4.3) and (4.7), we let

$$G(t) := \max_{\{\delta:(\delta,t(2-\delta)-1)\in L_t\}} h(\delta, t(2-\delta)-1) \quad \text{for} \quad t\in[1/3,\infty) \quad (4.10)$$

with

$$\begin{aligned} &h(\delta, t(2-\delta)-1) \\ &= \begin{cases} -\dfrac{1}{t(2-\delta)-1} - \dfrac{\delta^2}{8\,(t(2-\delta)-1)^2}, & (\delta, t(2-\delta)-1)\in D_1, \\ \left(\dfrac{2}{\delta+\sqrt{\delta^2+4\,(t(2-\delta)-1)}}\right)^2, & (\delta, t(2-\delta)-1)\in D_2. \end{cases} \end{aligned} \quad (4.11)$$

According to (4.9) and (4.10), inequality (4.8) can be written as

$$r < \max_{t\in[1/3,\infty)} \min\{H(t), G(t)\}.$$

Next, we consider the maximum of $h(\delta, t(2-\delta)-1)$ in (4.11) for $(\delta, t(2-\delta)-1)\in D$.

For the points $(\delta, t(2-\delta)-1)\in L_t\cap D_2$, according to (4.11), we have

$$\begin{aligned} \frac{\partial h(\delta, t(2-\delta)-1)}{\partial\delta} &= \frac{-8\left(\sqrt{\delta^2+4\,(t(2-\delta)-1)}+\delta-2t\right)}{\left(\delta+\sqrt{\delta^2+4\,(t(2-\delta)-1)}\right)^3\sqrt{\delta^2+4\,(t(2-\delta)-1)}} \\ &= \frac{4\left(\sqrt{\delta^2+4\,(t(2-\delta)-1)}+\delta-2\right)^2}{(2-\delta)\left(\delta+\sqrt{\delta^2+4\,(t(2-\delta)-1)}\right)^3\sqrt{\delta^2+4\,(t(2-\delta)-1)}} \geq 0. \end{aligned}$$

Notice that $t(2-\delta)-1=\eta\in(-1,0)$ and $\delta^2+4\eta$ is positive. Therefore, $h(\delta, t(2-\delta)-1)$ is increasing with respect to δ in the secant line $L_t\cap D_2$, which implies that

the maximum of $h(\delta, t(2-\delta)-1)$ is attained at the point on the curve $\frac{3}{16}\delta^2 = -\eta$. Notice also that this curve lies in D_1. Hence, we only need to consider the points $(\delta, t(2-\delta)-1) \in D_1$.

For points $(\delta, t(2-\delta)-1) \in L_t \cap D_1$, in view of (4.11), we see that

$$\begin{aligned}\frac{\partial h(\delta, t(2-\delta)-1)}{\partial \delta} &= -\frac{1}{4\,(t(2-\delta)-1)^3}\left[\delta\,(t(2-\delta)-1) + \left(\delta^2 + 4\,(t(2-\delta)-1)\right)t\right] \\ &= -\frac{1}{4\,(t(2-\delta)-1)^3}\rho(\delta)\end{aligned}$$

with

$$t(2-\delta)-1 = \eta \in (-1, 0), \quad \rho(\delta) := -\left(4t^2 - 2t + 1\right)\delta + 8t^2 - 4t.$$

Notice that ρ is a decreasing function of δ since it is linear and $4t^2 - 2t + 1 = 4(t - \frac{1}{4})^2 + \frac{3}{4} > 0$.

If $t \in [1/3, 1/2)$, then $\rho(\delta) < \rho(0) = 8t^2 - 4t < 0$. Therefore, $h(\delta, t(2-\delta)-1)$ is decreasing with respect to δ and attains its maximum on the secant segment L_t at $\delta = 0$. From (4.7), we infer that $\eta = 2t - 1$. According to (4.3), we have

$$r < h(\delta, \eta) = -\frac{1}{\eta} = \frac{1}{1-2t}. \tag{4.12}$$

If $t \in (1/2, \infty)$, then $8t^2 - 4t > 0$. The root $\delta^\star$ of ρ is

$$\delta^\star = \frac{4t(2t-1)}{4t^2 - 2t + 1}. \tag{4.13}$$

If $\delta \in (0, \delta^\star)$, then $\rho(\delta) > 0$ and $h(\delta, t(2-\delta)-1)$ is increasing with respect to δ. If $\delta \in (\delta^\star, 2)$, then $\rho(\delta) < 0$ and $h(\delta, t(2-\delta)-1)$ is decreasing with respect to δ. Therefore, $h(\delta, t(2-\delta)-1)$ attains its maximum on the secant segment L_t at $\delta^\star$. From (4.7), we have

$$\eta^\star = t\left(2 - \delta^\star\right) - 1 = -\frac{(2t-1)^2}{4t^2 - 2t + 1}. \tag{4.14}$$

Therefore, from (4.3), we obtain

$$r < h(\delta^\star, \eta^\star) = -\frac{1}{\eta^\star} - \frac{(\delta^\star)^2}{8\,(\eta^\star)^2} = \frac{1}{2} + \frac{1}{2(2t-1)^2}. \tag{4.15}$$

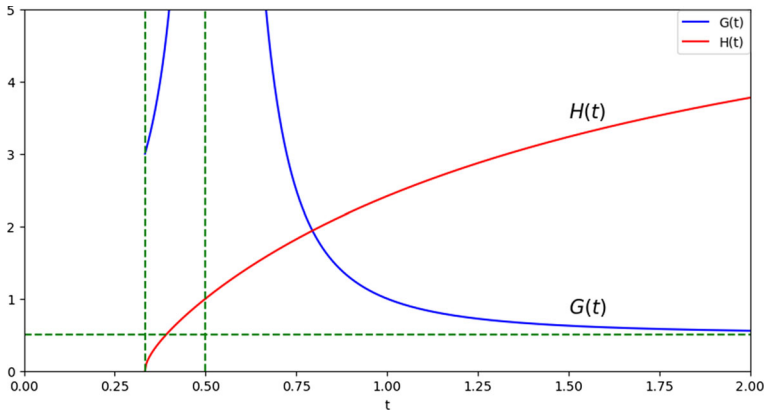

Fig. 2 The graphs of H and G; see (4.9) and (4.16)

Combining (4.10), (4.12) and (4.15), we have

$$G(t) = \begin{cases} \dfrac{1}{1-2t}, & t \in [1/3, 1/2)\,, \\ +\infty, & t = 1/2, \\ \dfrac{1}{2} + \dfrac{1}{2(2t-1)^2}, & t \in (1/2, +\infty). \end{cases} \tag{4.16}$$

It is easily seen from (4.9) that H is increasing with respect to $t \in [1/3, \infty)$. Furthermore, in view of (4.16), G is increasing in the interval $[1/3, 1/2)$ and decreasing in the interval $(1/2, \infty)$; see Fig. 2. Since $H(t) < H(1/2) = 1 < 3 = G(1/3) < G(t)$ for $t \in [1/3, 1/2)$, the graphs of H and G do not intersect for $t \in [1/3, 1/2)$.

On the other hand, there exists a unique optimal point of $H(t) = G(t)$ if $t \in (1/2, \infty)$. Indeed, from (4.9) and (4.16) for $t \in (1/2, \infty)$, we have

$$\frac{\sqrt{3t-1}}{2\sqrt{t} - \sqrt{3t-1}} = \frac{1}{2} + \frac{1}{2(2t-1)^2}, \quad t \in (1/2, \infty),$$

that is

$$23t^5 - 55t^4 + 55t^3 - 29t^2 + 8t - 1 = 0, \quad t \in (1/2, \infty).$$

Notice that the above polynomial has only one real root, namely $t \approx 0.794645365827$. Substituting this value of t in (4.13), (4.14), and (4.15), respectively, we obtain the optimal values

$$\begin{aligned} &\delta^\star \approx 0.967237837020572, \quad \eta^\star \approx -0.179320334471962, \\ &r^\star \approx 1.9398285699451. \end{aligned} \tag{4.17}$$

Let us mention that the multipliers $\delta^\star$ and $\eta^\star$ are admissible, i.e., they satisfy all conditions in our stability analysis; in particular, $2 - \delta^\star + 2\eta^\star \geq 0$, which is used in Lemma 3.2 but does not enter into the definition of the domain D.

Acknowledgements The first-named author is grateful to Prof. Stefano Serra-Capizzano for providing useful information concerning the asymptotic behavior of eigenvalues of symmetric, banded Toeplitz matrices.

Funding Open access funding provided by HEAL-Link Greece.

Declarations

Conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akrivis, G., Chen, M.H., Yu, F., Zhou, Z.: The energy technique for the six-step BDF method. SIAM J. Numer. Anal. **59**, 2449–2472 (2021). https://doi.org/10.1137/21M1392656
2. Becker, J.: A second order backward difference method with variable steps for a parabolic problem. BIT **38**, 644–662 (1998). https://doi.org/10.1007/BF02510406
3. Calvo, M., Grigorieff, R.D.: Time discretisation of parabolic problems with the variable 3-step BDF method. BIT **42**, 689–701 (2002). https://doi.org/10.1023/A:1021992101967
4. Crouzeix, M., Lisbona, F.J.: The convergence of variable-stepsize, variable formula, multistep methods. SIAM J. Numer. Anal. **21**, 512–534 (1984). https://doi.org/10.1137/0721037
5. Emmrich, E.: Stability and error of the variable two-step BDF for semilinear parabolic problems. J. Appl. Math. Comput. **19**, 33–55 (2005). https://doi.org/10.1007/BF02935787
6. Garoni, C., Serra-Capizzano, S.: Generalized Locally Toeplitz Sequences: Theory and Applications, vol. I. Springer, Cham (2017)
7. Grigorieff, R.D.: Stability of multistep-methods on variable grids. Numer. Math. **42**, 359–377 (1983). https://doi.org/10.1007/BF01389580
8. Grigorieff, R.D.: Time discretization of semigroups by the variable two-step BDF method. In: Strehmel, K. (ed.) Numerical Treatment of Differential Equations (NUMDIFF-5, Halle 1989), Teubner, Leipzig, 1991, pp. 204–216
9. Grigorieff, R.D.: On the variable two-step BDF method for parabolic equations, Preprint 426/1995, TU Berlin
10. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems, 2nd edn. Springer, Berlin (1993)
11. Liao, H.-L., Zhang, Z.: Analysis of adaptive BDF2 scheme for diffusion equations. Math. Comput. **90**, 1207–1226 (2021). https://doi.org/10.1090/mcom/3585
12. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, 2nd edn. Springer, Berlin (2006)

Authors and Affiliations

Georgios Akrivis[1,2] · Minghua Chen[3] · Jianxing Han[3] · Fan Yu[3] · Zhimin Zhang[4]

Minghua Chen
chenmh@lzu.edu.cn

Jianxing Han
hanjx2023@lzu.edu.cn

Fan Yu
yuf20@lzu.edu.cn

Zhimin Zhang
ag7761@wayne.edu

1 Department of Computer Science and Engineering, University of Ioannina, Ioannina 451 10, Greece

2 Institute of Applied and Computational Mathematics, FORTH, 700 13 Heraklion, Crete, Greece

3 School of Mathematics and Statistics, Gansu Key Laboratory of Applied Mathematics and Complex Systems, Lanzhou University, Lanzhou 730000, People's Republic of China

4 Department of Mathematics, Wayne State University, Detroit, MI 48202, USA