The energy technique for BDF methods

Georgios Akrivis

joint work with Minghua Chen, Fan Yu and Zhi Zhou

INSTITUTE OF APPLIED & COMPUTATIONAL MATHEMATICS

Dep. Comp. Sci. & Engineering,
University of Ioannina, Greece

Institute Applied & Comp. Math.
Heraklion, Crete, Greece

Irish Numerical Analysis Forum

Energy technique

1. Main characteristic
   We take the inner product with a suitable test quantity (function or element).
2. Possible difficulty
   Suitable choice of test quantity that enables us to treat all terms that enter.
3. In the discrete case (Numerical Analysis)
   - For numerical methods that are stable for all parabolic equations, the choice of the test quantity is dictated by the stability properties of the method and is easy.
     (Algebraically stable Runge–Kutta methods, A-stable multistep methods)
   - For numerical methods that are stable for some parabolic equations the choice of the test quantity is in general difficult (and interesting!).
     (A(ϑ)-stable methods)

Main advantages of the energy technique

1. Simplicity
2. Powerfulness: it leads to several stability estimates
3. Flexibility: it can be easily combined with other stability techniques

Outline

1. An abstract parabolic equation
2. The energy technique for the q-step BDF method
3. $q = 1$ and $q = 2$: Trivial due to the A-stability of the methods (stable for all parabolic equations)
4. $q = 3, 4, 5$: Applicable via Nevanlinna–Odeh multipliers (stable for some parabolic equations)
5. $q = 6$: No Nevanlinna–Odeh multipliers exist
    - Can the Nevanlinna–Odeh requirements be relaxed?

Based on:

A., Minghua Chen, Fan Yu and Zhi Zhou: The energy technique for the six-step BDF method, SIAM J. Numer. Anal. **59** (2021) 2449–2472

1. An abstract parabolic equation

Let $T > 0$ and $u^0 \in H$. Consider the initial value problem

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

with A a positive definite, selfadjoint, linear operator on a Hilbert space $(H, (\cdot,\cdot))$ with domain $\mathscr{D}(A) := \{v \in H : Av \in H\}$ dense on H.

$|\cdot|$ norm on H

$V := \mathscr{D}(A^{1/2})$, $\|\cdot\|$ norm on V, $\|v\| := |A^{1/2}v|$.

Identify H with its dual and denote by V' the dual of V.

$\|\cdot\|_\star$ norm on V', $\|v\|_\star = |A^{-1/2}v|$.

$(\cdot,\cdot)$ inner product on H and antiduality pairing between V' and V.

Then, $\|v\| = (Av, v)^{1/2}$, $\|v\|_\star = (v, A^{-1}v)^{1/2}$, and $|(v,w)| \leqslant \|v\|_\star \|w\|$.

Example

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with smooth boundary $\partial\Omega$.

Then, the negative Dirichlet Laplacian

$$A := -\Delta : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega), \quad \Delta v = \sum_{i=1}^{d} v_{x_i x_i},$$

is a positive definite selfadjoint linear operator.

In this case we have

$$H = L^2(\Omega), \quad V = H_0^1(\Omega), \quad V' = H^{-1}(\Omega).$$

The energy technique for the differential equation

Testing the differential equation $u'(s) + Au(s) = f(s)$ by u, we have

$$\big(u'(s), u(s)\big) + \|u(s)\|^2 = \big(f(s), u(s)\big).$$

Now,

$$\big(u'(s), u(s)\big) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}|u(s)|^2 \quad \text{and} \quad \big(f(s), u(s)\big) \leqslant \frac{1}{2}\big(\|u(s)\|^2 + \|f(s)\|_\star^2\big),$$

whence

$$\frac{\mathrm{d}}{\mathrm{d}s}|u(s)|^2 + \|u(s)\|^2 \leqslant \|f(s)\|_\star^2.$$

Integrating this estimate from 0 to t, we obtain the stability property

$$\boxed{|u(t)|^2 + \int_0^t \|u(s)\|^2\,\mathrm{d}s \leqslant |u^0|^2 + \int_0^t \|f(s)\|_\star^2\,\mathrm{d}s, \quad 0 < t \leqslant T.}$$

Similarly, testing the differential equation $u'(s) + Au(s) = f(s)$ by u', we obtain a second stability estimate

$$\boxed{\|u(t)\|^2 + \int_0^t |u'(s)|^2 \, \mathrm{d}s \leqslant \|u^0\|^2 + \int_0^t |f(s)|^2 \, \mathrm{d}s, \quad 0 < t \leqslant T.}$$

Recall the stability estimates:

$$|u(t)|^2 + \int_0^t \|u(s)\|^2 \, \mathrm{d}s \leqslant |u^0|^2 + \int_0^t \|f(s)\|_\star^2 \, \mathrm{d}s$$

$$\|u(t)\|^2 + \int_0^t |u'(s)|^2 \, \mathrm{d}s \leqslant \|u^0\|^2 + \int_0^t |f(s)|^2 \, \mathrm{d}s$$

For $u^0 = 0$ and $f = 0$, we have $u = 0$. $\rightsquigarrow$ Uniqueness of the solution.

Continuous dependence from both the initial data and the forcing term.

Recall the stability estimates:

$$|u(t)|^2 + \int_0^t \|u(s)\|^2\,\mathrm{d}s \leqslant |u^0|^2 + \int_0^t \|f(s)\|_\star^2\,\mathrm{d}s$$

$$\|u(t)\|^2 + \int_0^t |u'(s)|^2\,\mathrm{d}s \leqslant \|u^0\|^2 + \int_0^t |f(s)|^2\,\mathrm{d}s$$

Goal: Derivation of discrete analogues for BDF methods.

In the discrete case, in the corresponding stability estimates:

- $u(t)$ is replaced by the approximation at a node of a partition.
- The integral is replaced by a sum.
- When we have additional starting approximations, they also enter into the stability estimates.

In the following, to simplify the notation, we consider the homogeneous equation, i.e., with $f=0$. The extension to inhomogeneous equations is very easy.

2. BDF methods

Consider the q-step BDF method, generated by the polynomials α and β,

$$\alpha(\zeta) = \sum_{j=1}^{q} \frac{1}{j}\zeta^{q-j}(\zeta-1)^j = \sum_{j=0}^{q} \alpha_j \zeta^j, \quad \beta(\zeta) = \zeta^q.$$

The BDF methods are $A(\vartheta_q)$-stable with $\vartheta_1 = \vartheta_2 = 90^\circ, \vartheta_3 \approx 86.03^\circ$, $\vartheta_4 \approx 73.35^\circ, \vartheta_5 \approx 51.84^\circ$ and $\vartheta_6 \approx 17.84^\circ$. (Exact values of the angles are also available.) The order of the q-step method is q.

Let $N \in \mathbb{N}$, $\tau := T/N$ be the time step, and $t^n := n\tau, n = 0, \dots, N$, be a uniform partition of the interval $[0, T]$. We recursively define a sequence of approximations u^m to the nodal values $u(t^m)$ by the q-step BDF method,

$$\sum_{i=0}^{q} \alpha_i u^{n+i} + \tau A u^{n+q} = 0 \quad \text{(unknown: } u^{n+q}), \quad n = 0, \dots, N-q,$$

assuming that starting approximations $u^0, \dots, u^{q-1}$ are given.

3. Energy technique

Let $(\mu_1,\dots,\mu_q)\in\mathbb{R}^q$. We test the q-step BDF method by $u^{n+q}-\mu_1u^{n+q-1}-\dots-\mu_qu^n$ and obtain

$$\Big(\sum_{i=0}^{q}\alpha_iu^{n+i},u^{n+q}-\sum_{j=1}^{q}\mu_ju^{n+q-j}\Big)+\tau\Big(Au^{n+q},u^{n+q}-\sum_{j=1}^{q}\mu_ju^{n+q-j}\Big)=0,$$

$n=0,\dots,N-q.$

First requirement: Assume that the polynomials $\alpha(\zeta)=\alpha_q\zeta^q+\dots+\alpha_0$ and $\mu(\zeta):=\zeta^q-\mu_1\zeta^{q-1}-\dots-\mu_q$ have no common divisor. Let $(\cdot,\cdot)$ be a real inner product with associated norm $|\cdot|$. If

$$\operatorname{Re}\frac{\alpha(\zeta)}{\mu(\zeta)}>0\quad\text{for }\ |\zeta|>1, \tag{A}$$

then there exists a positive definite symmetric matrix $G=(g_{ij})\in\mathbb{R}^{q,q}$ such that for $v^0,\dots,v^q$ in the inner product space,

$$\Big(\sum_{i=0}^{q}\alpha_iv^i,v^q-\sum_{j=1}^{q}\mu_jv^{q-j}\Big)\geqslant\sum_{i,j=1}^{q}g_{ij}(v^i,v^j)-\sum_{i,j=1}^{q}g_{ij}(v^{i-1},v^{j-1}). \tag{G}$$

Requirements (A) and (G) are equivalent (G. Dahlquist, 1978). They mean that the q-step scheme described by the parameters $\alpha_q, \dots, \alpha_0$, $1, -\mu_1, \dots, -\mu_q$ and the corresponding one-leg method are A- and G-stable, respectively.

Then, the first term on the left-hand side can be estimated from below using (G). With the notation $\mathcal{U}^n := (u^{n-q+1}, u^{n-q+2}, \dots, u^n)^\top$ and the norm $|\mathcal{U}^n|_G$ given by

$$|\mathcal{U}^n|_G^2 = \sum_{i,j=1}^{q} g_{ij}\big(u^{n-q+i}, u^{n-q+j}\big),$$

using (G), we have

$$\Big(\sum_{i=0}^{q} \alpha_i u^{n+i}, u^{n+q} - \sum_{j=1}^{q} \mu_j u^{n+q-j}\Big) \geqslant |\mathcal{U}^{n+q}|_G^2 - |\mathcal{U}^{n+q-1}|_G^2.$$

Therefore, we have

$$|\mathcal{U}^{n+q}|_G^2 - |\mathcal{U}^{n+q-1}|_G^2 + \tau I_{n+6} \leqslant 0$$

with

$$I_{n+6} := \Big\langle u^{n+q}, u^{n+q} - \sum_{j=1}^{q} \mu_j u^{n+q-j} \Big\rangle.$$

We use the notation $\langle \cdot, \cdot \rangle$ for the inner product on V, $\langle v, w \rangle := (A^{1/2}v, A^{1/2}w)$.

Standard approach: Estimate I_{n+q} from below,

$$I_{n+q} \geqslant \Big(1 - \frac{1}{2}\sum_{i=1}^{q} |\mu_i|\Big)\|u^{n+q}\|^2 - \frac{1}{2}\sum_{i=1}^{q} |\mu_i|\|u^{n+q-i}\|^2.$$

Then, we have

$$|\mathcal{U}^{n+q}|_G^2 - |\mathcal{U}^{n+q-1}|_G^2 + \tau\Big(1 - \frac{1}{2}\sum_{i=1}^{q} |\mu_i|\Big)\|u^{n+q}\|^2 \leqslant \tau\frac{1}{2}\sum_{i=1}^{q} |\mu_i|\|u^{n+q-i}\|^2.$$

Second requirement: To obtain stability with this approach we need

$$1 - |\mu_1| - \cdots - |\mu_q| > 0. \tag{P1}$$

A q-tuple $(\mu_1, \dots, \mu_q)$ satisfying (A) and (P1) is called Nevanlinna–Odeh multiplier for the q-step BDF method.

Nevanlinna and Odeh[1] introduced this technique and determined multipliers of the form $(\mu_1, 0 \ldots, 0)$ for the three-, four- and five-step BDF methods, with

- $\mu_1 = 0.0836$ for the three-step BDF method,
- $\mu_1 = 0.2878$ for the four-step BDF method,
- $\mu_1 = 0.8160$ for the five-step BDF method.

Optimal Nevanlinna–Odeh multipliers, i.e., such that $|\mu_1| + \cdots + |\mu_q|$ is as small as possible are given in[2].

[1] Nevanllina, Odeh: Numer. Funct. Anal. Optim. (1981)

[2] A., Katsoprinakis: Math. Comp. (2016)

4. $q = 6$: Nonexistence of Nevanlinna–Odeh multipliers

$$\text{(A)} \;\Rightarrow\; |\mu_1| + \cdots + |\mu_6| \geqslant \cos\vartheta_6 = 0.9516169$$

For the six-step BDF method, the A-stability condition (A) reads

$$\begin{aligned}
P(x) = {} & (-80x^5 + 208x^4 - 122x^3 - 82x^2 + 98x - 22)\\
& + (40x^4 - 104x^3 + 71x^2 + 15x + 8)\mu_1\\
& + (20x^3 - 52x^2 + 114x - 22)\mu_2 - (8 + 59x - 157x^2)\mu_3\\
& + (294x^3 - 66x^2 - 130x + 22)\mu_4\\
& + (588x^4 - 132x^3 - 417x^2 + 103x + 8)\mu_5\\
& + (1176x^5 - 264x^4 - 1128x^3 + 272x^2 + 146x - 22)\mu_6 \geqslant 0,
\end{aligned}$$

for $x \in [-1, 1]$.

Now,

$$P\left(\frac{3}{40}\right) < -15.156 + 13.735 \sum_{i=1}^{6} |\mu_i|.$$

Assuming $|\mu_1| + \cdots + |\mu_6| \leqslant 1$, we observe that

$$P\left(\frac{3}{40}\right) < -1.421 < 0.$$

Therefore, no Nevanlinna–Odeh multiplier exists.

5. Alternative approach

Idea: Instead of estimating I_{n+q} from below at every time level, sum over n and subsequently estimate the sum from below.

What will we achieve? We will relax the positivity condition

$$1 - |\mu_1| - \cdots - |\mu_q| > 0 \qquad \text{(P1)}$$

of Nevanlinna–Odeh for the q-step BDF method to the milder positivity condition

$$\boxed{1 - \mu_1 \cos x - \cdots - \mu_q \cos(qx) > 0 \quad \forall x \in \mathbb{R}.} \qquad \text{(P2)}$$

Gain? Such multipliers do exist also for the six-step BDF method.

What is the role of (P2)? It ensures that banded symmetric Toeplitz matrices of bandwidth $2q+1$, of any dimension $m \geqslant 2q+1$, with generating function $(1-\varepsilon) - \mu_1 \cos x - \cdots - \mu_q \cos(qx)$ are, for sufficiently small ε, positive definite.

Technical details

Summing from $n=0$ to $n=m-q-1$, we obtain

$$|\mathcal{U}^m|_G^2+\tau\sum_{n=q}^{m} I_n \leqslant |\mathcal{U}^{q-1}|_G^2. \tag{1}$$

It remains to estimate the sum $\sum_{n=q}^{m} I_n$ from below; we have

$$\sum_{n=q}^{m} I_n=\sum_{n=q}^{m}\Big\langle u^n, u^n-\sum_{j=1}^{q}\mu_j u^{n-j}\Big\rangle. \tag{2}$$

With $\mu_0=\varepsilon-1,$ we rewrite (2) as

$$\sum_{n=q}^{m} I_n=\varepsilon\sum_{n=q}^{m}\|u^n\|^2+J_m,\ \ J_m:=-\sum_{j=0}^{q}\mu_j\sum_{i=1}^{m-q+1}\langle u^{q-1+i}, u^{q-1+i-j}\rangle. \tag{3}$$

Rewrite J_m in a suitable form to estimate it from bellow. To this end, we introduce the lower triangular Toeplitz matrix $L = (\ell_{ij}) \in \mathbb{R}^{m-q+1,m-q+1}$ with entries

$$\ell_{i,i-j} = -\mu_j, \quad j = 0, \dots, q, \quad i = j+1, \dots, m-q+1,$$

and all other entries equal zero. With this notation, we have

$$\sum_{i,j=1}^{m-q+1} \ell_{ij} \langle u^{q-1+i}, u^{q-1+j} \rangle = -\sum_{j=0}^{q} \mu_j \sum_{i=j+1}^{m-q+1} \langle u^{q-1+i}, u^{q-1+i-j} \rangle,$$

i.e.,

$$\sum_{i,j=1}^{m-q+1} \ell_{ij} \langle u^{q-1+i}, u^{q-1+j} \rangle = J_m + \sum_{j=1}^{q} \mu_j \sum_{i=1}^{j} \langle u^{q-1+i}, u^{q-1+i-j} \rangle. \qquad (4)$$

The last term can be easily estimated by the Cauchy–Schwarz and arithmetic–geometric inequalities since $q-1+i-j \leqslant q-1$.

We obtain

$$|\mathcal{U}^m|_G^2 + \frac{\varepsilon}{2}\tau\sum_{n=q}^{m}\|u^n\|^2$$

$$+ \boxed{\tau\sum_{i,j=1}^{m-q+1}\ell_{ij}\langle u^{q-1+i}, u^{q-1+j}\rangle} \leqslant |\mathcal{U}^{q-1}|_G^2 + C_\varepsilon\tau\sum_{j=0}^{q-1}\|u^j\|^2.$$

Question: What can we do with the boxed term?

Consider the symmetric part $L_s = (L + L^\top)/2$ of L. The generating function of the banded Toeplitz matrix L_s is

$$\varphi(x) := (1-\varepsilon) - \mu_1\cos x - \cdots - \mu_q\cos(qx).$$

The eigenvalues of L_s are bounded from below by the minimum of φ (Grenander–Szegő theorem).

Now, for $z = (z_0, \dots, z_{m-q})^\top \in \mathbb{C}^{m-q+1}$, we have

$$(L_s z, z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi(x) \Big| \sum_{k=0}^{m-q} z_k \, \mathrm{e}^{\mathrm{i}\, kx} \Big|^2 \mathrm{d}x$$

and

$$(z, z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big| \sum_{k=0}^{m-q} z_k \, \mathrm{e}^{\mathrm{i}\, kx} \Big|^2 \mathrm{d}x.$$

Therefore,

$$(L_s z, z) \geqslant \min_x \varphi(x)(z, z).$$

Thus L_s is positive definite and, consequently, L is also positive definite ($(Lx, x) = (L_s x, x)$ for $x \in \mathbb{R}^{m-q+1}$).

Conclusion: The boxed term is nonnegative and thus the Nevanlinna–Odeh requirement

$$1 - |\mu_1| - \cdots - |\mu_q| > 0$$

can be relaxed to

$$\boxed{1 - \mu_1 \cos x - \cdots - \mu_q \cos(qx) > 0 \quad \forall x \in \mathbb{R}}. \tag{P2}$$

Final stability estimate:

$$c_1 |u^m|^2 + \frac{\varepsilon}{2} \tau \sum_{n=q}^{m} \|u^n\|^2 \leqslant c_2 C_\varepsilon \sum_{j=0}^{q-1} \big(|u^j|^2 + \tau \|u^j\|^2 \big).$$

The constants are independent of A, T, m and τ (but the norm $\| \cdot \|$ depends on A).

Second stability estimate

Similarly, we obtain a discrete analogue of the second stability property of the parabolic equation:

$$\|u^n\|^2 + \tau \sum_{\ell=q}^{n} |\dot{u}^\ell|^2 \leqslant C \sum_{j=0}^{q-1} \|u^j\|^2, \quad n = q, \dots, N,$$

with

$$\dot{v}^{n+q} := \frac{1}{\tau} \sum_{i=0}^{q} \alpha_i v^{n+i}, \quad n = 0, \dots, N-q.$$

The constant is independent of A, T, m and τ.

Table: Multipliers for the six-step BDF method.

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
$\frac{13}{9}$	$-\frac{25}{36}$	$\frac{1}{9}$	0	0	0
1.6	-0.92	0.3	0	0	0
0.8235	-0.855	0.38	0	0	0
1.67	-1	0.4	-0.1	0	0
0.8	-0.7	0.2	0.1	0	0
1.118	-1	0.6	-0.2	0.2	0
0.6708	-0.2	-0.2	0.6	-0.2	0
0.735	-0.2	-0.4	0.8	-0.4	0.2

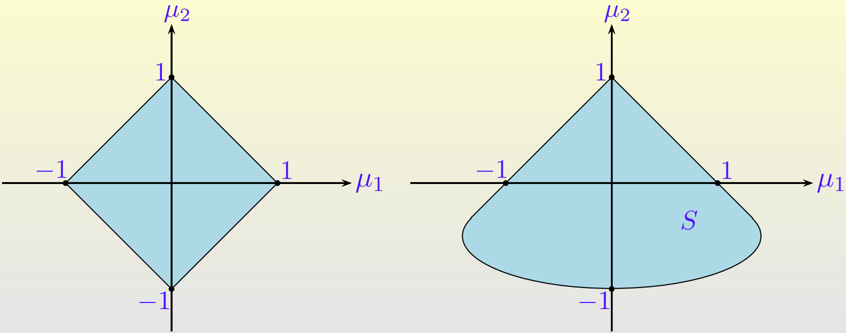

Figure: Illustration of the conditions (P1) and (P2), left and right, respectively, for $\mu_3 = \cdots = \mu_6 = 0$.

$$S = \Big\{(\mu_1, \mu_2) : -\frac{1}{3} \leqslant \mu_2 < 1 - |\mu_1|\Big\} \cup \Big\{(\mu_1, \mu_2) : 4\Big(\mu_2 + \frac{1}{2}\Big)^2 + \frac{1}{2}\mu_1^2 < 1\Big\}$$

Thank you very much!