Optimal Stability for Trapezoidal-Backward Difference Split-Steps

Sohan Dharmaraja, Yinghui Wang,
and Gilbert Strang

Department of Mathematics, MIT

I. Introduction. The trapezoidal method is A-stable. When the equation $u' = au$ has $\mathrm{Re}\ a \le 0$, the difference approximation has $|U_{n+1}| \le |U_n|$:

$$\frac{U_{n+1} - U_n}{\Delta t} = \frac{a\,U_{n+1} + a\,U_n}{2} \quad \text{leads to} \quad U_{n+1} = \frac{1 + a\,\Delta t/2}{1 - a\,\Delta t/2}\,U_n = GU_n. \tag{1}$$

That growth factor G has A-stability:

$$|G| \le 1 \quad \text{whenever} \quad \mathrm{Re}\,(a\,\Delta t) \le 0.$$

The accuracy is second order: $U_n - u(n\,\Delta t)$ is bounded by $C(\Delta t)^2$ for $n\,\Delta t \le T$. But the stability is very close to the edge; $|G| = 1$ if a is imaginary. Nonlinearities can push us over the edge, and the trapezoidal method can fail. In practice (when the iterations to compute U_{n+1} stop early), more stability is often needed.

Extra safety from additional damping can be achieved in different ways. Here we begin by alternating the trapezoidal method with backward differences (BDF2, also second-order accurate):

BDF2 $$\frac{U_{n+2} - U_{n+1}}{\Delta t} + \frac{U_{n+2} - 2U_{n+1} + U_n}{2\,\Delta t} = f(U_{n+2}). \tag{2}$$

This split-step method is self-starting (the trapezoidal method determines U_1 from U_0, and then U_2 comes from BDF2). It is a stabilized option that was proposed in [ ] for circuit simulation. Now it is available in the ADINA finite element code [ ] and elsewhere. An alternative is the Hilber-Hughes-Taylor integrator used successfully by ABAQUS [ ]. It is important to control high-frequency ringing produced by changes in the stepsize Δt.

The computing time in these implicit methods will often be dominated by the solution of a nonlinear system for U_{n+1} and then U_{n+2}. Some variant of Newton's method is the normal choice. So we need an exact or approximate Jacobian of the "implicit parts" in equations (1) and (2), when a nonlinear vector $f(U)$ replaces the scalar test case $f = au$. Writing f' for the matrix $\partial f_i/\partial u_j$, the Jacobians in the two cases are

$$\text{(Trapezoidal)} \quad I - \frac{\Delta t}{2}\,f' \qquad \text{(BDF2)} \quad \frac{3}{2}I - \Delta t\,f'$$

It would be desirable if those Jacobians were equal or proportional. With the same Δt in the two methods, they are not.

The neat idea in [ ] was to allow different steps $\alpha\,\Delta t$ and $(1-\alpha)\,\Delta t$ for trapezoidal and BDF2. The trapezoidal method will now produce $U_{n+\alpha}$ instead of U_{n+1}:

Trapezoidal $$U_{n+\alpha} - U_n = \frac{\alpha\,\Delta t}{2}\big(f(U_{n+\alpha}) + f(U_n)\big). \tag{3}$$

Then BDF2 determines U_{n+1} from U_n and $U_{n+\alpha}$. To maintain second-order accuracy, this requires coefficients A, B, C that depend on α:

BDF2α $$AU_{n+1} - BU_{n+\alpha} + CU_n = (1-\alpha)\Delta t\,f(U_{n+1})\,. \tag{4}$$

Here $A = 2 - \alpha$, $B = 1/\alpha$, and $C = (1-\alpha)^2/\alpha$. The standard choice $\alpha = \frac{1}{2}$ produces $A = \frac{3}{2}$, $B = 2$, $C = \frac{1}{2}$ in agreement with (2). (The step Δt moved to the right side is now $\Delta t/2$.) These values of A, B, C are chosen to give the exact solutions $U = t$ and $U = t^2$ when the right sides are $f = 1$ and $f = 2t$.

The Jacobians in (3) and (4), for $U_{n+\alpha}$ and then U_{n+1}, become

$$\text{(Trapezoidal)} \quad I - \frac{\alpha\,\Delta t}{2} f' \qquad \text{(BDF2}\alpha\text{)} \quad (2-\alpha)I - (1-\alpha)\Delta t\, f' \tag{5}$$

When $\alpha = 2 - \sqrt{2}$ and f' is a constant matrix, J_{BDF} in (4) matches $\sqrt{2}\, J_{\mathsf{Trap}}$ in (3):

$$\sqrt{2}\left[I - \frac{(2-\sqrt{2})\Delta t}{2} f'\right] = \left[\sqrt{2}\, I - (\sqrt{2}-1)\Delta t\, f'\right]. \tag{6}$$

Let $c = 2-\sqrt{2}$ denote this "magic choice" for α. It is known to give the least truncation error [ ] among all α (as well as proportional Jacobians). Our goal in this paper is to identify one more property that makes this choice magic: $\alpha = 2-\sqrt{2}$ *also gives the largest stability region*.

It is recognized that f' would normally be evaluated at U_n in the trapezoidal step and at $U_{n+\alpha}$ in the BDF2 step. In our limited experience, the saving in not computing an extra Jacobian more than compensates for this imperfect start in Newton's method. For linear constant-coefficient dynamics, when f' is the same matrix throughout, Bathe [ ] confirmed the desirability of $\alpha = 2 - \sqrt{2}$. In that case the Jacobian is factored into LU once and for all.

II. The growth factors G_α and G_c.

Stability is tested here, as usual, on the scalar equation $u' = au$. The number a may be complex—it represents any of the eigenvalues in a constant-coefficient linear system.

The trapezoidal method had a growth factor G in equation (1): $U_{n+1} = G(a\,\Delta t)\, U_n$. The split-step combination in (3-4) will have a growth factor G_α, computed now. The particular choice $\alpha = c = 2 - \sqrt{2}$ will then have the growth factor G_c. These factors are ratios of simple polynomials in $z = a\,\Delta t$.

The tests for stability in the model problem $u' = au$ are $|G_\alpha(z)| \le 1$ and $|G_c(z)| \le 1$. We will show that the special choice $|G_c(z)|$ minimizes $|G_\alpha(z)|$.

To compute these growth factors with $f(U) = aU$, substitute $U_{n+\alpha}$ from the trapezoidal step (3) into (4):

$$A\, U_{n+1} - B\, \frac{1+\alpha z/2}{1-\alpha z/2}\, U_n + CU_n = (1-\alpha)\, zU_{n+1}. \tag{7}$$

With $A = 2-\alpha$, $B = 1/\alpha$, and $C = (1-\alpha)^2/\alpha$, this simplifies to $U_{n+1} = G_\alpha\, U_n$:

$$G_\alpha(z) = \frac{2\alpha - 4 - (2 - 2\alpha + \alpha^2)z}{\alpha(\alpha-1)z^2 + (2-\alpha^2)z + 2\alpha - 4}. \tag{8}$$

Lemma 1. Suppose z is real and $Q < \alpha < 1$. Then $|G_\alpha(z)| \le 1$ if and only if $z \le 0$ or $z \ge (4-2\alpha)/(\alpha - \alpha^2)$. This ratio $Q(\alpha)$ gives the edge of the stability region for real z.

Lemma 2. The minimum of $Q(\alpha) = (4-2\alpha)/(\alpha-\alpha^2)$ is at $\alpha = c = 2 - \sqrt{2}$.

Lemma 1 will be proved later, and Lemma 2 now:

$$Q'(\alpha) = \frac{2(2 - 4\alpha + \alpha^2)}{(\alpha - \alpha^2)^2} = 0 \quad \text{at} \quad \alpha = 2 \pm \sqrt{2}. \tag{9}$$

On the interval $0 < \alpha < 1$, $Q(\alpha)$ is large near the endpoints and decreases to its minimum values at $\alpha = c = 2 - \sqrt{2}$. That minimum value Q_c (which gives the largest stability region $z \geq Q_c$ on the real line) is $6 + 4\sqrt{2}$:

$$\min Q_\alpha = Q_c = \frac{4 - 2c}{c - c^2} = 6 + 4\sqrt{2}. \tag{10}$$

Now we turn to complex $z = x + iy$. The growth factor $G_\alpha(z)$ in (8) has numerator N and denominator D. Compute $|N|^2$ and $|D|^2$:

$$|N|^2 = \big[2\alpha - 4 - (2 - 2\alpha + \alpha^2)x\big]^2 + (2 - 2\alpha + \alpha^2)^2 y^2$$
$$|D|^2 = \big[\alpha(\alpha - 1)(x^2 - y^2) + (2 - \alpha^2)x + 2\alpha - 4\big]^2 + \big[2\alpha(\alpha - 1)x + (2 - \alpha^2)\big]^2 y^2$$

Write m for y^2. For fixed x, set $g_\alpha(m) = |D|^2 - |N|^2$. The split-step method is stable when $g_\alpha(m) \geq 0$, which means $|D| \geq |N|$ and $|G_\alpha| \leq 1$.

Since $g_\alpha(m)$ is quadratic in m, we have explicit expressions for its roots $m_1(\alpha)$ and $m_2(\alpha)$. The final step is to maximize the stability region in the imaginary direction y, for fixed real part x. This minimization of $m_1(\alpha)$ and $m_2(\alpha)$ exceeds the patience of humans (the present authors in particular) but not of *Maple*. Computer algebra finds six roots of $m_1'(\alpha) = 0$ and $m_2'(\alpha) = 0$, and two of them are $\alpha = 2 \pm \sqrt{2}$ (checkable by hand).