Optimal Stability for Trapezoidal-Backward Difference Split-Steps

Sohan Dharmaraja, Yinghui Wang,
and Gilbert Strang

Department of Mathematics, MIT

Cambridge MA 02139 USA

Dedicated to Ron Mitchell for a lifetime of leadership by example

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

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

That growth factor G has A-stability:

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

The accuracy is second order: $U_n - u(n\,\Delta t)$ is bounded by $C(\Delta t)^2$ for $n\,\Delta t \leq T$. 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 [1] for circuit simulation. Now it is available in the ADINA finite element code [2–3] and elsewhere. A good alternative is the Hilber-Hughes-Taylor integrator used successfully by ABAQUS [7]. 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 [1] was to allow different steps $\alpha\,\Delta t$ and $(1-\alpha)\,\Delta t$ for trapezoidal and BDF2, with $0 < \alpha < 1$. 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 the part-way value $U_{n+\alpha}$. To maintain second-order accuracy, this requires coefficients A, B, C that depend on α:

BDF2$\boldsymbol{\alpha}$
$$AU_{n+1} - BU_{n+\alpha} + CU_n = (1-\alpha)\Delta t\,f(U_{n+1})\,. \tag{4}$$

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 [1] 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.*

This optimal method is analyzed in [5]: a valuable paper. The recommended start for Newton's method is the current U_n in the trapezoidal step to U_{n+c}, and then V_{n+c} in the BDF2 step:

$$V_{n+c} = (1.5 + \sqrt{2})U_n + (2.5 + 2\sqrt{2})U_{n+c} - (6 + 4.5\sqrt{2})(f_{n+c} - f_n).$$

This comes from a cubic Hermite extrapolation. A key point of [5] is the estimate of local truncation error to be used for the stepsize choice in stiff problems. TR-BDF2c as implemented in MATLAB is described by Hosea and Shampine as an "attractive implicit one-step method".

It is recognized that f' would normally be evaluated at different points in the two half-steps. In our limited experience, the saving in not computing an extra Jacobian at $U_{n+\alpha}$ more than compensates for this imperfect start in Newton's method: Evaluate J_{Trap} at U_n and multiply by $\sqrt{2}$ for J_{BDF}.

For linear constant-coefficient dynamics, when f' is the same matrix throughout, Bathe [2–3] confirmed the desirability of $\alpha = 2 - \sqrt{2}$. In that case the Jacobian is factored into LU once and for all.

With the choice $\alpha = 2-\sqrt{2}$, this split-step method is ode23b among the MATLAB solvers for systems of ordinary differential equations. The documentation mentions that "this solver may be more efficient than ode15s at crude tolerances." This is the correct context for many problems in applied dynamics. The spatial accuracy of finite elements would not justify a high-order method in time. We will use ode45 (Runge-Kutta with stepsize control) as a test standard in numerical simulations of a double pendulum.

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 $|G_c(z)| \le 1$. Those tests are passed for Re $z \le 0$, and they are also passed in parts of the half-plane Re $z > 0$. This is the split-step improvement on the trapezoidal method, shown in the graphs of Figure 1, and $\alpha = c$ gives the greatest improvement.

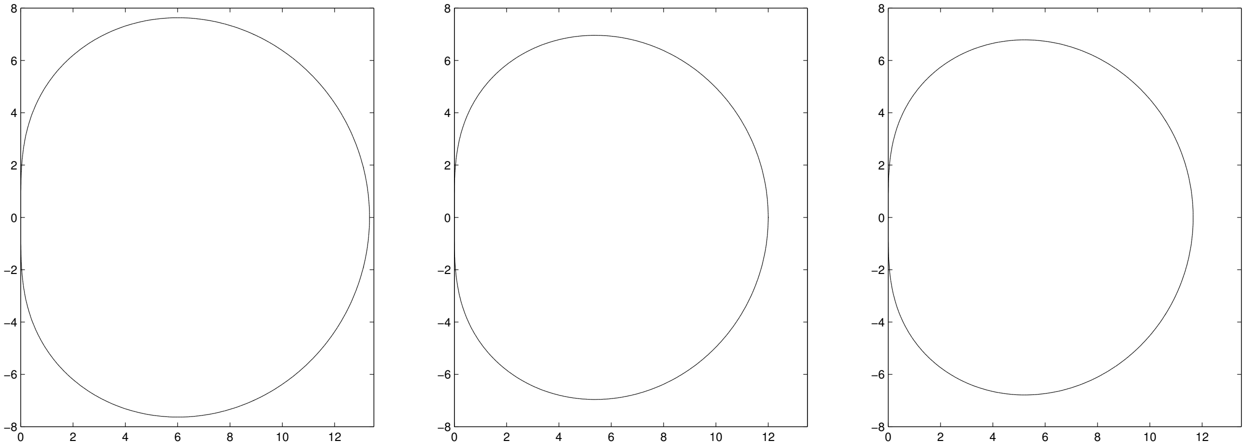
Theorem *The choice $\alpha = c$ gives the largest stability region: If there is an α with $|G_\alpha(z)| \le 1$ then $|G_c(z)| \le 1$.*

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 $U_{n+1} = G_\alpha\,U_n$:

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



$\alpha = .4$ $\alpha = .5$ $\alpha = 2 - \sqrt{2}$

Figure 1: The stability region outside the curve $|G_\alpha(z)| = 1$ is largest for $\alpha = 2 - \sqrt{2}$.

Lemma 1. Suppose z is real and $0 < \alpha < 1$. 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 $z = a\,\Delta t$.

Lemma 2. The minimum of $Q(\alpha)$ is $(4-2\alpha)/(\alpha-\alpha^2)$ at $\alpha = c = 2 - \sqrt{2}$. Thus the unstable real interval $(0, Q)$ is smallest for $\alpha = c$.

Lemma 1 will be proved later, and Lemma 2 now. The key point is the factor $2-4\alpha+\alpha^2$ which vanishes at the choice $\alpha = c$:

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

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

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

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 $|D|^2$:

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

Those are linear and quadratic in the variable $m = y^2$, for fixed x.

Here is an outline of our proof that $\alpha = c$ gives the largest stability region in the z-plane:

1. Set $g_\alpha(m) = |D|^2 - |N|^2$. Then $z = x+iy = a\,\Delta t$ gives stability if $g_\alpha(m) \geq 0$.

2. Fix the real part x between 0 and $6+4\sqrt{2}$. Since g_α is quadratic in $m = y^2$, it has two roots $m_1(\alpha)$ and $m_2(\alpha)$. The expressions for their sum $s(\alpha) = m_1 + m_2$ and their product $p(\alpha) = m_1 m_2$ are straightforward from the quadratic g_α:

$$s(\alpha) = \frac{2}{\alpha^2-\alpha}\left[8 - 4\,\alpha - (2-\alpha^2)x + (\alpha-\alpha^2)x^2\right]$$

$$p(\alpha) = x^4 - \frac{2}{(\alpha^2-\alpha)^2}\left[(2\alpha-4)^2 x + (2-\alpha^2)(\alpha-\alpha^2)x^3\right]$$

3. The derivatives $ds/d\alpha$ and $dp/d\alpha$ contain the factor $2 - 4\,\alpha + \alpha^2$ which vanishes at $\alpha = c = 2-\sqrt{2}$. The derivatives of the individual roots m_1 and m_2 also vanish at $\alpha = c$. Then this choice of α gives the smallest unstable interval $-Y(\alpha) < y < Y(\alpha)$ for each fixed x. Here $m = Y^2$ is the smaller root of $g_\alpha(m) = 0$.

The boundary points $x \pm iY$ of the stability region for $\alpha = c$ form the last (and smallest) curve in Figure 1. The points z inside the curve give instability for $\alpha = c$ and for all $0 < \alpha < 1$. The "magic choice" gives the largest stability region for the split-step method.

Computations for Step 3. The derivatives of $s = m_1 + m_2$ and $p = m_1 m_2$ are

$$\frac{ds}{d\alpha} = \frac{8-2x}{(\alpha^2-\alpha)^2}(2-4\alpha+\alpha^2)$$

$$\frac{dp}{d\alpha} = \frac{16(\alpha-2)x+2(\alpha^2-\alpha)x^2}{(\alpha^2-\alpha)^3}(2-4\alpha+\alpha^2)$$

These vanish at $\alpha = c = 2-\sqrt{2}$ since $2-4c+c^2=0$.

Then $s' = 0$ and $p' = 0$ yield

$$m_1' + m_2' = 0 \quad \text{and} \quad m_2 m_1' + m_1 m_2' = 0.$$

Those two equations imply $m_1' = 0$ and $m_2' = 0$ unless $m_1 = m_2$. This exceptional case also gives $m_1' = m_2' = 0$ by analysis of the discriminant of $g_\alpha(m)$.

May we acknowledge that the formulas for m_1, m_2, m_1', and m_2' were first produced by *Maple*, before it was realized that s and p would allow computations by hand.

Computations for Lemma 1. When $z = x$ is real and $y = 0$, the numerator and denominator of $G_\alpha = N/D$ have

$$\begin{aligned} N+D &= \alpha(\alpha-1)x^2 - 2\ \alpha(\alpha-1)x + 4\alpha - 8 \\ N-D &= \alpha(\alpha-1)x^2 + (4-2\alpha)x \end{aligned}$$

The sum $N+D$ remains negative for all $0 < \alpha < 1$, because its discriminant is $4\alpha(\alpha-1)(\alpha^2-5\alpha+8) < 0$. (That quadratic is positive because its discriminant is -7.) Then $|G_\alpha| = 1$ only when $N = D$. This occurs at $x = 0$ and at $x = (4-2\alpha)/(\alpha-\alpha^2)$. Between those two values we have instability. Lemma 2 says that this real instability interval is smallest when $\alpha = 2-\sqrt{2}$.

III. Examples: The Double Pendulum

We test the improved stability of the split-step TR-BDF2 combination by comparison with the trapezoidal method and BDF method alone. A simple pendulum of length ℓ is governed by the equation $\theta'' = (g/\ell)\sin\theta$. The mild nonlinearity of $\sin\theta$ is compounded when a second pendulum is attached (Figure 2). For simplicity we choose equal masses (which cancel from the equations of motion) and equal lengths ℓ. The tumbling of the second mass over the top of the first gives a convenient model of complex behavior—not easy for a finite difference method to follow.

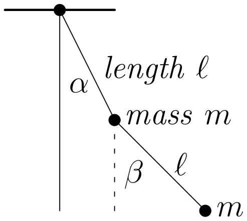


Figure 2: A double pendulum with enough energy can send the lower mass "over the top." The angle β can leave $(-\pi, \pi)$.

We write α and β for the angles from the vertical. The potential energy is lowest when these angles are zero:

$$V = -mg\ell\left(\cos\alpha + (\cos\alpha + \cos\beta)\right).$$

With angular velocities $A = \alpha'$ and $B = \beta'$, the kinetic energy of the double pendulum is

$$K = \frac{1}{2}m\ell^2(A^2 + 2\cos(\alpha - \beta)AB + B^2).$$

Then the Euler-Lagrange equations minimize the action integral of the Lagrangian $L = K - V$ and express Newton's second law:

$$(\partial L/\partial A)' = \partial L/\partial\alpha \quad \text{and} \quad (\partial L/\partial B)' = \partial L/\partial\beta. \tag{11}$$

The unknown is $u(t) = (\alpha, \beta, A, B)$. The system to solve is $u' = f(u) = (A, B, f_3, f_4)$, in which $f_3(u)$ and $f_4(u)$ come from solving equations (11) for A' and B'.

The Jacobian of $f(u)$ is a 4 by 4 matrix. It is evaluated at $u = U_n$ in the first iteration of Newton's method, for the trapezoidal rule $U_{n+\alpha} = U_n + \alpha\,\Delta t(f_n + f_{n+\alpha})/2$.

IV. Numerical Simulations

The double pendulum is evolved using four different time integrators:

1. Trapezoidal method (TR)

2. Second-order backward differences (BDF2)

3. The split-step method (TR-BDF2) with $\alpha = 2 - \sqrt{2}$

4. MATLAB's explicit Runge-Kutta code ode45.

We take ode45, with its higher accuracy, as the standard. The choice $\alpha = 0.5$ for the split-step gives answers very close to the "magic" choice—it is the difference in the Jacobian evaluation and factorization for Newton's method that is significant.

We report here on the motion of the lower pendulum, which is more erratic (physically and also computationally). The graphs will show β and β' against t, and the movement of β' against β in the phase plane (Poincaré plot). Many more graphs are on the website math.mit.edu/cse associated with the textbook [7] on computational science and engineering. (Section 2.6 of the book describes variants of Newton's method and a range of engineering applications.)

The masses will be initially raised to $\alpha(0) = 9\pi/10$ and $\beta(0) = \pi$ with angular velocities $\alpha'(0) = 0.7$ and $\beta'(0) = 0.4$. The time step is 0.02 and the simulation runs to $t = 10$. Near $t = 6.5$ both TR and BDF2 (but not the split-step combination) break away from the accurate solution given by ode45. The first graphs compare the angle $\beta(t)$ and the velocity $\beta'(t)$ of the second pendulum from BDF2 and TR-BDF2.

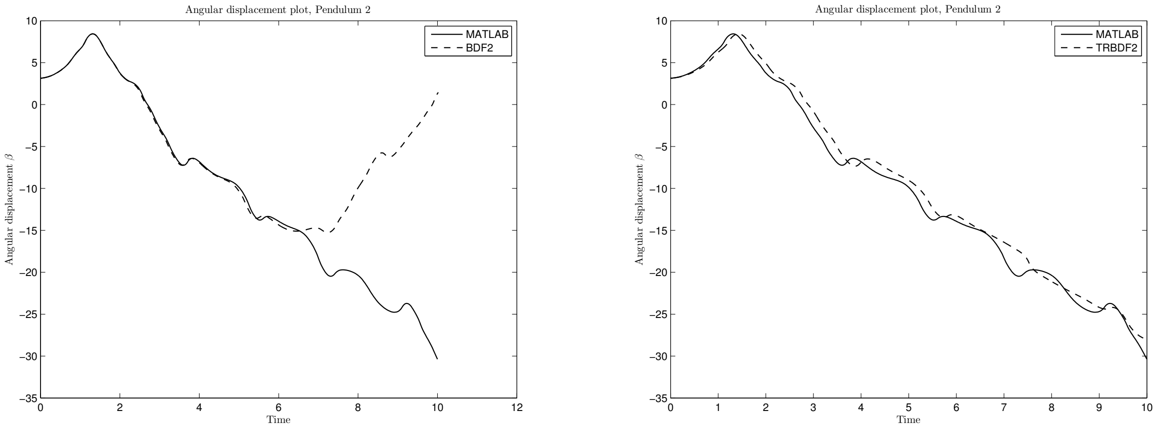


Figure 3: The split-step method stays near the solution from ode45.

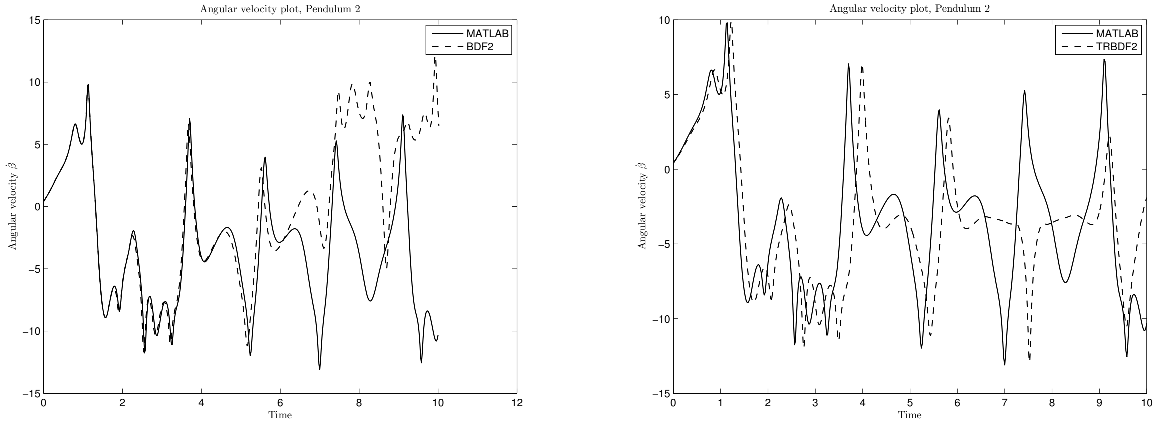


Figure 4: The sign of β' stays positive at $t = 6.5$ for TR-BDF.

Figure 5 shows the same experiments (now for TR versus TR-BDF2 in the $\beta - \beta'$ phase plane. This “Poincaré plot” starts at the right and moves left as time increases.

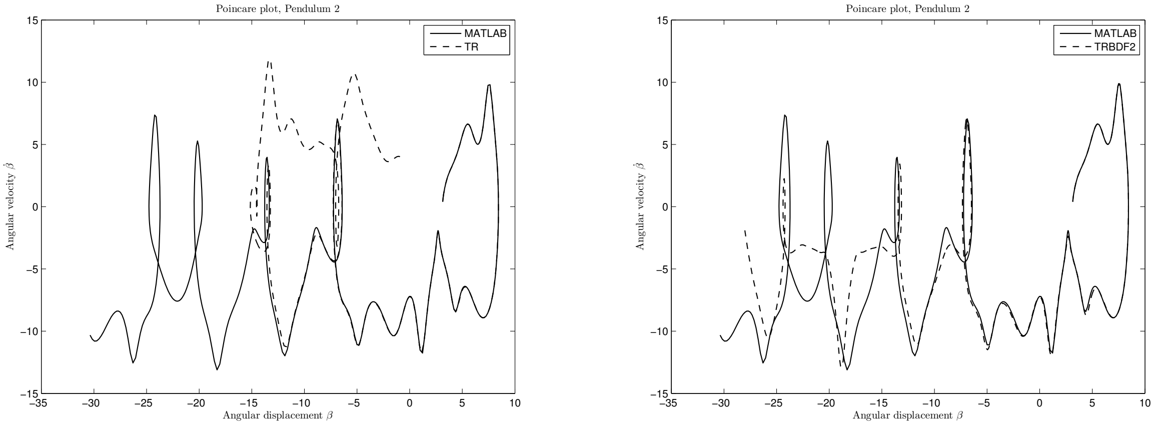


Figure 5: Divergence of the trapezoidal method, shown in the phase plane.

References

[1] Bank, R. E., W. C. Coughran, Jr., W. Fichtner, E. Grosse, D. Rose, and R. Smith, Transient simulation of silicon devices and circuits, *IEEE Trans. CAD* **4** (1985) 436–451 and *IEEE Trans. Electr. Devices* **32** (1985) 1992-2007.

[2] Bathe, K. J. and M. M. I. Baig, On a composite implicit time integration procedure for nonlinear dynamics, *Computers and Structures* **83** (2005) 2513-2524.

[3] Bathe, K. J., Conserving energy and momentum in nonlinear dynamics: A simple implicit time integration scheme, *Computers and Structures* **85** (2007) 437-445.

[4] Kuhl, D. and M. A. Crisfield, Energy-conserving and decaying algorithms in non-linear structural dynamics, *Int. J. Num. Meth. Eng.* **45** (1999) 569-599.

[5] Shampine, L. F. and M. E. Hosea, Analysis and implementation of TR-BDF2, *Applied Numerical Mathematics* **20** (1996) 21–37.

[6] Shampine, L. F., I. Gladwell, and S. Thompson, *Solving ODE's with MATLAB*, Cambridge University Press (2003).

[7] Strang, G., *Computational Science and Engineering*, Wellesley-Cambridge Press (2007).