CAAM 336 · DIFFERENTIAL EQUATIONS

Problem Set 8 · Solutions

Posted Monday 22 October 2012. Due Monday 29 October 2012, 5pm. Corrected 25 October 2012.

1. [50 points: 18 points for (a); 12 points for (b); 10 points each for (c) and (d)]
   Consider the following three matrices:

$$\text{(i)} \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \text{(ii)} \quad \mathbf{A} = \begin{bmatrix} -50 & 49 \\ 49 & -50 \end{bmatrix} \qquad \text{(iii)} \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

   (a) For each of the matrices (i)–(iii), compute the matrix exponential $e^{t\mathbf{A}}$.

   You may use `eig` for the eigenvalues and eigenvectors, but please construct the matrix exponential "by hand" (not with `expm`). For diagonalizable $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, recall the formula $e^{t\mathbf{A}} = \mathbf{V}e^{t\mathbf{\Lambda}}\mathbf{V}^{-1}$. If you encounter a complex eigenvalue $\lambda = \alpha + i\beta$, you may use the formula

$$e^{\lambda} = e^{\alpha+i\beta} = e^{\alpha}(\cos(\beta) + i\sin(\beta)).$$

   (b) Use your answers in part (a) to explain the behavior of solutions $\mathbf{x}(t)$ to the differential equation $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$ as $t \to \infty$, given that $\mathbf{x}(0) = [2,\ 0]^T$ (e.g., specify and explain exponential growth, exponential decay, or neither) for each of the three matrices (i)–(iii).

   (c) For the matrix (ii), describe how large one can choose the time step Δt so that the forward Euler method applied to $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t\mathbf{A}\mathbf{x}_k,$$

   will produce a solution that qualitatively matches the behavior of the true solution (i.e., the approximations $\mathbf{x}_k$ should grow, decay, or remain of the same size as the true solution does). Answer the same question for the backward Euler method

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t\mathbf{A}\mathbf{x}_{k+1}.$$

   (d) For the matrix in (iii), describe how the forward Euler method behaves *for all* Δt as $k \to \infty$ for $\mathbf{x}(0) = [1, 1]^T$. Now describe how the backward Euler method must behave as $k \to \infty$ for the same matrix and initial condition.

Solution.

[GRADERS: it is acceptable for students to use MATLAB to compute eigendecompositions, but they should not simply use the `expm` command. In particular, only give half credit if students computed $e^{t\mathbf{A}}$ for a fixed value of t. The correct answer should depend on the variable t.]

(a) We compute the matrix exponentials for each matrix in turn.

   (i) Note that $\det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 - 1 = (\lambda+1)(\lambda-1)$ and hence the eigenvalues of $\mathbf{A}$ are $\lambda_1 = -1$ and $\lambda_2 = 1$. The corresponding (normalized) orthogonal eigenvectors are

$$\mathbf{u}_1 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \mathbf{u}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

   As $\mathbf{A}$ is symmetric, if we set $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2]$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2)$, we have $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ and

$$e^{t\mathbf{A}} = \mathbf{U}e^{t\mathbf{\Lambda}}\mathbf{U}^* = \mathbf{U}\begin{bmatrix} e^{-t} & 0 \\ 0 & e^{t} \end{bmatrix}\mathbf{U}^*.$$

   Multiplying this out gives

$$e^{t\mathbf{A}} = \tfrac{1}{2}\begin{bmatrix} e^t + e^{-t} & e^t - e^{-t} \\ e^t - e^{-t} & e^t + e^{-t} \end{bmatrix}.$$

(ii) If we denote the matrix in part (a) as $\mathbf{A}_1$, then we find that the $\mathbf{A}$ in part (b) can be written as $\mathbf{A} = -50\mathbf{I} + 49\mathbf{A}_1$, from which it follows that $\mathbf{A}$ has eigenvalues $\lambda_1 = -99$ and $\lambda_2 = -1$ with the same eigenvectors as in part (a):

$$\mathbf{u}_1 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \mathbf{u}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Again $\mathbf{A}$ is symmetric, and we have that

$$e^{tA} = \mathbf{U}e^{t\mathbf{\Lambda}}\mathbf{U}^* = \tfrac{1}{2}\begin{bmatrix} e^{-t} + e^{-99t} & e^{-t} - e^{-99t} \\ e^{-t} - e^{-99t} & e^{-t} + e^{-99t} \end{bmatrix}.$$

(iii) [GRADERS: please be a bit more lenient with this problem, as this $\mathbf{A}$ is nonsymmetric, a case we didn't dwell excessively on in class.]

The characteristic polynomial is $\det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 + 1 = (\lambda - i)(\lambda + i)$, where $i^2 = -1$. Hence the eigenvalues are $\lambda_1 = -i$ and $\lambda_2 = i$. The corresponding normalized eigenvectors are

$$\mathbf{v}_1 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ -i \end{bmatrix}, \qquad \mathbf{v}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ i \end{bmatrix}.$$

Since $\mathbf{A}$ is not symmetric we write $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2]$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2)$, and the matrix exponential takes the form

$$e^{t\mathbf{A}} = \mathbf{V}e^{t\mathbf{\Lambda}}\mathbf{V}^{-1} = \tfrac{1}{2}\begin{bmatrix} e^{it} + e^{-it} & i(e^{-it} - e^{it}) \\ i(e^{it} - e^{-it}) & e^{it} - e^{-it} \end{bmatrix}.$$

Note that for real numbers t,

$$e^{it} = \cos(t) + i\sin(t)$$

and

$$e^{-it} = \cos(-t) + i\sin(-t) = \cos(-t) - i\sin(t),$$

and hence one could arrive at the simple formula (not required):

$$e^{t\mathbf{A}} = \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}.$$

Alternatively, one can note that $\mathbf{A}^2 = -\mathbf{I}$, $\mathbf{A}^3 = -\mathbf{A}$, $\mathbf{A}^4 = \mathbf{I}$, $\ldots$, to obtain from the series expression

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + \tfrac{1}{2}t^2\mathbf{A}^2 + \tfrac{1}{3!}t^3\mathbf{A}^3 + \cdots$$

that

$$e^{t\mathbf{A}} = \begin{bmatrix} 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 - \frac{1}{6!}t^6 + \cdots & t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 + \cdots \\ -t + \frac{1}{3!}t^3 - \frac{1}{5!}t^5 + \frac{1}{7!}t^7 - \cdots & 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 - \frac{1}{6!}t^6 + \cdots \end{bmatrix} = \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}.$$

Here we have spotted that the entries in this matrix are the Taylor series for $\sin(t)$ and $\cos(t)$.

(b) The behavior of $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0)$ for $\mathbf{x}(0) = [2, 0]^T$ depends on the matrix.

(i) For the specified initial condition, we have

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0) = \begin{bmatrix} e^t + e^{-t} \\ e^t - e^{-t} \end{bmatrix}.$$

Thus, as $t \to \infty$, the solution $\mathbf{x}(t)$ blows up. (In fact, it behaves like $e^t[1\ 1]^T$.)

(ii) For the given $\mathbf{x}(0)$, we have

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0) = \begin{bmatrix} e^{-t} + e^{-99t} \\ e^{-t} - e^{-99t} \end{bmatrix},$$

and hence $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$. This must be true since both eigenvalues of $\mathbf{A}$ are negative.

(iii) Notice that

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0) = 2\begin{bmatrix} \cos(t) \\ -\sin(t) \end{bmatrix},$$

so $\mathbf{x}(t)$ neither grows nor decays. (In fact, $\|\mathbf{x}(t)\|$ is constant!)

(c) The eigenvalues for the matrix given by (ii) are $\lambda_1 = -99$ and $\lambda_2 = -1$. Thus the solution to the equation $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$ will decay to zero as $t \to \infty$ for all choices of initial condition $\mathbf{x}(0)$.

We wish to choose the step size Δt for the forward Euler method so that the iterates $\mathbf{x}_k$ decay to zero as $k \to \infty$. For this equation

$$\mathbf{x}_k = (\mathbf{I} + \Delta t\mathbf{A})^k\mathbf{x}_0,$$

and so we need all eigenvalues of the symmetric matrix $\mathbf{I} + \Delta t\mathbf{A}$ to be less than one in magnitude. The eigenvalues of $\mathbf{I} + \Delta t\mathbf{A}$ are simply

$$\mu_1 = 1 + \Delta t\lambda_1 = 1 - 99\Delta_t, \qquad \mu_2 = 1 + \Delta t\lambda_2 = 1 - \Delta_t.$$

For all $0 < \Delta_t < 2$ we have $|\mu_2| < 1$, but to get $|\mu_1| < 1$ we have a stricter requirement:

$$0 < \Delta t < 2/99.$$

(Alternatively, one can simply look for Δt such that $\Delta t\lambda_1, \Delta t\lambda_2 \in (-2, 0)$.)

For the backward Euler method, we have

$$\mathbf{x}_k = (\mathbf{I} - \Delta t\mathbf{A})^{-k}\mathbf{x}_0,$$

and we need all eigenvalues of $(\mathbf{I} - \Delta t\mathbf{A})^{-1}$ to be less than one in magnitude. Those eigenvalues are

$$\mu_1 = \frac{1}{1 - \Delta t\lambda_1} = \frac{1}{1 + 99\Delta t}, \qquad \mu_2 = \frac{1}{1 - \Delta t\lambda_2} = \frac{1}{1 + \Delta t}.$$

These values are less than one in magnitude for all $\Delta t > 0$, so there is no restriction on Δt to obtain $\mathbf{x}_k \to 0$ as $k \to \infty$.

(d) The diagonalizable matrix $\mathbf{A}$ given in (iii) has eigenvalues $\lambda_\pm = \pm i$. It follows that the forward Euler iterations, given by

$$\mathbf{x}_k = (\mathbf{I} + \Delta t\mathbf{A})^k\mathbf{x}(0)$$

will behave as $k \to \infty$ like eigenvalues of $(\mathbf{I} + \Delta t\mathbf{A})^k$, i.e., like $(1 + \Delta t\lambda_\pm)^k$. Since

$$|1 + \Delta t\lambda_\pm| = |1 \pm i\Delta t| = \sqrt{1 + (\Delta t)^2} > 1,$$

we conclude that the forward Euler iterates will always blow up as $k \to \infty$ *for any choice of* $\Delta t > 0$.

On the other hand, the backward Euler iterates,

$$\mathbf{x}_k = (\mathbf{I} - \Delta t\mathbf{A})^{-k}\mathbf{x}(0)$$

will behave as $k \to \infty$ like eigenvalues of $(\mathbf{I} - \Delta t\mathbf{A})^{-k}$, i.e., like $(1 - \Delta t\lambda_\pm)^{-k}$. Since

$$\left|\frac{1}{1 - \Delta t\lambda_\pm}\right| = \frac{1}{|1 \mp i\Delta t|} = \frac{1}{\sqrt{1 + (\Delta t)^2}} < 1,$$

we conclude that the backward Euler iterates will always decay as $k \to \infty$ *for any choice of* $\Delta t > 0$.

---

2. [50 points: 12 points for (a) and (b); 16 points for (c); 10 points for (d)]
There exist a host of alternatives to the forward and backward Euler methods for approximating the solution of the differential equation $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$. For example, the *trapezoid method* has the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tfrac{1}{2}\Delta t\mathbf{A}(\mathbf{x}_k + \mathbf{x}_{k+1}),$$

where $\Delta t > 0$ is the time-step, and $\mathbf{x}_k \approx \mathbf{x}(t_k)$ for $t_k = k\Delta t$.

(a) Like backward Euler, the trapezoid method is an *implicit* technique: $\mathbf{x}_{k+1}$ appears on both the right and left hand side of the formula above that defines it. Describe how to find $\mathbf{x}_{k+1}$ given $\mathbf{x}_k$. In particular, what linear system of algebraic equations needs to be solved at each step? (For comparison, the backward Euler method requires the solution of the system $(\mathbf{I} - \Delta t\mathbf{A})\mathbf{x}_{k+1} = \mathbf{x}_k$ for the unknown $\mathbf{x}_{k+1}$ at each step.)

(b) Consider the matrix and initial condition

$$\mathbf{A} = \begin{bmatrix} -1 & 10 \\ 0 & -2 \end{bmatrix}, \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Approximate the solution to $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$ on the interval $t \in [0, 5]$ for time step $\Delta t = .05$. Produce a `semilogy` plot showing $t_k = k\Delta t$ versus $\|\mathbf{x}_k\|$ for $k = 0, \ldots, 100$. (Use the `norm` command in MATLAB to compute $\|\mathbf{x}_k\|$.)

(c) We wish to understand how the error in our approximation at time $t = 1$ improves as we run the simulation with smaller and smaller Δt values. Produce a `loglog` plot showing Δt versus the error in the trapezoid rule and backward Euler approximations for the matrix and initial condition in part (b) at time $t = 1$. To compute the error, first find the exact solution $\mathbf{x}(1) = e^{\mathbf{A}}\mathbf{x}(0)$ using the `expm` command, then compute the norms $\|\widehat{\mathbf{x}} - \mathbf{x}(1)\|$, where $\widehat{\mathbf{x}}$ denotes your approximation to $\mathbf{x}(1)$ from the trapezoid or backward Euler methods. Start your plot with $\Delta t = 1/2$ and use sufficiently many smaller values of Δt to make the trend in your plot clear. For which method does the error decay more rapidly as $\Delta t \to 0$?

(d) Forward Euler iterates can be written as $\mathbf{x}_k = (\mathbf{I} + \Delta t\mathbf{A})^k\mathbf{x}_0$, while backward Euler iterates can be written as $\mathbf{x}_k = (\mathbf{I} - \Delta t\mathbf{A})^{-k}\mathbf{x}_0$.

Work out a similar formula for the iterates $\mathbf{x}_k$ generated by the trapezoid method.

Suppose $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ is symmetric, and all of its eigenvalues λ_j, $j = 1, \ldots, n$, are negative. How must you choose the time step Δt to ensure that the iterates $\mathbf{x}_k$ generated by the trapezoid method converge to zero, $\|\mathbf{x}_k\| \to 0$, as $k \to \infty$?

Solution.

(a) Given the trapezoid rule

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tfrac{1}{2}\Delta t\mathbf{A}(\mathbf{x}_k + \mathbf{x}_{k+1}),$$

group all terms involving $\mathbf{x}_{k+1}$ on the left and $\mathbf{x}_k$ on the right to obtain

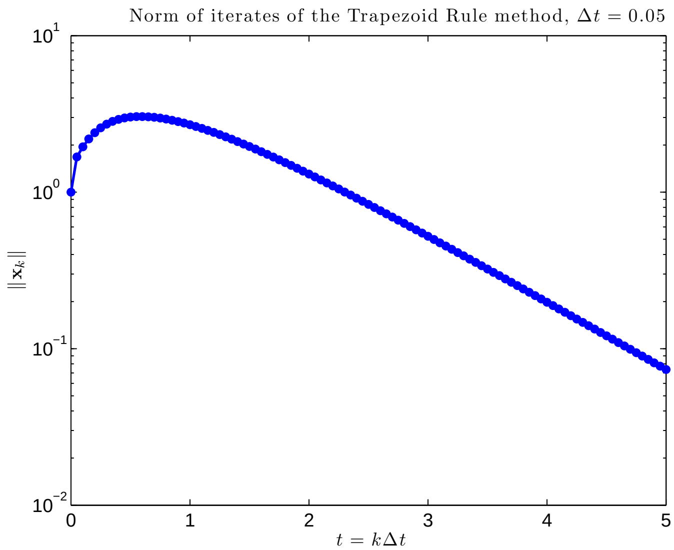
$$(\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})\mathbf{x}_{k+1} = (\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})\mathbf{x}_k.$$

One can solve then solve this system for $\mathbf{x}_{k+1}$ (e.g., using MATLAB's 'backslash' command). Alternatively, invert the matrix on the left to obtain a formula for $\mathbf{x}_{k+1}$:

$$\mathbf{x}_{k+1} = (\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})^{-1}(\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})\mathbf{x}_k.$$

(Either of these forms is acceptable for full credit.)

(b) The plot below shows the growth of norm of the solution $\mathbf{x}_k$ as a function of k for $\Delta t = .05$. Though both eigenvalues of $\mathbf{A}$ are negative, there is transient growth in $\mathbf{x}_k$ before the eventual decay.

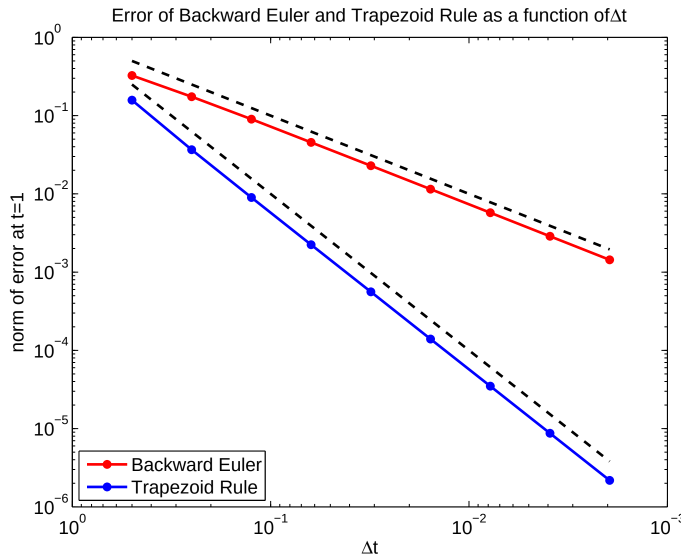



```
A = [-1 10; 0 -2];
x0 = [1;1];

tfinal = 5;
dt = .05;
I = eye(2);
x_trap = x0;
normx = 1;
for k=1:tfinal/dt
    x_trap  = (I-.5*dt*A)\(I+.5*dt*A)*x_trap;  % trapezoid rule
    normx = [normx;norm(x_trap)];
end
figure(1), clf
semilogy(0:dt:tfinal, normx, 'b.-','linewidth',2,'markersize',20)
xlim([0 tfinal])
set(gca,'fontsize',14)
xlabel('$t = k{\Delta}t$','fontsize',14,'interpreter','latex')
ylabel('$\| {\bf x}_k \|$','fontsize',14,'interpreter','latex')
title('Norm of iterates of the Trapezoid Rule method, $\Delta{t} = 0.05$',...
      'fontsize',14,'interpreter','latex')
print -depsc2 traprule1.eps
```

(c) The following plot shows the error in the backward Euler and trapezoid rule computations as a function of the step size Δt. (Note use of the `set('gca','Xdir','reverse')` command to reverse the direction of the horizontal axis so that Δt decreases from left to right.

[GRADERS: Please deduct 7 points if the two lines have the same slope. This error most likely comes from students comparing the approximate solution at time $t = 1 \pm \Delta t$ to the true solution at $t = 1$.]



The black dashed lines show Δt and $(\Delta t)^2$: Note that the backward Euler error decreases at the rate Δt, while the trapezoid rule decreases at the rate $(\Delta t)^2$. Thus, as Δt is cut in half, the trapezoid rule error is quartered. Thus, the trapezoid rule is considerably better.

(Though not necessary for this problem, a complete analysis would also consider the amount of work required at each step. On that count the trapezoid rule is a bit more expensive, as it requires an extra matrix-vector multiplication at each step.)

```
A = [-1 10; 0 -2];
x0 = [1;1];

dtvec = 2.^-[1:9]';
tfinal = 1;
I = eye(2);
err_euler = zeros(size(dtvec));
err_trap  = zeros(size(dtvec));

for j = 1:length(dtvec)
   dt = dtvec(j);
   x_euler = x0; x_trap = x0;
   normx = 1;
   for k=1:tfinal/dt
       x_euler = (I-dt*A)\x_euler;                % backward Euler
       x_trap  = (I-.5*dt*A)\(I+.5*dt*A)*x_trap; % trapezoid rule
   end
   x_true = expm(A*tfinal)*x0;
   err_euler(j) = norm(x_euler-x_true);
   err_trap(j)  = norm(x_trap-x_true);
end
figure(1), clf
loglog(dtvec, err_euler, 'r.-','linewidth',2,'markersize',20), hold on
loglog(dtvec, err_trap, 'b.-','linewidth',2,'markersize',20)
loglog(dtvec, dtvec, 'k--','linewidth',2)
loglog(dtvec, dtvec.^2, 'k--','linewidth',2)
legend('Backward Euler', 'Trapezoid Rule', 3)
set(gca,'fontsize',14,'xdir','reverse')
xlabel('{\Delta}t'), ylabel('norm of error at t=1')
title('Error of Backward Euler and Trapezoid Rule as a function of{ }{\Delta}t')
print -depsc2 traprule2.eps
```

(d) The formula in part (a) enables the computation we need to make for this part. First, by applying k steps of the trapezoid method, we have the formula

$$\mathbf{x}_k = (\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})^{-k}(\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})^k\mathbf{x}_0.$$

Next, the problem asks how the method will behave as $k \to \infty$ if $\mathbf{A}$ is symmetric with all eigenvalues negative. In this case write $BA = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, so that

$$\begin{aligned} \mathbf{x}_k &= (\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})^{-k}(\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})^k \mathbf{x}_k \\ &= \mathbf{V}(\mathbf{I} - \tfrac{1}{2}\Delta t\boldsymbol{\Lambda})^{-k}(\mathbf{I} + \tfrac{1}{2}\Delta t\boldsymbol{\Lambda})^k \mathbf{V}^T \mathbf{x}_k. \end{aligned}$$

The diagonal entries in

$$(\mathbf{I} - \tfrac{1}{2}\Delta t\boldsymbol{\Lambda})^{-k}(\mathbf{I} + \tfrac{1}{2}\Delta t\boldsymbol{\Lambda})^k$$

have the form

$$\frac{(1 + \frac{1}{2}\Delta t\lambda)^k}{(1 - \frac{1}{2}\Delta t\lambda)^k} = \left(\frac{1 + \frac{1}{2}\Delta t\lambda}{1 - \frac{1}{2}\Delta t\lambda}\right)^k,$$

and so the behavior of $\mathbf{x}_k$ as $k \to \infty$ will be controlled by

$$\left|\frac{1 + \frac{1}{2}\Delta t\lambda}{1 - \frac{1}{2}\Delta t\lambda}\right|.$$

If this quantity is less than one for all eigenvalues λ of $\mathbf{A}$, then $\mathbf{x}_k \to \mathbf{0}$ as $k \to \infty$. If any of these quantities is larger than one, there exist initial conditions for which $\|\mathbf{x}_k\| \to \infty$ as $k \to \infty$.

[GRADERS: Please deduct 3 points if the student does not make the following key observation.]

We have not yet used the fact that $\lambda < 0$. What do we learn from this? If $\lambda < 0$, then

$$|1 - \frac{1}{2}\Delta t\lambda| = 1 + \frac{1}{2}\Delta t|\lambda| > |1 + \frac{1}{2}\Delta t\lambda|,$$

and so

$$\left|\frac{1 + \frac{1}{2}\Delta t\lambda}{1 - \frac{1}{2}\Delta t\lambda}\right| < 1.$$

Hence, $\mathbf{x}_k \to \mathbf{0}$ for all initial conditions.