Solutions to selected exercises and additional examples for my book Numerical Methods for Evolutionary Differential Equations

Uri Ascher

July 2, 2009

In this file I have collected solutions to selected exercises appearing in my book. Some of these solutions extend beyond what is strictly required in the question. Others leave details out. These solved exercises serve as additional examples for the text as well. It is seriously recommended that you try to solve the posed problems *before* taking a peak at the answers.

This file represents work in progress, and additions of solved examples (include the source in latex and figures in `eps` or `jpg`, please) will be gratefully received and incorporated. However, the rest of the exercises appearing in the book, even those for which I do have solutions that may be shared with friends and teachers using the text, are intended to remain largely without publicly posted solutions, just like research problems in real life.

Note: References to exercises, sections, equations, figures and tables of the book appear here like there, generally in the form m.n, where m indicates the chapter number, e.g., Figure 1.8. On the other hand, references and labels to equations, tables and figures of this document have only a single counter, e.g., Figure 2.

Chapter 1

1.4 This relates to Section 1.1.

Here $P(\imath\xi) = -\nu\xi^2 - \imath a\xi$. Hence

$$|e^{P(\imath\xi)t}| = |e^{-\nu\xi^2}||e^{-\imath a\xi}| = |e^{-\nu\xi^2}| \le 1 \quad \forall \xi.$$

When $\nu \to 0$ we obtain the advection equation and $|e^{P(\imath\xi)t}| \to 1$.

1.5 This relates to Section 1.2.

(a) As in the text we must consider

$$|\gamma_1(\zeta)| = |1 + \mu a - \mu a[\cos(\zeta) + \imath \sin(\zeta)].$$

Now, for $\zeta = \pi/2$,

$$|\gamma_1(\zeta)| \geq |\mathcal{R}\mathrm{e}\gamma(\zeta)| = |1 + \mu a| > 1.$$

Hence instability is present for any $\mu > 0$. [Additionally, imagine Figure 1.8 with $a > 0$.]

(b) Exactly the same analysis resulting in γ_1 is applied here, yielding

$$\gamma(\zeta) = 1 - \mu a + \mu a e^{-\imath\zeta}.$$

This is the same as γ_1 with $(-a)$ replacing a (the negative sign in the exponent merely implies that we trace the same dotted line in Figure 1.8 in the opposite direction). Hence the same stability results hold with a replaced by $-a$.

(c) The upwind scheme simply picks the stable guy, according to the preceding analysis, between the forward and the backward differences. Hence it is stable for any value of a provided $\mu|a| \leq 1$.

(d) For $u_t - u_x = 0$ we have $a = -1 < 0$. Hence the upwind scheme chooses the forward differencing. The extra column is therefore the same as the 'Error in (1.15a)' in Table 1.1.

1.7 We have

$$\mathbf{u}_t = \begin{pmatrix} 0 & -\partial_{xx} + c \\ \partial_{xx} - c & 0 \end{pmatrix} \mathbf{u}.$$

Hence,

$$P(\imath\xi) = \begin{pmatrix} 0 & \xi^2 + c \\ -(\xi^2 + c) & 0 \end{pmatrix}.$$

The eigenvalus of a general 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ are given by

$$\frac{1}{2}\left[a + d \pm \sqrt{(a+d)^2 - 4(ad - bc)}\right].$$

Here we get the *purely imaginary* eigenvalues $\pm\imath(\xi^2 + c)$. Since they are different the matrix is diagonalizable. So, the situation is precisely as for a hyperbolic PDE system considered in §1.1.2. This IVPDE is well-posed, and it is conservative (i.e., no smoothing, as in Figure 1.3).

Incidentally, there is no need to write the PDE as a real-variable system in order to reach the above conclusions.

Chapter 2

This chapter can really be learned before (or independently of) Chapter 1.

2.2 (a) Obviously,

- $\theta = 0$ gives forward Euler,
- $\theta = 1$ gives backward Euler,
- $\theta = 1/2$ gives the trapezoidal method.

(b) For the test equation $y' = \lambda y$, the θ-method reads

$$y_{n+1} = y_n + z(\theta y_{n+1} + (1-\theta)y_n)$$

where $z = \lambda k$, so

$$y_{n+1} = R(z)y_n = \frac{1+(1-\theta)z}{1-\theta z} y_n \,.$$

The method is therefore A-stable when

$$1/2 \le \theta \le 1 \,.$$

(c) For stiff decay we must have $R(-\infty) = 0$. Clearly

$$R(-\infty) = \frac{\theta - 1}{\theta},$$

so this happens only when $\theta = 1$. For the backward Euler method indeed the method has stiff decay.

(d) The question here is for what values of θ

$$|R(z)| = |\frac{1+(1-\theta)z}{1-\theta z}| \le \delta$$

as $z \to -\infty$? Substituting for $R(-\infty)$ we get

$$\theta \ge \frac{1}{1+\delta}.$$

(e) As a general Runge-Kutta method, note that there are two evaluations of f, one of which is shared with the next step. So let $K_1 = f(y_n)$, $K_2 = f(y_{n+1}) = f(y_n + k[\theta K_2 + (1-\theta)K_1])$. This gives the tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array}$$

(f) Let us check the first few order conditions:

$$\begin{aligned} \mathbf{b}^T C^0 \mathbf{1} &= b_1 + b_2 = 1, \\ \mathbf{b}^T C \mathbf{1} &= b_2 = \theta, \\ \mathbf{b}^T C^2 \mathbf{1} &= b_2 = \theta, \\ \mathbf{b}^T A \mathbf{1} &= \mathbf{b}^T C \mathbf{1} = \theta. \end{aligned}$$

Hence the order is 1 for all θ except $\theta = 1/2$ for which the order is 2.

(g) The advantage is that the method is almost second order and introduces only little damping, which is good for approximating differential problems that have (almost) no damping. On the other hand, $R(-\infty) = \frac{-1+2\varepsilon}{1+2\varepsilon}$, which is barely below 1 in magnitude, so the typical oscillatory behavior of the trapezoidal method is essentially reproduced in the stiff limit.

2.4

$$y_{n+1} = y_{n-1} + 2zy_n, \quad z = k\lambda.$$

As in the leapfrog treatment in Chapter 1 we guess $y_n = \kappa^n$ and obtain the quadratic equation

$$\kappa^2 - 2z\kappa - 1 = 0,$$

with the solutions $\kappa_{1,2} = z \pm \sqrt{z^2+1}$. Both these roots must satisfy $|\kappa_{1,2}| \leq 1$ for stability.

If $z < 0$ is real then obviously $z - \sqrt{z^2+1} < -1$ and stability cannot hold.

For the more general case where λ, hence z, is a complex scalar, note that the two roots of the quadratic equation satisfy

$$\kappa_1 \kappa_2 = -1,$$

so if stability is to hold then $|\kappa_l| = 1,\ l = 1, 2$. For each we can therefore write

$$\kappa_l = e^{\imath\theta_l},$$

so from the quadratic equation

$$z = e^{\imath\theta_l} - e^{-\imath\theta_l} = 2\imath \sin\theta_l.$$

Stability therefore implies that z, hence λ, must be purely imaginary.

2.9 (a) Here $f_y = 2y$, so

$$L = \sup|2y| = \sup_{0 \leq t \leq 1-\varepsilon} |2/(1-t)| = \frac{2}{\varepsilon}.$$

Theorem 2.3 holds, albeit with a large Lipschitz constant L, for any *fixed, positive* ε. There is a unique solution, and the perturbation bounds hold as well although they become less meaningful as $\varepsilon \to 0$.

$k = \varepsilon$	FEuler	RK2	Rk4
.1	5.7	1.9	7.1e-2
.01	7.6e+1	2.1e+1	7.1e-1
.001	8.3e+2	2.1e+2	7.1
.0001	8.8e+3	2.1e+3	7.1e+1
.00001	9.0e+4	2.1e+4	7.1e+2

Table 1: Maximum errors for Exercise 2.9.

(b) The required results are listed in Table 1. There is an added row for $k = \varepsilon = 10^{-5}$ and an added column for the explicit trapezoidal method. When quickly comparing forward Euler against RK4 one might conclude that there is a fundamental difference in the performance of the two methods. But in fact, although the error in RK4 is much smaller it is not fundamentally different, and the error in forward Euler arises not because of an impending blowup but because the results it produces are *below* the exact values. The *relative error* measured over all mesh points, which is more meaningful here, is below 1 for forward Euler, about .21 for the explicit trapezoidal, and about .007 for RK4 for all values of $k = \varepsilon$ listed.

(c) Here what corresponds to λ in the test equation is $f_y = 2y > 1$, so the problem (not the numerical method) is unstable. In general one cannot expect a good pointwise numerical error for an unstable problem. Indeed in chaotic systems, which do contain segments (i.e., subintervals in time) of instability, we may not expect a quality pointwise numerical error after a while unless exponentially small time steps and floating point accuracy are employed.

2.11 (a) We plug the exact x and v into (2.45). By Taylor expansion, assuming as much smoothness on the solution as necessary, we have

$$x(t \pm k) = x(t) \pm kx'(t) + \frac{k^2}{2}x''(t) \pm \frac{k^3}{6}x'''(t) + \frac{k^4}{24}x''''(t) + \dots.$$

Thus,

$$\begin{aligned}[x(t+k) - 2x(t) + x(t-k)]/k^2 &= x'' + \frac{k^2}{12}x''''(t) + O(k^4), \\ &= f(x,v) + \frac{k^2}{12}x''''(t) + O(k^4).\end{aligned}$$

Likewise the approximation for v is 2nd order.

(b) The choice $\gamma = .5$ centers the formula (2.46b) and makes it 2nd order. Then, in (2.46a), for any $0 \le \beta \le 1$ we have $(1-\beta)f(x(t)) + \beta f(x(t+k)) = f(x(t)) + O(k)$, so the formula agrees with Taylor's expansion up to $O(k^3)$ terms, which corresponds to 2nd order accuracy.

(c) With $\gamma = .5$ we have a trapezoidal rule in (2.46b). Now with $\beta = .5$ we have in (2.46a)

$$\begin{aligned} x_{n+1} &= x_n + \frac{k}{2}v_n + \frac{k}{2}(v_n + k(.5f_n + .5f_{n+1})) \\ &= x_n + \frac{k}{2}(v_n + v_{n+1}). \end{aligned}$$

So the whole scheme is just the trapezoidal rule for the first order form (2.44).

(d) In (2.46b) it's the same trapezoidal rule. In (2.46a) we have

$$x_{n+1} = x_n + kv_n + \frac{k^2}{2}f_n.$$

This formula is explicit. The advantage is in case that $f = f(x)$, i.e., when f does not depend on v. Then (2.46b) becomes explicit too, and no nonlinear equations need be solved.

Chapter 3

3.2 (a) The results are tabulated in the first two rows of Table 2. Clearly the

method	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
(3.38a)	2.6e-2	2.5e-3	2.5e-4	2.5e-5	2.5e-6
(3.38b)	8.4e-4	8.3e-6	8.3e-8	8.3e-10	1.5e-11
2nd deriv	1.7e-2	1.7e-3	1.7e-4	1.7e-5	8.3e-8

Table 2: Errors in approximating first and second derivatives of e^x at $x = 0$ on nonuniform meshes $h_j = 10^{-l},\ h_{j-1} = .5h_j$.

first method is first order and the second is second order accurate. The rightmost result for (3.38b) is polluted by floating point cancellation error.

(b) The formula (3.38b) is the result of interpolating at the points x_{j-1}, x_j, x_{j+1} by a quadratic polynomial and differentiating the result at $x = x_j$. Recall the Review section on polynomial interpolation. The derivative of the error according to (2.38) at x_j is therefore what we are after. Differentiating the product and substituting $x = x_j$, the only term that does not vanish is

$$u[x_{j-1}, x_j, x_{j+1}, x](x_j - x_{j-1})(x_j - x_{j+1}) = -\frac{u_{xxx}(\xi)}{6}h_j h_{j-1}.$$

(c) This formula is centered at $.5(x_{j-1} + x_{j+1}) = x_j + h/4$. Therefore the calculated value v satisfies $v = u_x(x_j + h/4) + O(h^2)$. But $u_x(x_j + h/4) =$

$u_x(x_j) + h/4u_{xx}$ where the latter is evaluated at a nearby point. Since $u_{xx} \neq 0$ in the neighborhood, there is a positive constant such that $|u_x(x_j + h/4) - u_x(x_j)| > ch$. Hence

$$|v - u_x(x_j)| \geq |u_x(x_j + h/4) - u_x(x_j)| - |u_x(x_j + h/4) - v| \geq ch + O(h^2).$$

Chapter 4

4.2 The amplification factor is

$$g(\zeta) = \cos(\zeta) - \imath\mu a \sin(\zeta).$$

Thus, assuming $\mu|a| \leq 1$ we have

$$|g(\zeta)|^2 = \cos^2(\zeta) + \mu^2 a^2 \sin^2(\zeta) \leq \cos^2(\zeta) + \sin^2(\zeta) = 1$$

for any ζ.

4.4 This follows straight from the definitions. For a general scheme

$$\sum_{i=-l}^{r} \gamma_i \mathbf{v}_{j+i}^{n+1} = \sum_{i=-l}^{r} \beta_i \mathbf{v}_{j+i}^{n},$$

where β_i and γ_i are $s \times s$ constant matrices, we have the amplification matrix

$$G(\zeta) = (\sum_{i=-l}^{r} \gamma_i e^{\imath i\zeta})^{-1} \sum_{i=-l}^{r} \beta_i e^{\imath i\zeta}.$$

Consistency requires that the local truncation error

$$\tau(t, x) = k^{-1}[\sum_{i=-l}^{r} \gamma_i u(t + k, x + ih) - \sum_{i=-l}^{r} \beta_i u(t, x + ih)]$$

tend to 0 as $k, h \to 0$. Take $u \equiv 1$. Then for consistency we must have

$$\sum_{i=-l}^{r} \gamma_i = \sum_{i=-l}^{r} \beta_i.$$

Therefore,

$$G(0) = I,$$

which yields in particular that $\rho(G(0)) = 1$.

4.8 Stability requires that both roots of the quadratic equation

$$\kappa^2 + 2\imath\mu\left(\rho - \frac{4\nu}{h^2}\sin^2(\zeta/2)\right)\sin(\zeta) + 1 = 0$$

be bounded by 1 in magnitude, where $\mu = k/h$. This happens (indeed their magnitude equals 1) if

$$\mu^2\left(\rho - \frac{4\nu}{h^2}\sin^2(\zeta/2)\right)^2 \sin^2(\zeta) \le 1.$$

The process for getting here is the same as seen before in Chapter 4 and in Exercise 2.3.

Deriving a precise yet simplified condition is hard, but the requirement

$$\mu\left(|\rho| + \frac{4|\nu|}{h^2}\right) \le 1,$$

is clearly a simple sufficient condition for constant-coefficient stability.

Chapter 5

5.4 The obtained errors for the 2nd order schemes are comparable for $\mu < 1$. The resulting solution curves are generally good, and the oscillations observed for the square wave initial data are much diminished (although they are more present in the non-dissipative box scheme). The Lax-Friedrichs scheme produces much less accurate results. For $\mu > 1$ only the box scheme produces reasonable results, of course.

However, the observed convergence rate is well below what may be expected from 2nd order methods. This is because of the low smoothness of $u_0(x)$ at $x = 0, 1, -1$. These cusps propagate also for $t > 0$ at the phase speed $c = -1$.

5.13 Below is a table of computed errors. The solution is depicted in Figure 1.

h	k	$\|error\|_2$	$\|error\|_\infty$
$2^{-7}\pi$	h	.24	.075
	2h	.55	.18
	3h	*	*
$2^{-8}\pi$	h	.0090	.0027
	2h	.092	.025
	3h	*	*

The errors for the finer spatial step are significantly better when $k = h$, more than just by the expected factor $2^4 = 16$. For $k = 3h$ the method is unstable, as

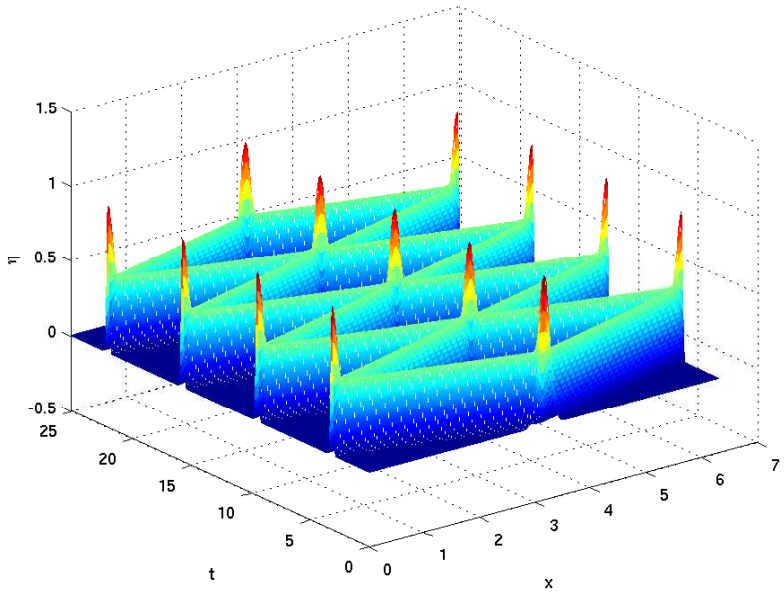


Figure 1: Exercise 5.13: solution using $h = k = 2^{-8}\pi$.

expected for RK4. The solution for the coarser $h = k$ is depicted in Figure 2(a). Compare this to the solution for the finer $h = k$ in Figure 2(b).

This suggests that the coarser h is simply not sufficient to resolve this sharp solution profile, whereas the finer one is.

5.14 (a) Applying a Fourier transform as usual we get

$$\begin{aligned} \hat{u}_t &= \imath\xi\hat{w}, \\ (1+\nu\xi^2)\hat{w}_t &= \imath\xi\hat{u}. \end{aligned}$$

The symbol matrix is therefore

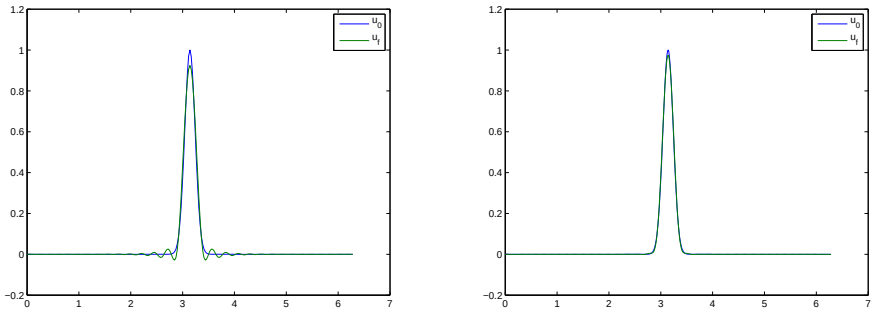
$$P(\imath\xi) = \begin{pmatrix} 0 & \imath\xi \\ \frac{\imath\xi}{1+\nu\xi^2} & 0 \end{pmatrix}.$$

The eigenvalues are

$$\pm\frac{\imath\xi}{\sqrt{1+\nu\xi^2}}.$$

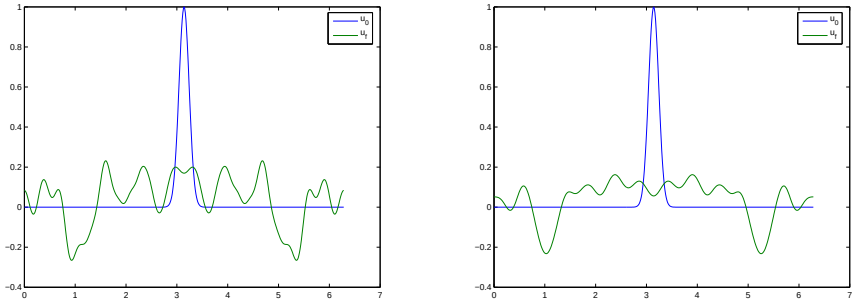
Since they are both imaginary the claims to be proved are obtained.

(b) Here we don't know an exact error. The method is l_2-stable even for $k = 10h$, because the high wave numbers are attenuated, as we see in



(a) using $h = k = 2^{-7}\pi$ (b) using $h = k = 2^{-8}\pi$

Figure 2: Exercise 5.13: solutions at $t = 0$ and $t = 8\pi$.



(a) using $h = k = 2^{-8}\pi$ (b) using $h = 2^{-8}\pi$ and $k = 10h$

Figure 3: Exercise 5.14: solutions at $t = 0$ and $t = 8\pi$.

Part (a). But the solution is not physical. The solution for the finer $h = k$ is depicted in Figure 3(a). The solutions for $h = 2^{-8}\pi$, $k = 2h$, and for $h = k = 2^{-7}\pi$ do not look very different. But the solution for $h = 2^{-8}\pi$, $k = 10h$, depicted in Figure 3(b), does look different. The error dangerously masquerades as an honest wave.

```
% Exercise 5.14
% Shallow water regime. If b = 0 then becomes the
% classical wave equation written as a 1st order system
%       v_t = u_x,     w_t = cv_x,     (c = r2/r1 - 1)
%       w = u - b u_{xx}               (b = r2/r1 sqrt(beta) )
%       periodic conditions on [0,2pi]
%       v(0,x) = exp(-a(x-pi)^2), a=50, u(0,x) = 0.

% problem coefficients
clear all
```

```
r1 = 1; r2 = 2; c = r2/r1 - 1;
b = 0.01;
a = 50;

tf = 8*pi;                % final time
J = 2^9;
h = 2*pi/J;           % spatial step size
k = 2*h/sqrt(c);        % time step size
mu = k/h;
N = floor(tf / k);
xx  = [0:h:2*pi];     % spatial mesh: identify xx(1) with xx(end)
tt = 0:k:N*k;

% construct the almost tridiagonal matrix: note it's only J x J
r = b/h^2;
s = 1+2*r;

e = ones(J,1);
M = spdiags([-r*e s*e -r*e], -1:1, J, J);
M(1,J) = -r;
M(J,1) = -r;

% initial conditions
v(1,:) = exp(-a*(xx-pi).^2);
u(1,:) = zeros(1,length(xx));
w(1,:) = zeros(1,length(xx));            % w is the auxuliary variable

% RK4
for n=1:N
  Yv1 = v(n,:);
  Yu1 = u(n,:);
  Yw1 = w(n,:);
  Kv1(3:J-1) = (8*(Yu1(4:J)-Yu1(2:J-2)) -Yu1(5:J+1)+Yu1(1:J-3) ) / (12*h);
  Kv1(2) = (8*(Yu1(3)-Yu1(1)) -Yu1(4)+Yu1(J) ) / (12*h);
  Kv1(1) = (8*(Yu1(2)-Yu1(J)) -Yu1(3)+Yu1(J-1) ) / (12*h);
  Kv1(J) = (8*(Yu1(J+1)-Yu1(J-1)) -Yu1(2)+Yu1(J-2) ) / (12*h);
  Kv1(J+1) = Kv1(1);
  Yv2 = Yv1 + k/2 * Kv1;
  Kw1(3:J-1) = c * (8*(Yv1(4:J)-Yv1(2:J-2)) -Yv1(5:J+1)+Yv1(1:J-3) ) / (12*h)
  Kw1(2) = c * (8*(Yv1(3)-Yv1(1)) -Yv1(4)+Yv1(J) ) / (12*h);
  Kw1(1) = c * (8*(Yv1(2)-Yv1(J)) -Yv1(3)+Yv1(J-1) ) / (12*h);
  Kw1(J) = c * (8*(Yv1(J+1)-Yv1(J-1)) -Yv1(2)+Yv1(J-2) ) / (12*h);
```

```
Kw1(J+1) = Kw1(1);
Yw2 = Yw1 + k/2 * Kw1;
Yu2(1:J) = (M \ Yw2(1:J)')';  Yu2(J+1) = Yu2(1);

Kv2(3:J-1) = (8*(Yu2(4:J)-Yu2(2:J-2)) -Yu2(5:J+1)+Yu2(1:J-3) ) / (12*h);
Kv2(2) = (8*(Yu2(3)-Yu2(1)) -Yu2(4)+Yu2(J) ) / (12*h);
Kv2(1) = (8*(Yu2(2)-Yu2(J)) -Yu2(3)+Yu2(J-1) ) / (12*h);
Kv2(J) = (8*(Yu2(J+1)-Yu2(J-1)) -Yu2(2)+Yu2(J-2) ) / (12*h);
Kv2(J+1) = Kv2(1);
Yv3 = Yv1 + k/2 * Kv2;
Kw2(3:J-1) = c * (8*(Yv2(4:J)-Yv2(2:J-2)) -Yv2(5:J+1)+Yv2(1:J-3) ) / (12*h)
Kw2(2) = c * (8*(Yv2(3)-Yv2(1)) -Yv2(4)+Yv2(J) ) / (12*h);
Kw2(1) = c * (8*(Yv2(2)-Yv2(J)) -Yv2(3)+Yv2(J-1) ) / (12*h);
Kw2(J) = c * (8*(Yv2(J+1)-Yv2(J-1)) -Yv2(2)+Yv2(J-2) ) / (12*h);
Kw2(J+1) = Kw2(1);
Yw3 = Yw1 + k/2 * Kw2;
Yu3(1:J) = (M \ Yw3(1:J)')';  Yu3(J+1) = Yu3(1);

Kv3(3:J-1) = (8*(Yu3(4:J)-Yu3(2:J-2)) -Yu3(5:J+1)+Yu3(1:J-3) ) / (12*h);
Kv3(2) = (8*(Yu3(3)-Yu3(1)) -Yu3(4)+Yu3(J) ) / (12*h);
Kv3(1) = (8*(Yu3(2)-Yu3(J)) -Yu3(3)+Yu3(J-1) ) / (12*h);
Kv3(J) = (8*(Yu3(J+1)-Yu3(J-1)) -Yu3(2)+Yu3(J-2) ) / (12*h);
Kv3(J+1) = Kv3(1);
Yv4 = Yv1 + k * Kv3;
Kw3(3:J-1) = c * (8*(Yv3(4:J)-Yv3(2:J-2)) -Yv3(5:J+1)+Yv3(1:J-3) ) / (12*h)
Kw3(2) = c * (8*(Yv3(3)-Yv3(1)) -Yv3(4)+Yv3(J) ) / (12*h);
Kw3(1) = c * (8*(Yv3(2)-Yv3(J)) -Yv3(3)+Yv3(J-1) ) / (12*h);
Kw3(J) = c * (8*(Yv3(J+1)-Yv3(J-1)) -Yv3(2)+Yv3(J-2) ) / (12*h);
Kw3(J+1) = Kw3(1);
Yw4 = Yw1 + k * Kw3;
Yu4(1:J) = (M \ Yw4(1:J)')';  Yu4(J+1) = Yu4(1);

Kv4(3:J-1) = (8*(Yu4(4:J)-Yu4(2:J-2)) -Yu4(5:J+1)+Yu4(1:J-3) ) / (12*h);
Kv4(2) = (8*(Yu4(3)-Yu4(1)) -Yu4(4)+Yu4(J) ) / (12*h);
Kv4(1) = (8*(Yu4(2)-Yu4(J)) -Yu4(3)+Yu4(J-1) ) / (12*h);
Kv4(J) = (8*(Yu4(J+1)-Yu4(J-1)) -Yu4(2)+Yu4(J-2) ) / (12*h);
Kv4(J+1) = Kv4(1);
v(n+1,:) = Yv1 + k/6 * (Kv1+2*Kv2+2*Kv3+Kv4);
Kw4(3:J-1) = c * (8*(Yv4(4:J)-Yv4(2:J-2)) -Yv4(5:J+1)+Yv4(1:J-3) ) / (12*h)
Kw4(2) = c * (8*(Yv4(3)-Yv4(1)) -Yv4(4)+Yv4(J) ) / (12*h);
Kw4(1) = c * (8*(Yv4(2)-Yv4(J)) -Yv4(3)+Yv4(J-1) ) / (12*h);
Kw4(J) = c * (8*(Yv4(J+1)-Yv4(J-1)) -Yv4(2)+Yv4(J-2) ) / (12*h);
```

```
  Kw4(J+1) = Kw4(1);
  w(n+1,:) = Yw1 + k/6 * (Kw1+2*Kw2+2*Kw3+Kw4);
  u(n+1,1:J) = (M \ w(n+1,1:J)')'; u(n+1,J+1) = u(n+1,1);

end

figure(1)
mesh(xx,tt,v)
xlabel('x')
ylabel('t')
zlabel('\eta')
axis([0 7 0 tf -.5 1.5])

figure(2)
plot(xx,v(1,:),xx,v(N+1,:))
legend('u_0','u_f')
```

Chapter 6

6.4 The Hamiltonian system can generally be written as

$$\mathbf{y}' = J\boldsymbol{\nabla} H(\mathbf{y}),$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

The important properties of the matrix J are that it is *skew-symmetric*, i.e., $J^T = -J$, and that it is nonsingular and constant in the independent variable t. Denote the Jacobian matrix by

$$Y(t;\mathbf{c}) = \frac{\partial \mathbf{y}(t;\mathbf{c})}{\partial \mathbf{c}},$$

with $\mathbf{y}(0;\mathbf{c}) = \mathbf{c}$ for some (arbitrary) initial time. The flow is called *symplectic* if

$$Y^T J^{-1} Y = J^{-1}, \quad \forall t.$$

Differentiating the ODE for $\mathbf{y}$ with respect to $\mathbf{c}$ we have

$$Y' = J(\nabla^2 H)Y, \quad Y(0) = I.$$

Now, the midpoint method reads at time step n

$$\frac{\mathbf{y}_{n+1} - \mathbf{y}_n}{k} = J\boldsymbol{\nabla} H(\mathbf{y}_{n+1/2}),$$

where $\mathbf{y}_{n+1/2} = (\mathbf{y}_n + \mathbf{y}_{n+1})/2$ and $\mathbf{y}_0$ is given by the initial data. Then also

$$\begin{aligned} \frac{Y_{n+1} - Y_n}{k} &= J\nabla^2 H(\mathbf{y}_{n+1/2}) Y_{n+1/2}, \\ Y_{n+1/2} &= \frac{Y_{n+1} + Y_n}{2}, \quad n = 0, 1, \ldots, \end{aligned} \tag{1}$$

and $Y_0 = I$ the identity.

Consider

$$Z = (Y_{n+1} - Y_n)^T J^{-1} (Y_{n+1} + Y_n) + (Y_{n+1} + Y_n)^T J^{-1} (Y_{n+1} - Y_n).$$

Opening up parentheses, the cross terms cancel and we have

$$Z = 2(Y_{n+1}^T J^{-1} Y_{n+1} - Y_n^T J^{-1} Y_n).$$

On the other hand, by (1)

$$\begin{aligned} \frac{Y_{n+1} + Y_n}{2} J^{-1} \frac{Y_{n+1} - Y_n}{k} &= Y_{n+1/2}^T \nabla^2 H(\mathbf{y}_{n+1/2}) Y_{n+1/2}, \\ \frac{(Y_{n+1} - Y_n)^T}{k} J^{-1} \frac{Y_{n+1} + Y_n}{2} &= Y_{n+1/2}^T \nabla^2 H(\mathbf{y}_{n+1/2}) J^T J^{-1} Y_{n+1/2} \\ &= -Y_{n+1/2}^T \nabla^2 H(\mathbf{y}_{n+1/2}) Y_{n+1/2}. \end{aligned}$$

Therefore

$$Z = 0.$$

Hence

$$Y_{n+1}^T J^{-1} Y_{n+1} = Y_n^T J^{-1} Y_n = \ldots = Y_0^T J^{-1} Y_0 = J^{-1}.$$

6.5 (a) Half a step of forward Euler gives

$$y_{n+1/2} = y_n + k/2 \; f(y_n).$$

Following this by half a step of backward Euler gives

$$y_{n+1} = y_{n+1/2} + k/2 \; f(y_{n+1}).$$

Adding these two equations up clearly yields the trapezoidal step.

(b) Half a step of backward Euler first gives

$$y_{n+1/2} = y_n + k/2\ f(y_{n+1/2}).$$

Then forward Euler gives

$$y_{n+1} = y_{n+1/2} + k/2\ f(y_{n+1/2}).$$

Adding these equations up gives

$$y_{n+1} = y_n + k\ f(y_{n+1/2}).$$

Subtracting these equations from one another confirms that indeed

$$y_{n+1/2} = (y_n + y_{n+1})/2$$

so the midpoint step is obtained.

(c) Since each trapezoidal step is half a forward-Euler step (denote f) followed by half a backward-Euler step (denote b), N trapezoidal steps give the sequence $(fb)(fb)(fb)(fb)(fb)(fb)\ldots(fb)$. Writing this sequence as f followed by the rest $(bf)(bf)(bf)(bf)(b\ldots(bf)$ and one more b, we obtain the required result.

Another way: Writing the trapezoidal scheme

$$y_{n+1} = y_n + k/2(f_n + f_{n+1}),\ \ n = 0, 1, 2, \ldots N-1$$

and summing these expressions up from $n = 0$ to $n = N-1$ yields

$$y_N - y_0 = k/2\ f_0 + \ k\sum_{n=1}^{N-1} f_n\ + k/2\ f_N.$$

On the other hand, doing half a forward-Euler step, $y_{1/2} = y_0 + k/2 f_0$, followed by $N-1$ midpoint steps from one midstep to the next

$$y_{n+1/2} = y_{n-1/2} + kf_n,\ \ n = 1, \ldots, N-1,$$

and then a half backward-Euler step $y_N = y_{N-1/2} + k/2 f_N$ and summing all these up gives the same expression for $y_N - y_0$.

(d) An example can be derived by considering as simple an ODE as $y' = \lambda(t)y$. Consider the scalar problem

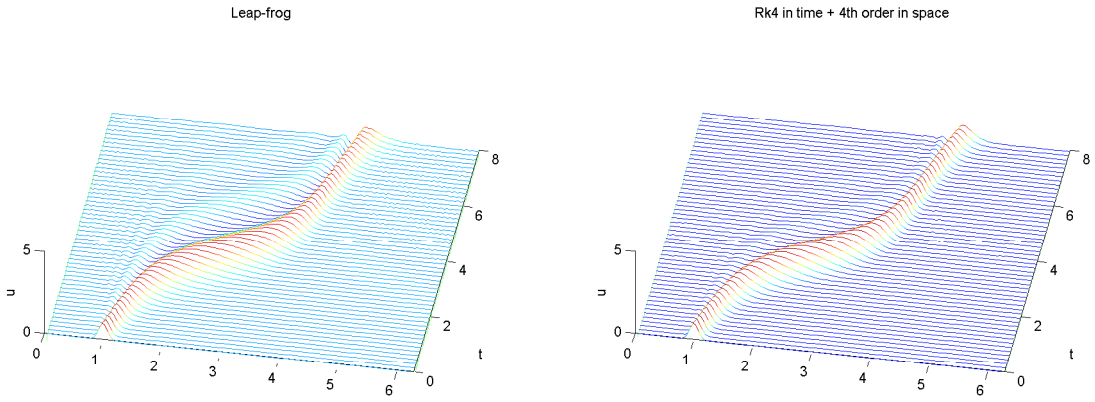
$$y' = 10^{18}(t-1)y, \quad 0 < t < 1 - 10^{-12}$$

and $y(0) = 10^{-6}$. Use a uniform step $k = 10^{-4}$. The exact solution and the midpoint solution both remain below 10^{-6} in magnitude, so they are close to each other by less than 10^{-5}. For the trapezoidal scheme, on the other hand, we have $(t-1)10^{18}k < -10^6 k = -10^2 \ll -2$, so

$$y_N = \frac{2 + k\lambda(t_{N-1})}{2 - k\lambda(t_N)} y_{N-1} \approx -\frac{\lambda(t_{N-1})}{\lambda(t_N)} y_{N-1} \approx \ldots \approx (-1)^N \frac{\lambda(t_0)}{\lambda(t_N)} y_0.$$

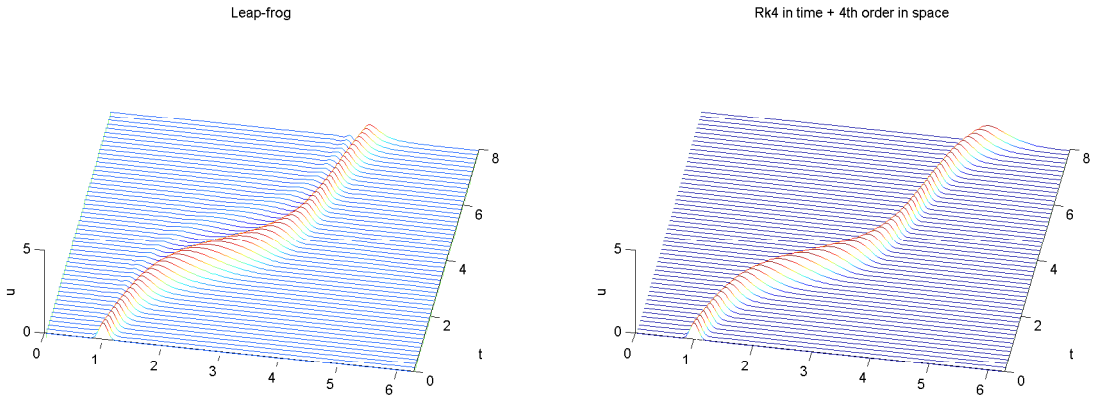
Hence

$$|y_N| \approx 10^{12-6} = 10^6.$$



(a) Order (2,2) (b) Order (4,4)

Figure 4: Solutions for Exercise 7.4, $J = 128$.



(a) Order (2,2) (b) Order (4,4)

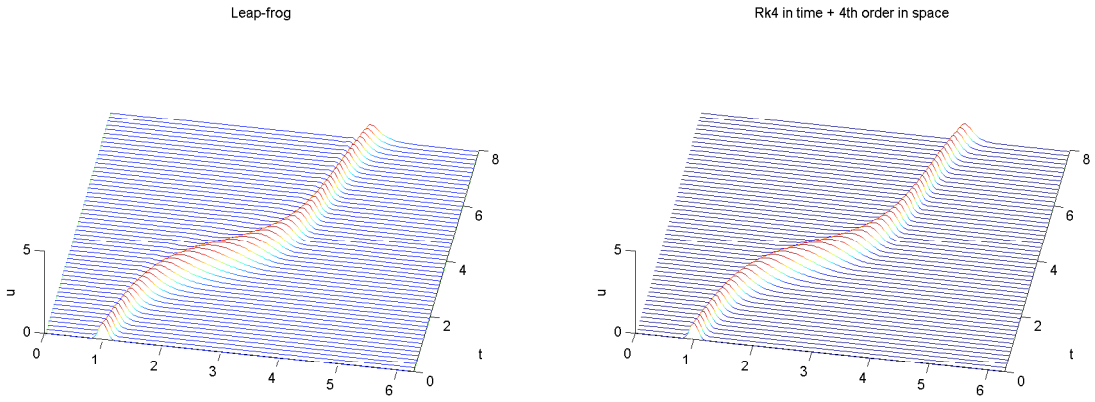
Figure 5: Solutions for Exercise 7.4, $J = 192$.

Chapter 7

7.4 The required plots are displayed in Figures 4–6. Clearly there is dispersion visible; however, this effect diminishes as the approximation improves, either by raising the order or by decreasing the step sizes k and h. For leapfrog there is still visible dispersion even for $J = 256$. For the slightly dissipative and more accurate RK4 the effect is less pronounced and indeed invisible to the naked eye for $J = 256$.

7.5 (a) We have the ODE system

$$\frac{d^2 v_j}{dt^2} = \frac{c^2}{h^2}\left(v_{j-1} - 2v_j + v_{j+1}\right), \quad j = 1, \ldots, J,$$



(a) Order (2,2)

(b) Order (4,4)

Figure 6: Solutions for Exercise 7.4, $J = 256$.

with $v_0 = v_{J+1} = 0$. This can apparently be written as $\frac{d\mathbf{v}}{dt} = -B\mathbf{v}$ with

$$B = \frac{c^2}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

(b) The required matrix C is the result of a Cholesky decomposition. For example, in MATLAB, `C = chol(B)`.

(c) Obviously the sought matrix

$$L = \begin{pmatrix} 0 & C^T \\ -C & 0 \end{pmatrix}$$

is skew-symmetric.

Let $\frac{d\mathbf{z}}{dt} \equiv \mathbf{z}' = L\mathbf{z}$. Multiplying both sides by $\mathbf{z}^T$, it follows that in the 2-norm

$$\frac{1}{2}\frac{d}{dt}\|z(t)\|^2 = \sum_i z_i z_i' = \mathbf{z}^T L\mathbf{z} = 0.$$

Hence $\|\mathbf{z}(t)\| = \|\mathbf{z}(0)\|$ for all t, as in Section 5.3.2. Translating notation we get the required result.

(d) I have calculated $\mathbf{w}^n$ from

$$C^T\mathbf{w}^n = \frac{\mathbf{v}^{n+1} - \mathbf{v}^{n-1}}{2k},$$

but it can equally well be calculated from

$$\mathbf{w}^{n+1/2} = \mathbf{w}^{n-1/2} - kC\mathbf{v}^n$$

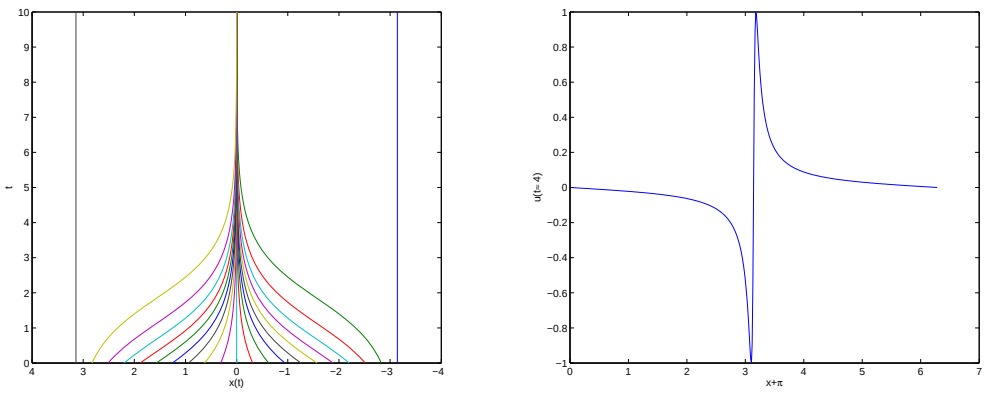
followed by averaging. This introduces an $O(k^2)$ error which is relatively independent of dispersion effects. The error in the invariant arises mainly around when t is a multiple of 10, which is when the pulse reaches the boundary and is reflected off it. Thus, the error in the invariant is less affected than the error in the solution by dispersion, but as we take both k and h smaller with μ fixed this error does not decrease as fast as the error in the solution because points closer and closer to the boundary are sampled. If we keep h fixed and let k go smaller then the invariant error decreases like $O(k^2)$, as expected, and becomes much smaller than the error in the solution.

Chapter 10

10.2 (a) The information is carried along characteristics, and the PDE is linear, but the characteristics are not straight lines. They are defined by

$$\frac{dx}{dt} = -\sin x.$$

For $-\pi < x(0) < \pi$ these curves all tend towards 0, see Figure 7(a).



(a) characteristic curves (b) solution at an early time

Figure 7: Exercise 10.2.

They do not cross! Hence there is no discontinuity in the analytical sense. However, the solution develops a very steep profile, depicted in Figure 7(b) before it becomes much steeper. Note that for $x > 0$, $u_0(x) = \sin x > 0$ and for $x < 0$, $u_0(x) < 0$. Thus, all positive value range of $\sin x$ gets crammed around $x = 0^+$ and all negative range of $\sin x$ gets crammed around $x = 0^-$.

(b) A method on a uniform mesh based on centered spatial discretization and RK4, say, would be fine with $h = .1$ for $t = 1$ but not for $t = 10$. At the latter time oscillations develop. An upwind discretization does better, producing no oscillation, although the solution maximum and minimum shrink in absolute value as time progresses.

A characteristics method whereby we (approximately) find the characteristics $x_j(t)$ for values of an initially uniform mesh, $x_j(0) = -\pi + jh,\ j = 1, \ldots, J,\ Jh = 2\pi$, and assign $u(x_j(t)) = \sin(-\pi + jh)$, is very accurate at these highly nonuniform mesh points.