

Physics 234: Computational Physics

Final Exam

Friday, April 17, 2009 / 14:00–17:00 / V103

Student's Name: Kevin Beach

Instructions

There are ten questions. You should attempt all of them. Mark your response on the test paper in the space provided. For those question that ask you to sketch a diagram, please provide meaningful labels. You are allowed one 8.5×11 sheet of notes (written on both sides). No other aids are permitted.

Good luck!

| <i>question</i> | <i>points awarded</i> |
|-----------------|-----------------------|
| 1 | /5 |
| 2 | /6 |
| 3 | /3 |
| 4 | /3 |
| 5 | /5 |
| 6 | /9 |
| 7 | /3 |
| 8 | /4 |
| 9 | /5 |
| 10 | /7 |
| | 50 / 50 |

Exam questions

1. Let's consider a fictitious 20-bit floating-point type that has 1 sign bit, 7 exponent bits, and 12 fraction bits. The exponent field is interpreted with an offset of 63. In other words,

$$[s/e_6e_5 \dots e_1e_0/f_{11}f_{10} \dots f_1f_0] = (-1)^s \times 2^{(e_6e_5 \dots e_1e_0)_2 - 63} \times (1 + .f_{11}f_{10} \dots f_1f_0)_2.$$

This representation is simplified in that it has *no reserved values* (for inf, nan, etc.) and *only properly normalized numbers are used*. As usual, the leading significand bit is hidden.

- (a) (1 point) In base 2,

$$\pi \doteq 11.001001000011111101101010100010 \dots$$

How is this number stored in the 20-bit floating-point format described above? (Use rounding rather than truncation.)

0 / 10000000 / 100100100010

- (b) (1 point) What are the largest and smallest positive values that can be represented?

0 / 11111111 / 1111111111111111

0 / 00000000 / 0000000000000000

- (c) (1 point) Is it possible to represent zero? **No**

- (d) (2 points) What is the floating-point difference between π and $25/8$?

□ / 01110001 / 000100000000

$$\frac{25}{8} = \frac{24+1}{8} = 3 + \frac{1}{8} = (11.001)_2$$

$$\begin{array}{r} 11.00100100010 \\ - 11.00100000000 \\ \hline 0.00000100010 \end{array}$$

$$\begin{aligned} 0.00000100010 &= 1.00010 \times 2^{-6} \\ &= 1.00010 \times 2^{57-63} \\ &\quad \uparrow \quad \quad \quad \uparrow \\ &\quad \text{hidden bit} \quad \quad \quad 2 \end{aligned}$$

| | |
|---------|----------|
| 0111111 | 63 |
| 0111110 | 62 |
| 0111100 | 60 |
| 0111000 | 56 |
| 0111001 | 57 |
| ↑ | exponent |

2. Consider the sequence of functions

$$f_0(x) = x$$

$$f_1(x) = (1+x)x(1-x)$$

$$f_2(x) = \frac{1}{4}(2+x)(1+x)x(1-x)(2-x)$$

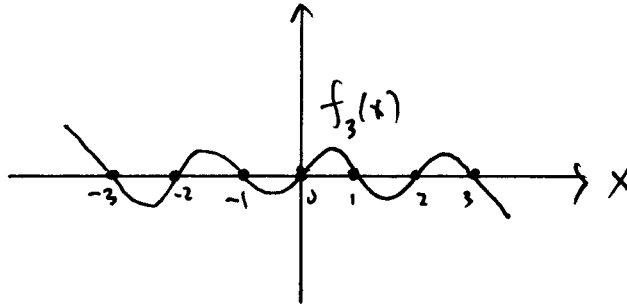
⋮

$$f_n(x) = \frac{1}{(n!)^2}(n+x) \cdots (2+x)(1+x)x(1-x)(2-x) \cdots (n-x),$$

which converges to the period-2 sinusoid

$$\lim_{n \rightarrow \infty} f_n(x) = \frac{1}{\pi} \sin(\pi x).$$

(a) (1 point) Sketch $f_3(x)$.



(b) (1 point) The n th function in the sequence can be written as a product in two ways:

$$f_n(x) = \underbrace{\frac{x}{(n!)^2} \prod_{k=1}^n (k^2 - x^2)}_{\text{version 1}} = x \underbrace{\prod_{k=1}^n \left[1 - \left(\frac{x}{k} \right)^2 \right]}_{\text{version 2}}.$$

Consider the corresponding C++ implementations listed below.

```
double f_v1(double x, unsigned long int N)
{
    if (N == 0) return x;
    double prod = 1.0-x*x;
    unsigned long int fac = 1;
    for (unsigned long int n = 2; n <= N; ++n)
    {
        prod *= n*n-x*x;
        fac *= n;
    }
    const double fac1 = double(fac);
    const double fac2 = fac1*fac1;
    return x*prod/fac2;
}
```

```

double f_v2(double x, unsigned long int N)
{
    if (N == 0) return x;
    double prod = 1.0-x*x;
    for (unsigned long int n = 2; n <= N; ++n)
    {
        const double xn = x/n;
        prod *= 1.0-xn*xn;
    }
    return x*prod;
}

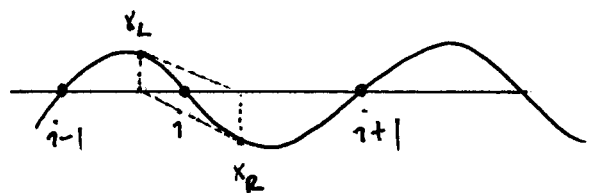
```

Which of these implementations fails at $N = 13$? Explain why.

Version one fails. The explicit computation of $n!$ overflows.

- (c) (2 points) The function $\sin(\pi x)$ has roots at all integer values of its argument. Show that the Newton-Raphson basis of attraction $(i - w/2, i + w/2)$ around each root i has a width w given by the equation

$$w = \frac{1}{\pi} \tan\left(\frac{\pi w}{2}\right).$$



$$x_L = x_R - \frac{f(x_R)}{f'(x_R)}$$

$$\Rightarrow i - \frac{w}{2} = i + \frac{w}{2} - \frac{\sin \pi(i + \frac{w}{2})}{\pi \cos \pi(i + \frac{w}{2})}$$

$$\text{or } w = \frac{1}{\pi} \tan \frac{\pi w}{2}$$

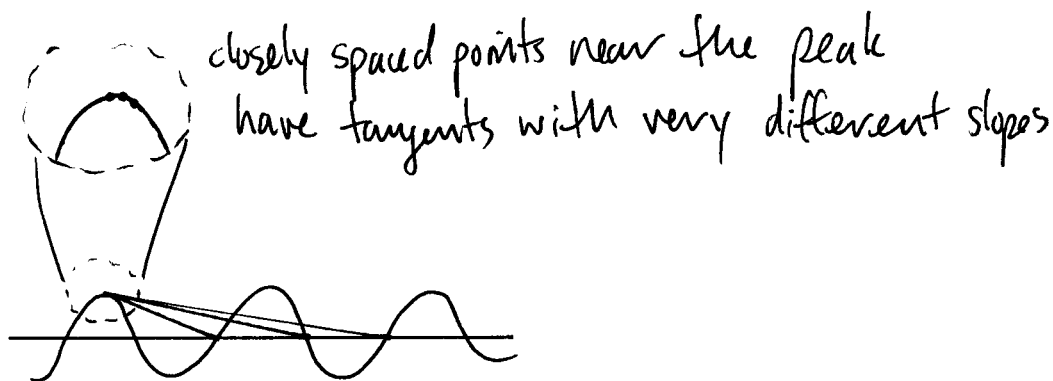
(d) (1 point) Explain how you might go about determining w numerically.

$w = \frac{1}{\pi} \tan\left(\frac{\pi w}{2}\right)$ can itself be written as a root-finding problem on $g(w) = w - \frac{1}{\pi} \tan\left(\frac{\pi w}{2}\right)$

Hence, guess w_0 and generate subsequent

$$w_{n+1} := w_n - \frac{g(w_n)}{g'(w_n)} = w_n - \frac{w_n - \frac{1}{\pi} \tan\left(\frac{\pi w_n}{2}\right)}{\frac{\pi}{2} (\tan^2\left(\frac{\pi w_n}{2}\right) - 1)}$$

(e) (1 point) Draw a diagram that illustrates why it is so hard to predict which root Newton's method will converge to when the initial guess lies outside these basins of attraction.



3. Read over the following code.

```
#include <cmath>
using std::fabs;
using std::sqrt;

#include <algorithm>
using std::sort;

double mystery(double u[3])
{
    double v[3] = { fabs(u[0]), fabs(u[1]), fabs(u[2]) };
    sort(v,v+3);
    const double w0 = v[0]/v[2];
    const double w1 = v[1]/v[2];
    return v[2]*sqrt(1.0+w0*w0+w1*w1);
}
```

(a) (2 points) What does the `mystery` function compute?

$$\vec{v} = (v_1, v_2, v_3) \xrightarrow{\substack{\text{3-vector} \\ \text{norm}}} |\vec{v}| = \sqrt{v_1^2 + v_2^2 + v_3^2}$$

(b) (1 point) Why is it beneficial to organize the calculation this way?

Numerically stable in the case where one component of the vector dominates the others

4. (3 points) Explain why the expression

$$\frac{f(x+h) - f(x)}{h},$$

computed with double-precision floating point arithmetic, has roughly

$$16 + \log_{10} \left| \frac{f'(x)h}{f(x)} \right|$$

significant decimal digits.

Represented as a double, $f(x)$ has roughly 16 digits of decimal precision, all of which are significant. The same holds for $f(x+h)$.

$$\begin{aligned} \text{If } h \text{ is small, then } f(x+h) &\approx f(x) + f'(x)h \\ &= f(x) \left(1 + \frac{f'(x)h}{f(x)} \right) \end{aligned}$$

Subtraction of the two quantities leads to catastrophic loss of significance since they are nearly equal.

The number of digits lost is $-\log_{10} \left| \frac{f'(x)h}{f(x)} \right|$, so

that

$$16 - \left(-\log_{10} \left| \frac{f'(x)h}{f(x)} \right| \right) = 16 + \log_{10} \left| \frac{f'(x)h}{f(x)} \right| \text{ remain.}$$

5. Suppose we try to solve the nonlinear equation $e^x = 2x^2 + \sin x$ using the secant method with the first two terms of the sequence taken to be $x_0 = 0$, $x_1 = 1$.

(a) (2 points) Show that the next term is

$$x_2 = \frac{1}{3 + \sin(1) - e} \doteq 0.8903.$$

$$\text{let } f(x) = 2x^2 + \sin x - e^x$$

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) = 1 - \frac{(1-0)}{2 + \sin(1) - e + 1} (2 + \sin(1) - e)$$

$$= \frac{(3 + \sin(1) - e) - (2 + \sin(1) - e)}{3 + \sin(1) - e} = \frac{1}{3 + \sin(1) - e}$$

(b) (1 point) Explain why two initial guesses are needed to get this method started.

The secant is formed by the line passing through the coordinates $(x_{n-1}, f(x_{n-1}))$ and $(x_n, f(x_n))$ of the previous two guesses.

(c) (2 points) The sequence (x_n) , generated by the secant method iteration rule, converges to the limiting value $L \doteq 0.9317317339$. We showed in class that the residuals $r_n = L - x_n$ obey the recursion relation $r_{n+1} \sim r_n r_{n-1}$. (In other words, the number of significant bits $b_n = -\log_2 |r_n/L|$ approaches a Fibonacci rule $b_{n+1} = b_n + b_{n-1}$ at large n .) We can characterize the convergence behaviour in terms of the smallest value q such that

$$\lim_{n \rightarrow \infty} \frac{|r_{n+1}|}{|r_n|^q} > 0.$$

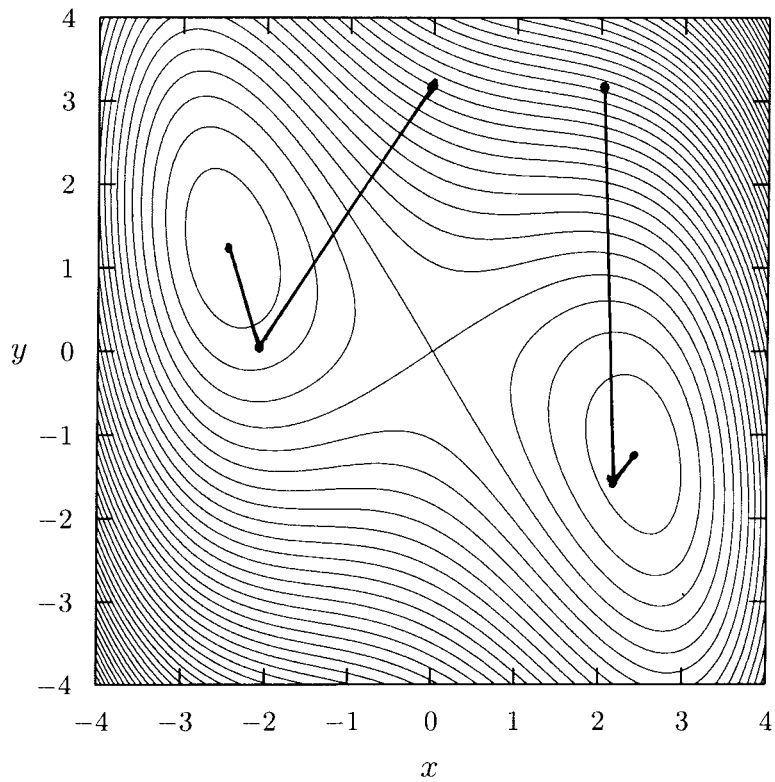
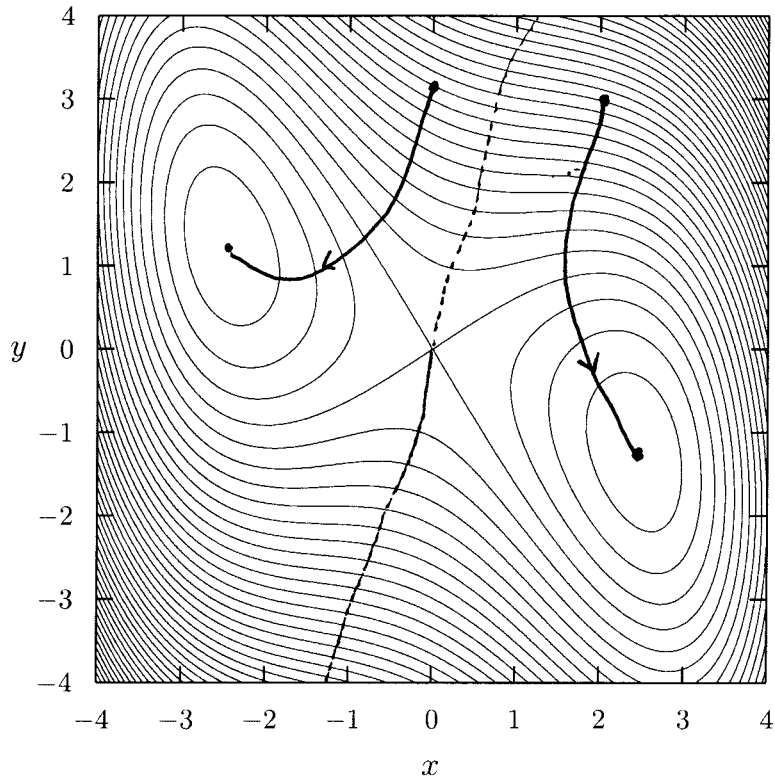
(The equivalent statement is that the number of accurate bits goes up by a factor of q with each iteration.) What is the value of q ?

$$\text{Assume } b_n = b_0 q^n$$

$$\text{Then } b_{n+1} = b_n + b_{n-1} \Rightarrow q^2 = q + 1$$

$$\Rightarrow q = \frac{1 + \sqrt{5}}{2}$$

6. The function $f(x, y) = 9xy - 10x^2 + x^4 + 9y^2$ gives rise to the following contour plot (shown twice for the benefit of questions 5(c) and 5(d)):



(a) (3 points) Identify the three extremal points of f by solving $\nabla f = 0$.

$$f(x, y, z) = 9xy - 10x^2 + x^4 + 9y^2$$

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 9y - 20x + 4x^3 \\ 9x + 18y \end{pmatrix} = 0$$

$$\Rightarrow x + 2y = 0 \quad \text{or} \quad y = -\frac{x}{2}$$

Then

$$-\frac{9x}{2} - 20x + 4x^3 = \left(-\frac{49}{2} + 4x^2\right)x = 0$$

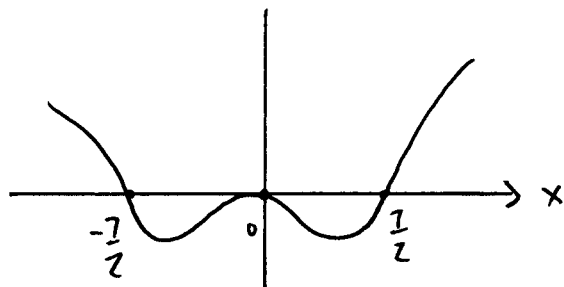
$$\text{has solutions } x = \pm \frac{7}{2\sqrt{2}}, \quad y = \mp \frac{7}{4\sqrt{2}}$$

$$\text{and } x = y = 0$$

- (b) (2 points) Find an expression for $f(x, y)$ evaluated along the line $y = -x/2$. Sketch the resulting one-dimensional function.

substitute $y = -\frac{x}{2}$

$$\begin{aligned}
 & 9xy - 10x^2 + x^4 + 9y^2 \\
 & -9\frac{x^2}{2} - 10x^2 + x^4 + 9\frac{x^2}{4} \\
 & = x^4 - \frac{40x^2}{4} - \frac{10x^2}{4} + \frac{9x^2}{4} \\
 & = x^4 - \frac{49x^2}{4} = x^2 \left(x^2 - \left(\frac{7}{2}\right)^2 \right)
 \end{aligned}$$



- (c) (2 points) Draw directly onto the upper contour plot (the top figure on page 9) two steepest-descent paths starting from the initial points $\mathbf{r} = (2, 3)$ and $\mathbf{r} = (0, 3)$. Each of these will terminate in one of the local minima. Assume that the step size has been taken arbitrarily small, so that the path is continuous: $\dot{\mathbf{r}} \sim -\nabla f(\mathbf{r})$. (Remember that the gradient is always directed perpendicular to the contour lines!) Draw a dotted line that separates the locus of points that flow toward each well.
- (d) (2 points) Draw directly onto the lower contour plot (the bottom figure on page 9) your educated guess as to what the first few Newton's method iterations

$$\mathbf{r}_{n+1} := \mathbf{r}_n - (H^{-1}\nabla f)_{\mathbf{r}_n}$$

might look like starting from the same two points. No calculation is required.

7. (3 points) Show that the expression

$$5x^{10} - 4x^7 - 2x^4 + x^2 - 1$$

can be evaluated with as few as 6 multiplication operations and 4 addition/subtraction operations. (Assume that you can store intermediate results.)

$$(5x^8 - 4x^5 - 2x^2 + 1)x^2 - 1$$

$$= [(5x^6 - 4x^3 - 2)x^2 + 1]x^2 - 1$$

$$= [((5 \cdot x^3 - 4) \cdot x^3 - 2) \cdot x^2 + 1] \cdot x^2 - 1$$

$\begin{matrix} \text{I} & & \text{II} & & \text{III} & & \text{IV} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \uparrow 3 & & \uparrow 4 & & \uparrow 5 & & \uparrow 6 \end{matrix}$

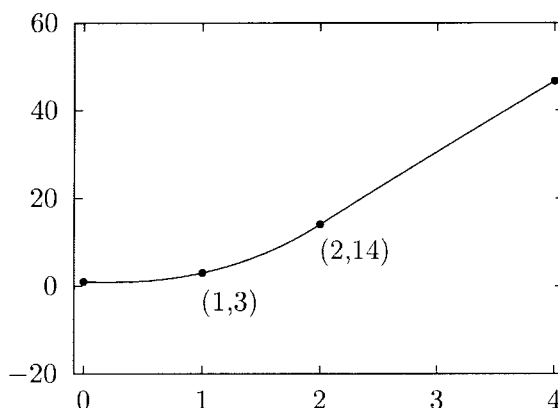
with $x^2 := x \cdot x$

$$x^3 := x^2 \cdot x$$

\uparrow
2

8. (4 points) Consider the piecewise-defined polynomial

$$s(x) = \begin{cases} 1 - x + 2x^2 + x^3 & 0 \leq x \leq 1 \\ x + x^2 + x^3 & 1 < x \leq 2 \\ -\frac{68}{3} + 20x - x^2 + \frac{1}{12}x^3 & 2 < x \leq 4 \end{cases}$$



Determine whether $s(x)$ is a cubic spline on the interval $[0, 4]$. If so, is it natural or clamped? If not, what requirement does it fail to meet?

$$s'(x) = \begin{cases} -1 + 4x + 3x^2 \\ 1 + 2x + 3x^2 \\ 20 - 2x + \frac{1}{4}x^2 \end{cases}$$

$$s''(x) = \begin{cases} 4 + 6x \\ 2 + 6x \\ -2 + \frac{1}{2}x \end{cases}$$

match function and first two derivatives at $x=1$ and $x=2$

$$C^0: s(1) = 1 - 1 + 2 \cdot 1 + 1 = 3$$

$$s(1) = 1 + 1 + 1 = 3 \quad \checkmark$$

$$s(2) = 2 + 4 + 8 = 14$$

$$s(2) = -\frac{68}{3} + 20 \cdot 2 - 4 + \frac{1}{12} \cdot 8$$

$$= 36 - \frac{68}{3} + \frac{2}{3}$$

$$= 36 - 22 = 14 \quad \checkmark$$

$$C^1: s'(1) = -1 + 4 + 3 = 6$$

$$s'(1) = 1 + 2 + 3 = 6 \quad \checkmark$$

$$s'(2) = 1 + 4 + 3 \cdot 4 = 17$$

$$s'(2) = 20 - 4 + 1 = 17 \quad \checkmark$$

$$C^2: s''(0) = 4 + 6 \cdot 0 = 4$$

$$s''(1) = 4 + 6 = 10 \quad \times$$

$$s''(1) = 2 + 6 = 8 \quad \times$$

$$s''(2) = -2 + 12 = 10 \quad \times$$

$$s''(2) = -2 + 1 = -1 \quad \times$$

$$s''(4) = 0$$

9. An ideal spring obeys Hooke's force law $F = -k(x - x_0)$, where x_0 is the equilibrium length and k is the spring constant. Suppose we measure the force $F_i \pm \delta F_i$ that results when the spring is stretched and held fixed at each of the displacements x_i ($i = 1, 2, \dots, N$). The δF_i values represent the experimental uncertainty.

- (a) (3 points) A linear regression analysis leads to estimates

$$\hat{k} = \frac{\langle x \rangle \langle F \rangle - \langle xF \rangle \langle 1 \rangle}{\langle x^2 \rangle \langle 1 \rangle - \langle x \rangle^2}, \quad \hat{x}_0 = \frac{\langle x^2 \rangle \langle F \rangle - \langle x \rangle \langle xF \rangle}{\langle x \rangle \langle F \rangle - \langle xF \rangle \langle 1 \rangle},$$

where the angled brackets denoted the weighted average

$$\langle A \rangle = \frac{1}{N} \sum_{i=1}^N \frac{A_i}{(\delta F_i)^2}.$$

Write an expression for the non-negative quantity that is minimized for this fit.

linear regression \Rightarrow weighted sum of squares of residuals is minimized

$$\chi^2 = \left\langle \left(F_i + k(x_i - x_0) \right)^2 \right\rangle = \frac{1}{N} \sum_{i=1}^N \left(\frac{F_i + k(x_i - x_0)}{\delta F_i} \right)^2$$

- (b) (1 point) Under what circumstances is this minimized quantity exactly zero?

When $N < 3$, since the system of equations is underdetermined

- (c) (1 points) Explain how you would use a Jackknife analysis to estimate the uncertainties on the fitted values \hat{k} and \hat{x}_0 .

Recompute fit N times by removing each data point in turn from the set. Treat the resulting \hat{k} and \hat{x}_0 values as a distribution.

10. A mass m is attached to a spring with spring constant k and equilibrium length x_0 . The entire system is immersed in a viscous liquid with drag coefficient b . The mass experiences a force $F = -k(x - x_0) - bv$ that depends on both its current position x and velocity v . Its Newtonian dynamics can be described by two coupled first-order equations

$$\dot{v} = \frac{F}{m} = -\frac{1}{m}[k(x - x_0) + bv],$$

$$\dot{x} = v.$$

To solve this system of equations numerically, let's discretize the time variable using a uniform time step Δt . We define $v_n \equiv v(t_n) \equiv v(n \cdot \Delta t)$ and $x_n \equiv x(t_n) \equiv x(n \cdot \Delta t)$. The nature of the iteration scheme will depend on how we go about computing the following integrals:

$$v_{n+1} = v_n + \frac{1}{m} \int_{t_n}^{t_{n+1}} dt F(x(t), v(t)),$$

$$x_{n+1} = x_n + \int_{t_n}^{t_{n+1}} dt v(t).$$

- (a) (4 points) Approximate the integrals using the trapezoid rule. Show that the appropriate iteration rule for this scheme is

$$v_{n+1} := \frac{v_n \left(1 - \frac{b\Delta t}{2m} - \frac{k(\Delta t)^2}{4m}\right) - \frac{k\Delta t}{m}(x_n - x_0)}{\left(1 + \frac{b\Delta t}{2m} + \frac{k(\Delta t)^2}{4m}\right)}$$

$$x_{n+1} := x_n + \frac{\Delta t}{2}(v_n + v_{n+1})$$

$$v_{n+1} = v_n + \frac{1}{m} \int_{t_n}^{t_{n+1}} dt F(x(t), v(t)) \approx v_n + \frac{\Delta t}{2m} (F(x_{n+1}, v_{n+1}) + F(x_n, v_n))$$

$$= v_n - \frac{\Delta t}{2m} \left[k(x_{n+1} - x_0) + bv_{n+1} + k(x_n - x_0) + bv_n \right]$$

$$x_{n+1} = x_n + \frac{1}{m} \int_{t_n}^{t_{n+1}} dt v(t) \approx x_n + \frac{\Delta t}{2m} (v_{n+1} + v_n)$$

Linear system of equations in x_{n+1} and v_{n+1} !

$$v_{n+1} = v_n - \frac{\Delta t}{2m} \left[k \left(x_n + \frac{\Delta t}{2m} (v_{n+1} + v_n) - x_0 \right) + b v_{n+1} + k (x_n - x_0) + b v_n \right]$$

$$\Rightarrow v_{n+1} \left(1 + \frac{b \Delta t}{2m} + \frac{k (\Delta t)^2}{4m} \right) = v_n \left(1 - \frac{b \Delta t}{2m} - \frac{k (\Delta t)^2}{4m} \right) - \frac{k \Delta t}{m} (x_n - x_0)$$

Update rules:

$$v_{n+1} := \frac{v_n \left(1 - \frac{b \Delta t}{2m} - \frac{k (\Delta t)^2}{4m} \right) - \frac{k \Delta t}{m} (x_n - x_0)}{1 + \frac{b \Delta t}{2m} + \frac{k (\Delta t)^2}{4m}}$$

$$x_{n+1} := x_n + \frac{\Delta t}{2m} (v_{n+1} + v_n)$$

- (b) (1 point) We seem to have solved the "knowledge of the future" problem. How did we do that here? And why can't we always do it?

The terms x_{n+1} and v_{n+1} show up on the rhs only in linear combinations.

- (c) (2 points) What's an appropriate choice of Δt ? Can you put rough bounds on how small it ought to be?

We want to ensure that both $\frac{b \Delta t}{2m}$ and

$\frac{k (\Delta t)^2}{4m}$ are small.

$$\Rightarrow \Delta t \ll \min \left(\frac{2m}{b}, \sqrt{\frac{4m}{k}} \right)$$