Name: SOLUTIONS

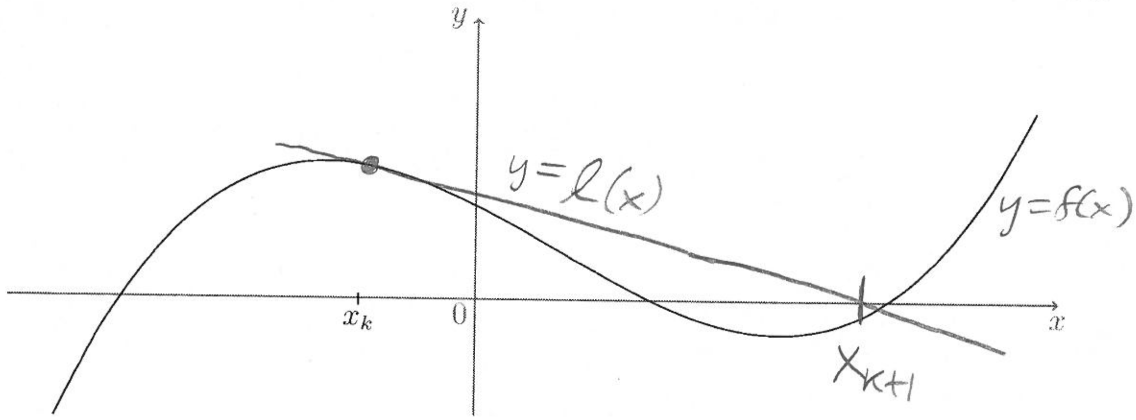
Math 310 Numerical Analysis (Bueler)

Thursday, 17 October 2019

Midterm Exam

In class. No book, electronics, or notes. 95 minutes maximum. 115 points possible.

1. (a) (*5 pts*) A differentiable function $f(x)$ and an iterate x_k are shown on the axes below. Sketch, with appropriate labeling, how Newton's method determines the next iterate x_{k+1}.



(b) (*10 pts*) Suppose $\ell(x)$ is the linearization of $f(x)$ at x_k. Give a formula for $\ell(x)$ and then a formula for where it crosses the x-axis. Write the result as Newton's method in the box below.

$$\ell(x) = f(x_k) + f'(x_k)(x - x_k)$$

$$\ell(x) = 0 \iff 0 = f(x_k) + f'(x_k)(x - x_k)$$

$$\iff x - x_k = \frac{-f(x_k)}{f'(x_k)}$$

$$\iff x = x_k - \frac{f(x_k)}{f'(x_k)}$$

$$\boxed{x_{k+1} = x_k - f(x_k)/f'(x_k)}$$

2. **(a)** (*10 pts*) Consider the equation $x^3 - 3x + 1 = 0$ and suppose $a_0 = -1$ and $b_0 = 1$ is a bracket. Apply two steps of the bisection method, reporting the bracket at the end of each step.

$$f(x) = x^3 - 3x + 1$$

$$f(a_0) = f(-1) = +3, \quad f(b_0) = f(1) = -1$$

$[a_0, b_0] = [-1, +1]$: $\quad c = 0, \quad f(c) = +1$

$[a_1, b_1] = [0, +1]$: $\quad c = \frac{1}{2}, \quad f(c) = \frac{1}{8} - \frac{3}{2} + 1 < 0$

$[a_2, b_2] = [0, \frac{1}{2}]$

(b) (*10 pts*) For the same equation as in **(a)**, with $x_0 = -1$ and $x_1 = 1$, apply one step of the secant method.

$$x_{k+1} = x_k - \frac{f(x_k)}{\underbrace{\left\{\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\right\}}_{\text{secant slope}}}$$

$$x_2 = x_1 - \frac{f(x_1)}{\left\{\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right\}}$$

$$= 1 - \frac{-1}{\left\{\frac{-1-(3)}{+1-(-1)}\right\}} = 1 - \frac{-1}{\left(\frac{-4}{2}\right)} = 1 - \frac{1}{2}$$

$$= \frac{1}{2}$$

3. (*15 pts*) Solve the following system by Gauss elimination and back-substitution:

$$3x_1 + x_2 + 2x_3 = 11$$
$$3x_1 + 4x_2 + x_3 = 14$$
$$-3x_1 + 5x_2 - 2x_3 = 1$$

Show your steps in an organized way. It must be clear that you are following the algorithm we considered in class. (*Hints: Pivoting is* not *requested. The numbers are integers at every stage if you follow the algorithm.*)

G.E.

$$\left.\begin{array}{ll} & 3x_1 + x_2 + 2x_3 = 11 \\ R_2 \Leftarrow R_2 - R_1 & 3x_2 - x_3 = 3 \\ R_3 \Leftarrow R_3 + R_1 & 6x_2 = 12 \end{array}\right] \text{stage 1}$$

$$\left.\begin{array}{ll} & 3x_1 + x_2 + 2x_3 = 11 \\ & 3x_2 - x_3 = 3 \\ R_3 \Leftarrow R_3 - 2R_2 & 2x_3 = 6 \end{array}\right] \text{Stage 2}$$

B.S.

$$x_3 = \frac{6}{2} = 3$$

$$x_2 = \frac{3 + x_3}{3} = 2$$

$$x_1 = \frac{11 - x_2 - 2x_3}{3} = \frac{3}{3} = 1$$

$$\therefore \quad \vec{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

4. *(5 pts)* Suppose we have used Gauss elimination with partial pivoting to factor some matrix A, so that $PA = LU$. Here P is a known permutation matrix, L is a known lower-triangular matrix, and U is a known upper-triangular matrix. Explain how to use this factorization to easily solve $A\mathbf{x} = \mathbf{b}$, assuming $\mathbf{b}$ is also given, and identify what algorithms are needed.

$Ax=b \iff PAx=Pb \iff LUx=Pb$

so ① solve $Ly=Pb$ by forward substitution

② solve $Ux=y$ by back substitution

5. *(10 pts)* Consider lower-triangular matrices with unit diagonal:

$$L = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \ell_{2,1} & 1 & 0 & & 0 \\ \ell_{3,1} & \ell_{3,2} & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ \ell_{n,1} & \ell_{n,2} & \dots & \ell_{n,n-1} & 1 \end{bmatrix}$$

$Ly = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$

Write a MATLAB/OCTAVE code, or a complete pseudocode, to solve systems $L\mathbf{y} = \mathbf{b}$, with L in the above form, by forward substitution.

```
function y = forwardsub(L,b)
```

```
n = length(b);
[optional check on inputs:   is L the right size?
   is L lower-triangular?   diag(L)==1?]
y = zeros(n,1);
y(1) = b(1);
for i=2:n
    s = b(i);
    for j=1:i-1
        s = s - L(i,j)*y(j);
    end
    y(i) = s;
end
```

6. Suppose that the IEEE standard for floating point representation discussed in class had an 8 bit version. It might look like this:

s	e_1	e_2	e_3	b_1	b_2	b_3	b_4

These 8 bits would represent the number

$$x = (-1)^s\,(1.b_1b_2b_3b_4)_2 \times 2^{(e_1e_2e_3)_2-3}.$$

However, normal numbers would not use exponents $(000)_2$ nor $(111)_2$, which have special uses.

(a) (*5 pts*) What is the representation of 1 (one) in this system? (*Give all the bits.*)

$$1 = (-1)^0\,(1.0000)_2 \times 2^{(011)_2-3}$$

0	0	1	1	0	0	0	0

(b) (*5 pts*) What is the largest number that this system can represent?

$$x = (-1)^0\,(1.1111)_2 \times 2^{(110)_2-3}$$

$$= \left(1+\tfrac{1}{2}+\tfrac{1}{4}+\tfrac{1}{8}+\tfrac{1}{16}\right) \times 2^{6-3} = \left(2-\tfrac{1}{16}\right)\times 2^3$$

$$= 2^4 - \tfrac{1}{2} = 15.5$$

(c) (*5 pts*) What is the value of "machine epsilon" in this system?

$$\varepsilon = (1.0001)_2 - (1.0000)_2 = \tfrac{1}{16}$$

(d) (*3 pts*) How would zero be represented? (*Give all the bits.*)

$0 =$

0	0	0	0	0	0	0	0

(e) (*2 pts*) Give the bits of a nonzero subnormal number. (*Just pick one.*)

0	0	0	0	0	1	0	1

7. **(a)** (*5 pts*) Find all the fixed points x_* of $\varphi(x) = \frac{1}{4}x^2 + \frac{3}{4}x - \frac{1}{2}$. (*Hint: There are two.*)

$$x = \varphi(x) = \frac{1}{4}x^2 + \frac{3}{4}x - \frac{1}{2}$$

$$4x = x^2 + 3x - 2$$

$$x^2 - x - 2 = 0$$

$$(x-2)(x+1) = 0$$

$$x_* = -1, +2$$

$$\varphi'(x) = \frac{1}{2}x + \frac{3}{4}$$

(b) (*10 pts*) Recall Theorem 4.5.1:

Theorem. Assume that $\varphi \in C^1$ and $|\varphi'(x)| < 1$ in some interval $[x_* - \delta, x_* + \delta]$ around a fixed point x_* of φ. If x_0 is in this interval then the fixed point iteration converges to x_*.

For the same function $\varphi(x)$ as in part **(a)**, will the iteration

$$x_{k+1} = \varphi(x_k)$$

converge to each fixed point x_* for all x_0 near x_*? (*Hint: Consider each x_* in turn.*)

$x_* = -1$:

$$\varphi'(-1) = -\frac{1}{2} + \frac{3}{4} = \frac{1}{4}$$

$|\varphi'(-1)| < 1$ so the iteration converges if x_0 is close to -1

$x_* = +2$:

$$\varphi'(+2) = \frac{1}{2} \cdot 2 + \frac{3}{4} = 1.75$$

$|\varphi'(+2)| > 1$ so the iteration does not have to converge if x_0 is close to $+2$

8. **(a)** (*5 pts*) State Taylor's theorem with remainder in the $n = 2$ case. Be sure to include the assumptions about the function $f(x)$.

Theorem. If f has three continuous derivatives on an interval including a and x then there is ξ between a and x so that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \frac{f'''(\xi)}{3!}(x-a)^3$$

(b) (*5 pts*) Using the Theorem above, find the quadratic (degree two) Taylor polynomial for $f(x) = \sqrt{x}$ using basepoint $a = 4$.

$$f(x) = \sqrt{x}$$
$$f'(x) = \tfrac{1}{2}x^{-1/2}$$
$$f''(x) = -\tfrac{1}{4}x^{-3/2}$$
$$f'''(x) = +\tfrac{3}{8}x^{-5/2}$$

$$P_2(x) = f(4) + f'(4)(x-4) + \frac{f''(4)}{2}(x-4)^2$$
$$= 2 + \frac{1}{4}(x-4) - \frac{1}{64}(x-4)^2$$

(c) (*5 pts*) The result of **(b)** is a polynomial $P_2(x)$ such that $f(x) \approx P_2(x)$. Use the Theorem in **(a)** to estimate the size of the error $|P_2(x) - f(x)|$ for all x in the interval $[3, 5]$.

$$-1 \le x-4 \le 1$$

$$|P_2(x) - f(x)| = \left| \frac{f'''(\xi)}{3!}(x-4)^3 \right|$$

$$\le \frac{\frac{3}{8}(3)^{-5/2}}{6} \cdot 1^3 = \frac{3}{8 \cdot 6 \cdot 3^{5/2}} = \frac{3}{8 \cdot 6 \cdot 3^2 \cdot 3^{1/2}}$$

$$= \frac{1}{144\sqrt{3}}$$

Extra Credit. (*3 pts*) Do one step of Newton's method to find the first-quadrant intersection of the circle $x^2 + y^2 = 4$ and the graph $y = e^x$. Start from $(x_0, y_0) = (1, 2)$, which is not too far from the intersection.

[one way!]

$$\left.\begin{aligned} \vec{x}_{k+1} &= \vec{x}_k + \vec{s} \\ J(\vec{x}_k)\,\vec{s} &= -\vec{F}(\vec{x}_k) \end{aligned}\right\} \text{Newton's method for systems}$$

$$\vec{F}(\vec{x}) = \begin{bmatrix} x_1^2 + x_2^2 - 4 \\ x_2 - e^{x_1} \end{bmatrix}, \quad J(\vec{x}) = \left(\frac{\partial F_i}{\partial x_j}\right) = \left[\begin{array}{c|c} 2x_1 & 2x_2 \\ \hline -e^{x_1} & 1 \end{array}\right]$$

$$\vec{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \therefore \vec{F}(\vec{x}_0) = \begin{bmatrix} 1 \\ 2-e \end{bmatrix}, \quad J(\vec{x}_0) = \left[\begin{array}{c|c} 2 & 4 \\ \hline -e & 1 \end{array}\right]$$

$$\left[\begin{array}{c|c} 2 & 4 \\ \hline -e & 1 \end{array}\right] \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} -1 \\ e-2 \end{bmatrix} \Rightarrow \vec{s} = \begin{bmatrix} (7-4e)/(2-4e) \\ (3e-4)/(2-4e) \end{bmatrix}$$

$$\Rightarrow \vec{x}_1 = \vec{x}_0 + \vec{s} = \begin{bmatrix} 1 + (7-4e)/(2-4e) \\ 2 + (3e-4)/(2-4e) \end{bmatrix}$$

BLANK SPACE BLANK SPACE