Name: SOLUTIONS

Math 426 Numerical Analysis (Bueler) Friday, 11 October 2024

Midterm Exam

In class. No book, notes, electronics, or internet. 65 minutes. 100 points.

1. **(a)** *(8 pts)* We want to find a root of $f(x) = x^3 - 2x - 2$. Show that $a_0 = 0$ and $b_0 = 2$ is a bracket. Then apply two steps of the bisection method, reporting the new bracket at the end of each step. (*If you need these facts, you may use them:* $f(0.5) = -2.875$ *and* $f(1.5) = -1.625$.)

- $f(a_0) = f(0) = -2 < 0$, $f(b_0) = f(2) = 8 - 4 - 2 = 2 > 0$ so $[a_0, b_0]$ is a bracket.

① • $c = 1$, $f(c) = 1 - 2 - 2 < 0$ so $[a_1, b_1] = [1, 2]$

② • $c = 1.5$, $f(1.5) = -1.625 < 0$ so $[a_2, b_2] = [1.5, 2]$

(in fact root is approximately $x_* = 1.7693$)

(b) *(8 pts)* For the same equation as in **(a)**, with $x_0 = 1$ and $x_1 = 2$, apply one step of the secant method, to compute x_2.

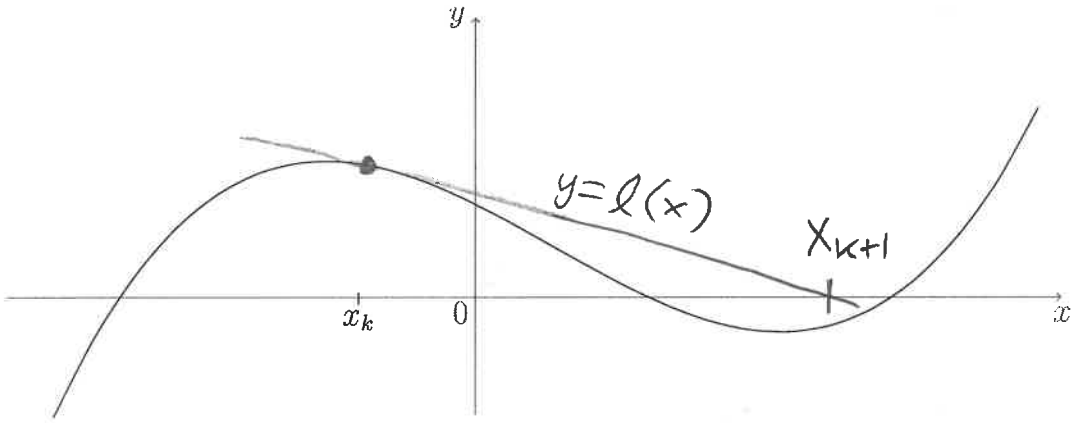
$$x_{k+1} = x_k - \frac{f(x_k)}{\left(\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\right)}$$

← think: like Newton, but with secant line slope)

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

$$= 2 - \frac{2}{\left(\frac{2 - (-3)}{2 - 1}\right)} = 2 - \frac{2}{5} = 1.6$$

2. **(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.

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

(↑ $\ell(x) = P_1(x)$ is Taylor polynomial of degree one)

$$0 = f(x_k) + f'(x_k)(x_{k+1} - x_k)$$

since we define x_{k+1} as where $\ell(x)$ crosses x-axis

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

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

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

$$2x_1 + 3x_2 + 4x_3 = 6$$
$$-2x_1 - 2x_2 - 3x_3 = -5$$
$$4x_1 + 8x_2 + 8x_3 = 12$$

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

$2x_1 + 3x_2 + 4x_3 = 6$

$R_2 \leftarrow R_2 - \left(\frac{-2}{2}\right)R_1$: $x_2 + x_3 = 1$ } step 1

$R_3 \leftarrow R_3 - \left(\frac{4}{2}\right)R_1$: $2x_2 + 0x_3 = 0$

GE

$2x_1 + 3x_2 + 4x_3 = 6$

$x_2 + x_3 = 1$ } step 2

$R_3 \leftarrow R_3 - \left(\frac{2}{1}\right)R_2$: $-2x_3 = -2$

$x_3 = \frac{-2}{-2} = 1$

$x_2 = \frac{1 - x_3}{1} = \frac{1-1}{1} = 0$

$x_1 = \frac{6 - 3x_2 - 4x_3}{2} = \frac{6 - 0 - 4}{2} = 1$

BS

checking: $2 + 4 = 6$ ✓

$-2 - 3 = -5$ ✓

$4 + 8 = 12$ ✓

4. (*9 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.

(0) $PA = LU$ (GE with partial pivoting; already done)

(0.5) multiply Pb

(1) solve $Ly = Pb$ by forward substitution

(2) solve $Ux = y$ by back substitution

[scratch: $PAx = Pb \Leftrightarrow L(Ux) = Pb$]

5. (*9 pts*) The following Matlab code solves a unit lower diagonal linear system $L\mathbf{y} = \mathbf{b}$. That is, L is an n by n lower-triangular matrix, with ones on the diagonal, and $\mathbf{b}$ is an n by 1 column vector. The output $\mathbf{y}$ is also an n by 1 column vector. Please count **exactly how many floating point operations it does**, and give the answer as a function of n.

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

n = length(b);                      % get size
y = zeros(n,1);                     % allocate space for answer
for i = 1:n                         % count through the rows
    y(i) = b(i);                    % start the numerator
    for j = 1:i-1
        y(i) = y(i) - L(i,j) * y(j);  % use y(j) which are already known
    end
end
```

2 flops

$$(\text{total flops}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{i-1} 2 \right)$$

$$= \sum_{i=1}^{n} 2(i-1) = 2\sum_{i=1}^{n} i - \sum_{i=1}^{n} 2$$

$$= 2\frac{n(n+1)}{2} - 2n = n^2 + n - 2n = n^2 - n$$

alternate method: see the pattern for small n cases

$n=1$: no work

$n=2$: $2 = 2 \cdot 1$

$n=3$: $2 + (2+2) = 6 = 3 \cdot 2$

$n=4$: $2 + (2+2) + (2+2+2) = 12 = 4 \cdot 3$

$n=5$: $2 + (2+2) + (2+2+2) + (2+2+2+2) = 20 = 5 \cdot 4$

$\therefore$ $n(n-1) = n^2 - n$

6. Suppose that the IEEE 754 standard for floating point representation had a 9 bit version. It might look like this:

s	e_1	e_2	e_3	e_4	b_1	b_2	b_3	b_4

These 9 bits would represent the number

$$x = (-1)^s \, (1.b_1b_2b_3b_4)_2 \times 2^{(e_1e_2e_3e_4)_2 - 7}.$$

$(7)_{10} = (0111)_2$

The exponent bits $(0000)_2$ nor $(1111)_2$ have special uses, so they are not allowed for regular, i.e. representable, nonzero numbers in this system.

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

$1 = (-1)^0 (1.0000)_2 \times 2^{(0)}$ ← $0 = 7 - 7$

0	0	1	1	1	0	0	0	0

(b) *(5 pts)* How is "machine epsilon" $\epsilon_{\text{machine}}$ defined, and what is its value in this system?

$\varepsilon_m = (1+\varepsilon_m) - 1 = (1.0001)_2 - (1.0000)_2$

$= (0.0001)_2 = \frac{1}{16} = (0.0625)_{10}$

$\frac{1}{2}\ \frac{1}{4}\ \frac{1}{8}\ \frac{1}{16}$

(c) *(5 pts)* What is the largest representable number in this system?

$X = (-1)^0 (1.1111)_2 \times 2^{(1110)_2 - (0111)_2}$

$= \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right) \times 2^{14-7} = \left(2 - \frac{1}{16}\right) \times 2^7$

$= \frac{31}{16} \cdot 128 = 31 \cdot 8 = 248$

(Extra Credit) *(1 pts)* Show the bits of an example (nonzero) subnormal number in this system.

0	0	0	0	0	1	0	1	0

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

$X = \frac{1}{4}x^2 + \frac{3}{4}$ $\quad 4x = x^2 + 3$

$x^2 - 4x + 3 = 0$

$(x-1)(x-3) = 0$

$x_* = 1, 3$

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

(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_*? Consider each x_* in turn.

$x_* = 1$: $|\varphi'(1)| = \frac{1}{2} < 1$ will converge

$x_* = 3$: $|\varphi'(3)| = \frac{3}{2} > 1$ will not converge

8. *(6 pts)* Find the quadratic (degree two) Taylor polynomial for $f(x) = x^{1/3}$ using basepoint $a = 8$.

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

$$= 2 + \frac{1}{12}(x-8) - \frac{1}{288}(x-8)^2$$

Scratch: $f(x) = x^{1/3}$

$f'(x) = \frac{1}{3}x^{-2/3}$

$f''(x) = \frac{-2}{9}x^{-5/3}$

$8^{1/3} = 2$

$\frac{1}{3}\cdot 8^{-2/3} = \frac{1}{3}\left(8^{1/3}\right)^{-2} = \frac{1}{3}\,\frac{1}{2^2} = \frac{1}{12}$

$\frac{-2}{9}\cdot(2)^{-5} = \frac{-2}{9\cdot 32} = \frac{-2}{288}$

Extra Credit. *(3 pts)* Consider the Newton iteration to solve $f(x) = 0$. This iteration can be regarded as a fixed-point iteration. Do this, and then show using Theorem 4.5.1 that it is a fast-converging fixed point iteration. Clearly state what assumptions are needed so that it converges fast.

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad \text{let } x_* \text{ be root } (f(x_*) = 0)$$

so

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

so

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

$$= \frac{f(x)f''(x)}{f'(x)^2}$$

$|\varphi'(x_*)| \ll 1$

if $f'(x_*) \neq 0$ then $\varphi'(x_*) = 0$ so $x_{k+1} = \varphi(x_k)$ converges fast