Name: SOLUTIONS

Math 310 Numerical Analysis (Bueler) December 2019

SAMPLE Final Exam

In class. No book or electronics. 1/2 sheet of notes allowed. 120 minutes maximum.

1. Write a MATLAB code for the Newton method applied to the problem $f(x) = 0$:

`function x = newton(f,dfdx,x0)`

The inputs are $f =$ `f`, the derivative $f' =$ `dfdx`, and an initial estimate $x_0 =$ `x0`. Stop the algorithm (*show this in the code!*) when $|f(x)| \le 10^{-6}$.

```
function x = newton(f, dfdx, x0)
x = x0;
for n=1:100
    if abs(f(x)) < 1e-6
        break
    end
    x = x - f(x)/dfdx(x);
end
```

2. (a) State the polynomial interpolation error theorem (with a remainder term). *Carefully state the hypotheses* and *the conclusion of the theorem.*

Thm If $f \in C^{n+1}[a,b]$ and $x_0 < x_1 < \dots < x_n$ are points in $[a,b]$, and if $p(x)$ is the unique degree n polynomial such that $p(x_i) = f(x_i)$ for all i, and if x is in $[a,b]$ then

$$f(x) = p(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n)$$

for some ξ in $[a,b]$.

(b) Assume the interval in question is $[-1,1]$ and that the interpolation points are the Chebyshev points $x_j = \cos(\pi j/n)$ for $n = 0,1,2,\dots,n$. What can you say about the remainder term that explains why the Chebyshev points are effective for interpolation? *Answer in a couple of complete sentences.*

The product $(x-x_0)\cdots(x-x_n)$ is small. In particular

$$|(x-x_0)\cdots(x-x_n)| \le \frac{1}{2^{n-1}}.$$

Thus $|f(x)-p(x)|$ is small if $|f^{(n+1)}|$ is not too big.

3. (a) Consider

$$f(x) = \frac{1}{x+2}.$$

Completely set up, but do not solve, the Vandermonde system to find the degree 3 polynomial $p(x)$ which interpolates $f(x)$ at the points $x_0 = -1.5, x_1 = -1, x_2 = 0, x_3 = 1$.

$p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$

$$\begin{bmatrix} 1 & (-1.5) & (-1.5)^2 & (-1.5)^3 \\ 1 & (-1) & (-1)^2 & (-1)^3 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1/2 \\ 1/3 \end{bmatrix}$$

(b) For the same $f(x)$ and interpolation points as in part **(a)**, write down Lagrange's form of the polynomial $p(x)$. Do not simplify.

$$p(x) = 2\,\frac{(x+1)(x)(x-1)}{(-1.5+1)(-1.5)(-1.5-1)} + 1\,\frac{(x+1.5)(x)(x-1)}{(-1+1.5)(-1)(-1-1)}$$

$$+ \frac{1}{2}\,\frac{(x+1.5)(x+1)(x-1)}{(1.5)(1)(-1)} + \frac{1}{3}\,\frac{(x+1.5)(x+1)(x)}{(1+1.5)(1+1)(1)}$$

4. Table 10.3 includes the error formula for Simpson's rule: if $f \in C^4[a,b]$ then

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] + \frac{1}{2880}(b-a)^5 f^{(4)}(\xi)$$

for some $\xi \in [a,b]$. Why does this fact show that Simpson's rule is exact if $f(x)$ is a cubic polynomial? *Answer in at least one complete sentence.*

If $f(x)$ is a cubic polynomial then

$f^{(4)}(x) = 0$ so

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

That is, the rule is exact.

5. (a) Find A_0 and A_1 so that the numerical integration rule

$$\int_{-1}^{1} f(x)\,dx \approx A_0 f(-\tfrac{1}{2}) + A_1 f(+\tfrac{1}{2})$$

is exact for all degree at most one polynomials. (I.e. for all linear functions.)

$2 = \int_{-1}^{1} x^0\,dx = A_0 \cdot 1 + A_1 \cdot 1 = A_0 + A_1$

$0 = \int_{-1}^{1} x^1\,dx = A_0(-\frac{1}{2}) + A_1 \cdot (\frac{1}{2}) = \frac{A_1 - A_0}{2}$

So $A_0 = A_1$ so $2 = 2A_0$ so

$A_0 = A_1 = 1$

(b) Show that the rule generated in part **(a)** is *not* exact for degree two polynomials.

$\frac{2}{3} = \int_{-1}^{1} x^2\,dx \stackrel{?}{=} 1 \cdot (-\frac{1}{2})^2 + 1 \cdot (\frac{1}{2})^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

not true!

6. Recall that if $\ell(x)$ is the piecewise-linear interpolant of $f \in C^2[a,b]$ at equally-spaced points $x_0 = a < x_1 < x_2 < \cdots < x_n = b$, with spacing $h = (b-a)/n$, then

$$|f(x) - \ell(x)| \le \frac{Mh^2}{8}$$

for all $x \in [a,b]$, where $M = \max_{x\in[a,b]} |f''(x)|$. Find n so that the error is at most 2×10^{-4} in using such equally-spaced linear interpolation for $f(x) = e^{-x}$ on $[a,b] = [0,2]$.

Want $|f(x) - \ell(x)| \le \frac{Mh^2}{8} \le 2 \times 10^{-4}$

where $f'' = +e^{-x}$ so $M = \max_{0 \le x \le 2} |e^{-x}| = 1$. So we want

$\frac{h^2}{8} \le 2 \times 10^{-4} \iff h \le \sqrt{16 \times 10^{-4}} = 4 \times 10^{-2} \iff \frac{2-0}{n} \le 4 \times 10^{-2}$

$\iff \frac{n}{2} \ge \frac{1}{4 \times 10^{-2}} = \frac{1}{4} \times 10^2 \iff n \ge \frac{1}{2} \times 10^2 = 50$

7. Do two steps of the Euler method, with step size $h = 1$, on the ODE IVP

$$y' = t - y, \qquad y(0) = 1.$$

$t_0 = 0, t_1 = 1, t_2 = 2$

$$y_{k+1} = y_k + h f(t_k, y_k) = y_k + h(t_k - y_k)$$

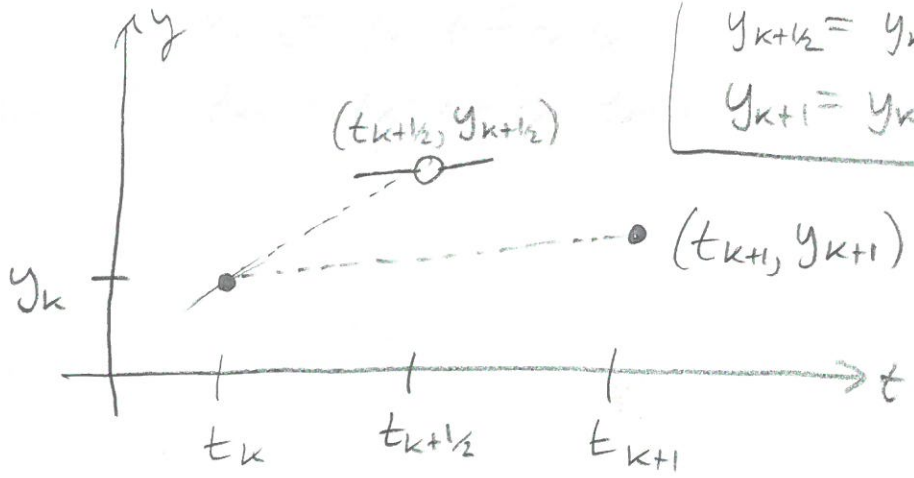
$k=0$: $y_1 = 1 + 1\cdot(0-1) = 1-1 = 0 \approx y(1)$

$k=1$: $y_2 = 0 + 1\cdot(1-0) = 1 \approx y(2)$

8. **(a)** Sketch one step of the midpoint method for the general ODE IVP

$$y' = f(t, y), \qquad y(t_0) = y_0$$

where $t_{k+1} - t_k = h$ is the step size. (Hints: *Your sketch will have t and y axes. Show the current iterate (t_k, y_k) and all the locations where a slope is computed. Show how to compute the new iterate y_{k+1}.*)



$$y_{k+1/2} = y_k + \tfrac{h}{2} f(t_k, y_k)$$
$$y_{k+1} = y_k + h f(t_{k+1/2}, y_{k+1/2})$$

(b) Show that the midpoint method is exact when solving the ODE IVP

$$y' = 2t - 8, \qquad y(2) = 3.$$

for any h,

$$y_1 = 3 + h\left(2(2+\tfrac{h}{2}) - 8\right) = 3 + h(4+h-8) = 3 - 4h + h^2$$

but

$$y(2+h) = y(2) + \int_2^{2+h} 2s - 8\, ds = 3 + \left[s^2 - 8s\right]_2^{2+h}$$

$$= 3 + \left((2+h)^2 - 8(2+h) - 2^2 + 8(2)\right) = 3 - 4h + h^2$$

9. Suppose the IEEE 754 standard for floating point representations had a 9 bit version:

s	e_1	e_2	e_3	b_1	b_2	b_3	b_4	b_5

representing the number

$$x = (-1)^s \, (1.b_1b_2b_3b_4b_5)_2 \; 2^{(e_1e_2e_3)_2 - 3_{10}}$$

Note the exception cases:

- exponent bits $(000)_2$ define the number zero or subnormal numbers
- exponent bits $(111)_2$ define the other exceptions: $\pm\infty$ and NaN (*... ignore the details*)

(a) What is the largest real number that this system can represent? (*State the number in decimal notation and show the bits.*)

0	1	1	0	1	1	1	1	1

$= +(1.11111)_2 \times 2^{6-3}$

$= (2 - \frac{1}{32}) \times 2^3 = 2^4 - 2^{-2} = 16 - \frac{1}{4} = 15\frac{3}{4}$

(b) What is the value of "machine epsilon" in this system? (*State the number in decimal notation.*)

$\varepsilon = (1.00001)_2 - (1.00000)_2 = (.00001)_2 = \frac{1}{32}$

10. Suppose we want to use Taylor's theorem to compute values of $\sin x$ for $|x| < 0.5$ to an accuracy of 10^{-3}. Use the Taylor theorem with remainder to determine how many terms, i.e. what n, is needed to do this.

$f(x) = \sin(x) \;\therefore\; f^{(n+1)}(x) = \pm \begin{matrix}\sin \\ \cos\end{matrix} x \;\therefore\; |f^{(n+1)}(\xi)| \le 1$

$f(x) = p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-0)^{n+1}$

↑ use base point $x_0 = 0$

$|f(x) - p_n(x)| \le |\ldots| \le \frac{1}{(n+1)!}(0.5)^{n+1} < 10^{-3}$

$n=1$: $\frac{0.5}{2} = \frac{1}{4}$

$n=2$: $\frac{(0.5)^2}{6} < 10^{-3}$

$\vdots$

$n=4$: $\frac{(\frac{1}{2})^5}{5!} = \frac{1}{32} < 10^{-3}$

$n = 4$

11. Solve the following system of linear equations by Gauss elimination with partial pivoting and back substitution. Show your steps.

$$2x_1 + 2x_2 = 6$$
$$4x_1 - 3x_2 = -2.$$

$R_1 \leftrightarrow R_2$: $4x_1 - 3x_2 = -2$
$2x_1 + 2x_2 = 6$

$R_2 \leftarrow R_2 - \frac{1}{2}R_1$: $4x_1 - 3x_2 = -2$
$\frac{7}{2}x_2 = 7$

back sub: $x_2 = \frac{7}{7/2} = 2$

$x_1 = \frac{-2+3\cdot 2}{4} = 1$

12. The high-level view of the Gauss elimination with partial pivoting algorithm is that, given a linear system

$$A\mathbf{x} = \mathbf{b},$$

it computes matrices P, L, U so that $PA = LU$. What properties do these matrices have? (*Write at least two complete sentences.*) Then explain how to solve the linear system, indicating how much work is required at each stage. (*Write at least two complete sentences.*)

The P matrix is a (row) permutating L is lower triangular with one on the diagonal, and U is upper triangular. To solve $Ax=b$ first multiply both sides by P and then break into two systems:

$$PAx = Pb \iff LUx = Pb \iff \begin{cases} Ly = Pb & (1) \\ Ux = y & (2) \end{cases}$$

Solve (1) by forward substitution and (2) by back substitution

13. (a) Write a MATLAB algorithm for multiplying a square $n \times n$ matrix A by an $n \times 1$ column vector $\mathbf{v}$. In particular, fill in the rest of the function below to compute

$$\mathbf{z} = A\mathbf{v}.$$

I have written the first line to get n. You may assume all sizes of the inputs are correct; there is no need to check these sizes. Do not use matrix-vector multiplication inside this routine; pretend that we are writing this for the first time and use `for` loops.

```
function z = mattimesvec(A,v)
% MATTIMESVEC multiplies A by v and gives z

n = length(v);
z = zeros(n,1);
for k=1:n
    for j=1:n
        z(k) = z(k) + A(k,j)*v(j);
    end
end
```

(b) Count the floating point operations in the above algorithm.

n^2 multiplications and n^2 additions

for $2n^2$ total operations

TABLE 10.3
Quadrature formulas and their errors.

Method	Approximation to $\int_a^b f(x)\,dx$	Error
Trapezoid rule	$\frac{b-a}{2}[f(a)+f(b)]$	$-\frac{1}{12}(b-a)^3 f''(\eta),\ \eta \in [a,b]$
Simpson's rule	$\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$	$\frac{1}{2880}(b-a)^5 f^{(4)}(\xi),\ \xi \in [a,b]$
Composite trapezoid rule	$\frac{h}{2}[f_0+2f_1+\ldots+2f_{n-1}+f_n]$	$O(h^2)$
Composite Simpson's rule	$\frac{h}{6}[f_0+4f_{1/2}+2f_1+\ldots+2f_{n-1}+4f_{n-1/2}+f_n]$	$O(h^4)$

[BLANK SPACE FOR SCRATCH WORK]