MATH 2130 Midterm Test 1 — Solutions

Prof. A. Willms, Dept. of Mathematics and Statistics, University of Guelph

Jan. 30, 2019

SECTION A.
ANSWER QUESTIONS 1–6 IN THE SPACE PROVIDED.

10 pts. 1. If v is a vector, the MATLAB command `norm(v)` will compute the Euclidean norm of v, that is, the length of v. Now suppose you have a matrix A in MATLAB. Write MATLAB code using a "for" loop that will compute the norm of each row of A. These computed norms should be stored in a vector L. (You will need to use a MATLAB command to determine the number of rows in A.)

Solution:
The following MATLAB code will do the requested task.

```
n = size(A,1);
for i=1:n
    L(i) = norm(A(i,:));
end
```

5 pts. 2. Suppose you had a computer using a floating point system with a small amount of precision. What formulas would you use to most accurately compute the roots of $2x^2 + 240x - 1 = 0$ on this computer? (You need not evaluate the formulas.)

Solution:
Since the quadratic has a positive coefficient for the linear term in x, the most accurate way to compute the two roots are to use the formulas:

$$x_1 = \frac{-240 - \sqrt{240^2 - 4(2)(-1)}}{2(2)} = \frac{-240 - \sqrt{240^2 + 8}}{4}$$

and

$$x_2 = \frac{-2(-1)}{240 + \sqrt{240^2 - 4(2)(-1)}} = \frac{2}{240 + \sqrt{240^2 + 8}}.$$

Students may also give the answer in terms of formulas as:

Since $b = 240 > 0$, the roots are accurately given by

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

and

$$x_2 = \frac{-2c}{b + \sqrt{b^2 - 4ac}},$$

where $a = 2$, $b = 240$, and $c = -1$.

5 pts. 3. A typical floating point number system uses a string of bits (each with value 0 or 1) of the form

$$s\ c_1\ c_2\ \cdots\ c_m\ f_1\ f_2\ \cdots\ f_p$$

to represent a number. Here s is the sign bit, the c_i are the exponent bits, and the f_i are the mantissa (or fraction) bits. What number does that string of bits represent? (You can assume it is a "normal" floating point number, that is, not one of the special cases.)

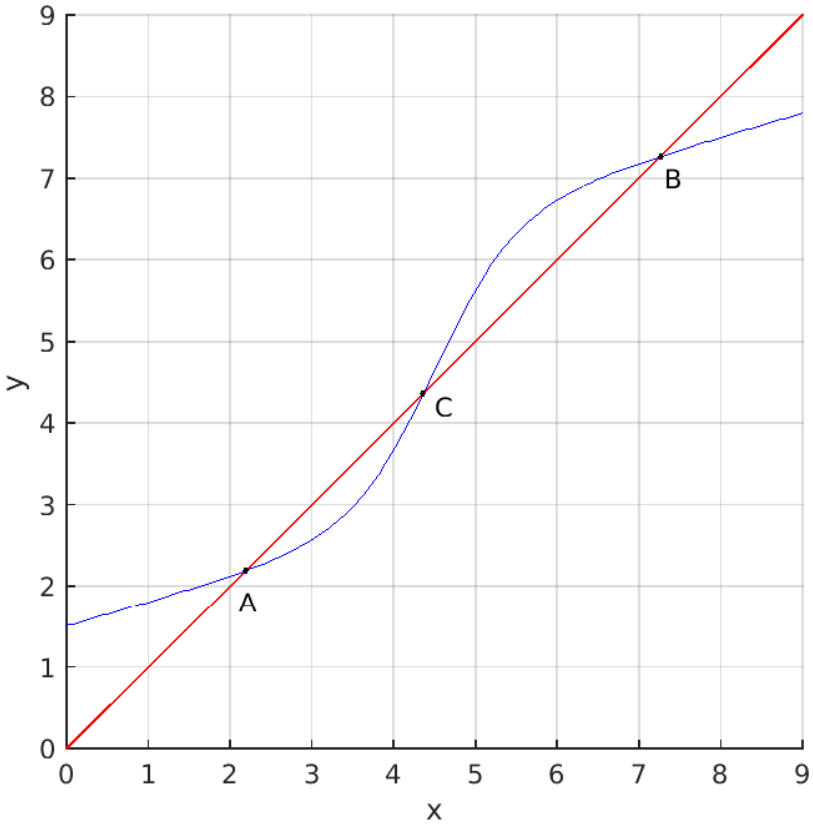
Solution:
This string of bits represents the number

$$(-1)^s (1.f_1\, f_2\, f_3\, \cdots\, f_p)_2 \times 2^{c-\left(2^{m-1}-1\right)},$$

where c is the binary unsigned integer defined by the c_i bits, that is, $c = (c_1\, c_2\, \cdots\, c_m)_2$.

10 pts. 4. Below is a sketch of a function $y = g(x)$ as well as the line $y = x$. Consider the fixed point iteration defined by

$$x_{n+1} = g(x_n).$$



(a) What is the largest interval $[a, b]$ where a and b are both integers between 0 and 9, such that the Fixed Point Interation Theorem will guarantee that the iteration will converge to the fixed point A provided the initial point x_0 is within $[a, b]$?

(b) Repeat (a) for the fixed point B.

(c) Is it possible for the iteration to converge to the fixed point C. Why or why not?

Solution:

(a) The largest such interval is $[0, 3]$.

(b) The largest such interval is $[6, 9]$.

(c) No it is not possible for the iteration to converge to C, because $g'(C) > 1$. Any point starting near $x = C$ will move away from $x = C$ on the next iteration.

10 pts. 5. Below is MATLAB code intended to implement the bisection method for finding a root of the equation $f(x) = 0$. There are six errors. Circle each error and, in the space below, write the corrections to the lines with errors.

```
function x = bisection(f,I,TOLX,TOLF)
% bisection - find a root by the bisection method
%
%  x = bisection(f,I,TOLX,TOLF) will find the root of f(x)=0 in the
%      interval I such that the relative error of the solution is less
```

```
%      than TOLX or the function value is within TOLF of zero
%
%  Input:
%    f : handle to a function that takes x and returns f(x)
%    I : length 2 vector defining the initial interval containing the root
%    TOLX : convergence criterion for x
%    TOLF : convergence criterion for f(x)
%
%  OUTPUT:
%    x : estimate solution to f(x) = 0
%
f1 = f(I(1));
f2 = f(I(2));
if f1 == 0
    x = I(1);
    return;
elseif f2 == 0
    x = I(2);
    return;
elseif f1*f2 < 0
    error('f evaluated at the end points of I does not alternate sign')
end
x = max(I);
fmid = f(x);
while abs(I(2) - x) > TOLX && abs(fmid) > TOLF
    test = fmid*f1;
    if test > 0
        f2 = fmid;
        I(1) = x;
    elseif test < 0
        I(2) = x;
    end
    x = mean(I);
    fmid = f1;
else
end  % end of function
```

Solution:
The errors were:

- last clause of "if/elseif" statement should have a $>$ sign':
  `elseif f1*f2 > 0`
- two lines prior to while loop, should be:
  `x = mean(I);`
- first part of while loop test should be a relative error, not absolute:
  `abs(I(2) - x) > TOLX*abs(x)`
- inside "if" clause of loop, `f2` should be `f1`: `f1 = fmid;`
- last line of while loop, `f1` should be `f(x)`: `fmid = f(x);`
- last line: `else` should be `end`: `end`

10 pts. 6. Show that Newton's method converges with order two.

Solution:
Newton's method is:

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

Expanding f around x_k by Taylor's formula gives

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{f''(\eta)}{2}(x - x_k)^2,$$

where η is some value between x and x_k. Assuming Newton's method converges to x^*, setting $x = x^*$ in the above equation yields

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\eta)}{2}(x^* - x_k)^2,$$

where we have used the fact that $f(x^*) = 0$. Dividing the above equation by $f'(x_k)$ and rearranging gives

$$-\frac{f(x_k)}{f'(x_k)} = (x^* - x_k) + \frac{f''(\eta)}{2f'(x_k)}(x^* - x_k)^2.$$

Using the above to substitute into (1) gives

$$x_{k+1} = x_k + (x^* - x_k) + \frac{f''(\eta)}{2f'(x_k)}(x^* - x_k)^2,$$

which implies

$$|x_{k+1} - x^*| = \left|\frac{f''(\eta)}{2f'(x_k)}\right| |x^* - x_k|^2,$$

and

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \lim_{k\to\infty} \left|\frac{f''(\eta)}{2f'(x_k)}\right| = \left|\frac{f''(x^*)}{2f'(x^*)}\right| = \lambda.$$

Since λ is finite (provided $f'(x^*) \neq 0$) we conclude that the order of convergence of Newton's method is (at least) two. (It will be larger than two in the case that $f''(x^*) = 0$.)

SECTION B.

TEAR OFF THIS PAGE. ANSWER THESE QUESTIONS USING THE BUBBLE SHEET.

1 pt. 1. What is the primary source of error when using a computer floating point system to do calculations with real numbers?

(a) Overflow.
(b) Underflow.
(c) Cancellation error.
(d) Round off error.
(e) Truncation error.

Solution:
The correct answer is (d).

1 pt. 2. What is the distance between the number 1.0 and the next (larger) floating point number in the IEEE double precision floating point system?

(a) 2^{-3}
(b) 2^{-11}
(c) 2^{-23}
(d) 2^{-52}
(e) 2^{-53}

Solution:
The correct answer is (d).

1 pt. 3. Let B, S, M, and N be the orders of convergence for the bisection method, the secant method, Muller's method, and Newton's method, respectively. Which of the following is true?

(a) $S \leq B < M < N$

(b) $M \leq B < N \leq S$

(c) $B < S < N \leq M$

(d) $B < S < M < N$

(e) $M \leq B < S < N$

Solution:
The correct answer is (d).

1 pt. 4. These methods for finding a zero of $f(x)$ employ a strategy whereby the zero is guaranteed to be bounded in a given interval.

(a) Bisection method and Muller's method

(b) Muller's method and Secant method

(c) Bisection method and Regula Falsi method

(d) Secant method and Bisection method

(e) Regula Falsi method and Muller's method

Solution:
The correct answer is (c).

1 pt. 5. Suppose you are using Newton's method to solve $f(x) = 0$ and suppose you knew (you wouldn't but suppose you did) that the estimates from the $(k-1)$st and kth iterations, x_{k-1} and x_k, were within 10^{-1} and 5×10^{-3} of the true solution, respectively. How close would you expect x_{k+1} to be from the true solution?

(a) 5×10^{-6}

(b) 1.25×10^{-6}

(c) 5×10^{-5}

(d) 1.25×10^{-5}

(e) 2.5×10^{-5}

Solution:
The correct answer is (d).

1 pt. 6. Muller's method for solving the scalar equation $f(x) = 0$ generates a sequence of estimates x_1, x_2, $x_3, \ldots, x_n$ of the solution. It uses the last three points, x_{n-2}, x_{n-1}, and x_n, and fits a quadratic polynomial through these points. How is the estimate x_{n+1} chosen?

(a) it is the midway point between x_{n-1} and x_n

(b) it is the zero of the quadratic that is between x_{n-2} and x_n

(c) it is the location where the quadratic has zero derivative

(d) it is the point where the quadratic intersects the line through $(x_{n-1}, f(x_{n-1}))$ and $(x_n, f(x_n))$

(e) it is the zero of the quadratic closest to x_n

Solution:
The correct answer is (e).

DID YOU ANSWER THE QUESTIONS IN SECTION B USING THE BUBBLE SHEET? IF NOT, YOU WILL GET ZERO FOR THEM.