MATH3230A Numerical Analysis (2018/19, First term)
Mid-term Examination 10th Oct 2018

- Time allowed: 2 hours – 4:30pm to 6:30 pm
- Please answer **all** questions and write down all detailed steps.
- Total mark is 55.

1. (a) State the definition of the following concepts:

   i. (2 marks) $x_k \to x^*$ Q-linearly with rate $\rho \in (0, 1)$.
   (Soln.)
   $$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \rho.$$

   ii. (2 marks) $x_k \to x^*$ R-sublinearly.
   (Soln.) If there exists a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ such that $\varepsilon_k \to 0$ Q-sublinearly, and for every $k \in \mathbb{N}$,
   $$|x_k - x^*| \le \varepsilon_k.$$

   iii. (2 marks) x_k Q-converges to x^* with order $p > 1$.
   (Soln.) There exists a positive constant ρ such that
   $$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \rho.$$

   iv. (2 marks) The absolute and relative error between approximation x_k and true value x^*.
   (Soln.) Absolute error is $|x_k - x^*|$ and relative error is $|x_k - x^*|/|x^*|$.

   (b) For the following sequences, compute their limit and the type of convergence (Q-linear, Q-superlinear, Q-sublinear, R-linear, R-superlinear, R-sublinear).

   i. (3 marks) $a_{2k} = 2^{-k}$, $a_{2k+1} = (1 + 2^k)^{-1}$.
   (Soln.) The limit is zero. $\frac{a_{k+1}}{a_k}$ does not have a limit, but $|a_{2k}| \le 2^{-k}$ and $|a_{2k+1}| \le 2^{-k}$ and so $|a_n| \le 2^{-n/2}$. The sequence $2^{-n/2}$ converges Q-linearly with rate $\rho = 1/\sqrt{2}$, and so a_k converges to zero R-linearly.

   ii. (3 marks) $a_n = \frac{n+4}{(2n+1)^3}$.
   (Soln.) The limit is zero and $\frac{a_{n+1}}{a_n} \to 1$ so the sequence converges Q-sublinearly.

   iii. (3 marks) $a_n = n^{-n}$.
   (Soln.) The limit is zero and $\frac{a_{n+1}}{a_n} = \frac{1}{n}\left(1 - \frac{1}{n+1}\right)^{n+1} \to 0$ the sequence converges Q-superlinearly.

2. Let $f : [a, b] \to \mathbb{R}$ be a continuous function such that $f(a) < 0$ and $f(b) > 0$.

   (a) (4 marks) Write down the Bisection algorithm for solving the nonlinear equation $f(x) = 0$.
   (Soln.) Set a stopping parameter δ, $a_0 = a$, $b_0 = b$, and $k = 0$. While $|b_k - a_k| > \delta$ set $x_k = (a_k + b_k)/2$ and do the following: if $f(x_k)f(a_k) > 0$ then set $a_{k+1} = x_k$ and $b_{k+1} = b_k$, otherwise set $a_{k+1} = a_k$ and $b_{k+1} = x_k$.

   (b) (3 marks) Let $a_0 = a$, $b_0 = b$, and $[a_1, b_1], [a_2, b_2], \ldots, [a_n, b_n], \ldots$ denote the successive intervals generated by the Bisection algorithm. Let $x_n = \frac{1}{2}(a_n + b_n)$ denote the midpoint of the interval $[a_n, b_n]$. Show that

$$|x_{n+1} - x_n| = 2^{-n-2}(b_0 - a_0).$$

   (Soln.) By definition x_{n+1} is the midpoint of $[a_n, x_n]$ or $[x_n, b_n]$, and so $|x_{n+1} - x_n| = \frac{1}{4}(b_n - a_n)$. Then, using that $b_n - a_n = 2^{-n}(b_0 - a_0)$ shows the result.

   (c) (3 marks) Consider the bisection algorithm applied to the starting interval $[2.5, 6.5] = [a_0, b_0]$. State (i) the width of the nth interval $[a_n, b_n]$, (ii) the maximum possible distance between x^* and the midpoint $x_n = \frac{1}{2}(a_n + b_n)$, and (iii) the number of iterations needed for the error $|x_n - x^*|$ to be less than 10^{-10} (an expression involving logarithms is acceptable).
   (Soln.) The width of the interval $[a_n, b_n]$ is $2^{-n}(b_0 - a_0) = 2^{-n+2}$. Using the error estimate for the Bisection algorithm, we have

$$|x_n - x^*| \le 2^{-(n+1)}(b_0 - a_0) = 2^{-n+1}.$$

   For $2^{-n+1} < 10^{-10}$ we have $n > 10\frac{\log 10}{\log 2} + 1$.

   (d) (3 marks) Let $f(x) = 2x^2 + 3x - 4$. Starting from the interval $[0, 1]$ write down the first three values x_1, x_2, x_3 generated by the Bisection algorithm.
   (Soln.) $x_1 = 0.5$, $x_2 = 0.75$, $x_3 = 0.875$.

3. (a) (2 mark) State the difference between Newton's method and Quasi-Newton method. Give one example of a Quasi-Newton method for solving the nonlinear equation $f(x) = 0$.
   (Soln.) In a Quasi-Newton's method, the derivative term $f'(x_k)$ is replaced by an approximation g_k that does not involve evaluating the derivative $f'(x)$. One example is the constant slope method $g_k = g$ or the Secant method $g_k = (f(x_k) - f(x_{k-1}))/(x_k - x_{k-1})$.

   (b) (3 marks) Let $f(x) = (x - x^*)^2 g(x)$ where $g(x^*) \neq 0$. Use the fixed-point iterative function $\varphi(x) = x - \frac{f(x)}{f'(x)}$ to show that Newton's method converges linearly with rate $\rho = \frac{1}{2}$.
   (Soln.) Evaluating $\varphi'(x^*)$ yields $\varphi'(x^*) = \frac{1}{2}$.

   (c) (3 marks) Consider the modified Newton's method

$$x_{k+1} = x_k - 2\frac{f(x_k)}{f'(x_k)}$$

   to solve $f(x) = (x - x^*)^2 g(x) = 0$ with $g(x^*) \neq 0$. For the fixed point iterative function $\varphi(x) = x - 2\frac{f(x)}{f'(x)}$ show that

$$\varphi'(x^*) = 0.$$

What can you deduce without further calculations about the order of convergence for the modified Newton's method?
(Soln.) The convergence is at least quadratic. Do not accept (superlinear).

(d) (3 marks) Let $\{e_k\}_{k\in\mathbb{N}}$ be a sequence of errors such that $e_k \to 0$ as $k \to \infty$. Furthermore, suppose there exists a constant $p > 0$ such that

$$e_{k+1} = e_k^p \text{ and } e_{k+2} = e_{k+1}e_k^2 \text{ for all } k \in \mathbb{N}.$$

Compute the value of p.
(Soln.) By the relations we have

$$(e_k)^{p^2} = e_{k+1}^p = e_{k+2} = e_{k+1}e_k^2 = e_k^{p+2}.$$

So $p^2 - p - 2$ should be zero. Taking the positive root gives $p = 2$.

4. For $s, m_1, \ldots, m_4, f_1, \ldots, f_6 \in \{0, 1\}$, consider a 11-bit format

$$s|m_1m_2m_3m_4|f_1f_2f_3f_4f_5f_6$$

(a) (2 marks) What is the range of the exponent $m = (m_1m_2m_3m_4)_2$? Excluding the cases $m = (0000)_2$ and $m = (1111)_2$ reserved for special values, what should be the number y so that

$$a = (-1)^s(1.f_1f_2\ldots f_6)_2 \times 2^{(m_1m_2m_3m_4)_2 - y}$$

has an equal number of choices for both positive and negative exponents?
(Soln.) The range of m is from $0 = (0000)_2$ to $15 = (1111)_2$, which yields 16 possibilities. Excluding the two endpoints we are left with 14 possibilities and so $y = 7$.

(b) (4 marks) What are the smallest normalized positive value and largest normalized finite value in this 11-bit format? [Recall the cases $m = (0000)_2$ and $m = (1111)_2$ are reserved for special values.]
(Soln.) The smallest positive value in this format is

$$0|0001|000000 \quad \mapsto \quad a_{min} = (1.000000)_2 \times 2^{(0001)_2 - 7} = 2^{-6}.$$

The largest finite value is

$$\begin{aligned} 0|1110|111111 \quad \mapsto \quad a_{max} &= (1.111111)_2 \times 2^{(1110)_2 - 7} \\ &= (2 - 2^{-6}) \times 2^{14-7} = (2 - 2^{-6}) \times 2^7 = 2^8 - 2. \end{aligned}$$

(c) (2 marks) What is the machine epsilon associated to this format?
(Soln.) $\varepsilon_M = 2^{-6}$ as we used 6-bits for the mantissa.

5. State whether the following statements are true or false, no justification is needed.

(a) (1 mark) Let $\{x_n\}_{n\in\mathbb{N}}$ and $\{y_n\}_{n\in\mathbb{N}}$ be two sequences that converge Q-linearly to x^* with rate $\rho_x, \rho_y \in (0, 1)$, respectively. If $\rho_x > \rho_y$, then $\{y_n\}_{n\in\mathbb{N}}$ converges faster than $\{x_n\}_{n\in\mathbb{N}}$.
(Soln.) True.

(b) (1 mark) Let $A \in \mathbb{R}^{2\times 2}$ be a matrix with nonnegative entries. If $\|A\|_1 = \|A\|_\infty$, then either $a_{12} = a_{21}$ or $a_{11} = a_{22}$.
(Soln.) True, since $\|A\|_1 = \max(a_{11} + a_{21}, a_{12} + a_{11})$ and $\|A\|_\infty = \max(a_{11} + a_{12}, a_{21} + a_{11})$, and a case analysis shows that $a_{12} = a_{21}$ and $a_{11} = a_{22}$.

(c) (1 mark) If a matrix is simultaneously lower triangular and upper triangular, then it must be a diagonal matrix.
(Soln.) True.

(d) (1 mark) For an invertible matrix A, the condition numbers $\kappa(A)$ and $\kappa(A^{-1})$ are the same.
(Soln.) True.

(e) (1 mark) Loss of significant will not occur when using the formula

$$x = \frac{1}{2a}\Big(- b + \sqrt{b^2 - 4ac}\Big)$$

to evaluate one root of the polynomial $ax^2 + bx + c = 0$ with $a = 1$, $b = 1$, $c = 10^{-20}$ with 5 digits of accuracy.
(Soln.) False, since we are subtracting $\sqrt{b^2 - 4ac} \approx b$ with b itself.

(f) (1 mark) Given a matrix $A \in \mathbb{R}^{n\times n}$, if $\max_{1\le i\le n} |\lambda_i| = 0$ where $\lambda_1, \dots, \lambda_n$ are its eigenvalues, then A must be the zero matrix.
(Soln.) False, take the matrix with entries 0,1,0,0.