AMSC/CMSC 460 Midterm Exam 1 – Solutions

Tuesday, Feb 27th, 2018

You have 75 minutes to complete this exam. **No** Calculators or cheat sheets are allowed. Submit each problem on a separate sheet. Show all work and explain your answers.

---

1. Recall a single precision floating point number x can be written as $x = (.1d_2 \dots d_{24})_2 2^e$ where $-125 \leq e \leq 128$. What is the smallest possible single precision floating point number that is greater than 32? You may write your answer as sums of powers of 2.

Solution: In normalized form we can write $32 = (.10\dots0)2^6$. Therefore the next floating point number bigger than 32 is

$$(.10\dots01)2^6 = (2^{-1} + 2^{-24})2^6 = 2^5 + 2^{-18} \approx 32.0000038146973$$

---

2. Consider the function

$$f(x) = \frac{1 - \cos x}{x^2}.$$

It is easy to check that $\lim_{x\to 0} f(x) = 1/2$. However MATLAB will claim that $f(x) = 0$ for $|x|$ any smaller than 10^{-8}. Explain why this is the case (why specifically 10^{-8}?). Hint: Use the fact that $\epsilon_m \approx 10^{-16}$ and $\cos(x) = 1 - \frac{1}{2}x^2 + \mathcal{O}(x^4)$.

Solution: The reason that MATLAB claims $f(x) = 0$ for $0 < |x| < 10^{-8}$ is because of cancellation in the numerator and the associated *loss of precision* of $1 - \cos x$. It is **not** due to underflow, which only occurs for numbers below $2.2250738585072 * 10^{-308}$. By the definition of the machine epsilon ϵ_m MATLAB rounds off any precision below the machine precision. Therefore $\epsilon_n/2$ is negligable *relative* to 1 and

$$\text{float}(1 + 2^{-1}\epsilon_m) = 1.$$

Using the Taylor series for $\cos x$ this means that for small x, $\cos(x) \approx 1 - \frac{1}{2}x^2$. Therefore if $|x| < \sqrt{\epsilon_m} \approx 10^{-8}$ then

$$\text{float}(\cos(x)) = \text{float}(1 - 2^{-1}\epsilon_m) = 1.$$

It follows that for $0 < |x| < 10^{-8}$ MATLAB will compute

$$\hat{f}(x) = \frac{\text{float}(1) - \text{float}(1 - 2^{-1}\epsilon_m)}{\text{float}(x)^2} = \frac{1-1}{x^2} = 0.$$

---

3. Consider the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

(a) Find the Cholesky factorization of $\mathbf{A}$.

(b) Using the fact that $\mathbf{A}$ has a Cholesky factorization, show that $\mathbf{A}$ is a positive definite matrix.

Solution: (a) To find the Cholesky factorization, we seek coefficients $a_{11}, a_{12}, a_{13}, a_{22}, a_{23}a_{33}$ such that

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{12} & a_{22} & 0 \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11}^2 & a_{11}a_{12} & a_{11}a_{13} \\ a_{11}a_{12} & a_{12}^2 + a_{22}^2 & a_{12}a_{13} + a_{22}a_{23} \\ a_{11}a_{13} & a_{12}a_{13} + a_{22}a_{23} & a_{13}^2 + a_{23}^2 + a_{33}^2. \end{bmatrix}$$

This requires us to solve the equations

$$\begin{aligned} a_{11}^2 &= 2 & a_{11}a_{12} &= -1 & a_{11}a_{13} &= 0 \\ a_{12}^2 + a_{22}^2 &= 2 & a_{12}a_{13} + a_{22}a_{23} &= -1 & a_{13}^2 + a_{23}^2 + a_{33}^2 &= 2 \end{aligned}$$

This gives

$$\begin{aligned} a_{11} &= \sqrt{2}, \quad a_{12} = -\frac{1}{\sqrt{2}} \quad a_{13} = 0 \\ a_{22} &= \sqrt{2 - a_{12}^2} = \sqrt{\frac{3}{2}}, \quad a_{23} = -1/a_{22} = -\sqrt{\frac{2}{3}} \\ a_{33} &= \sqrt{2 - a_{13}^2 - a_{23}^2} = \frac{2}{\sqrt{3}} \end{aligned}$$

Therefore we have found a Cholesky factorization $\mathbf{A} = \mathbf{U}^\top \mathbf{U}$, where $\mathbf{U}$ is given by

$$\mathbf{U} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & \sqrt{\frac{3}{2}} & -\sqrt{\frac{2}{3}} \\ 0 & 0 & \frac{2}{\sqrt{3}} \end{bmatrix}$$

(b) To show that $\mathbf{A}$ must be positive definite. We see that for any $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{U}^\top \mathbf{U}\mathbf{x} = (\mathbf{U}\mathbf{x})^\top (\mathbf{U}\mathbf{x}) = \|\mathbf{U}\mathbf{x}\|^2$$

Positive definiteness now follows from the fact that all of the diagonals on $\mathbf{U}$ are positive and therefore $\mathbf{U}$ is non-singular, meaning

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \|\mathbf{U}\mathbf{x}\|^2 > 0, \quad \text{for} \quad \mathbf{x} \neq 0.$$

---

4. Let $\mathbf{A}$ be a square matrix and let $c \in \mathbb{R}$ be a scalar. Let $\|\mathbf{A}\|$ denote the natural matrix norm induced from a vector norm and let $\kappa(\mathbf{A})$ be the associated condition number. Prove or disprove the following statements

   (a) $\|c\mathbf{A}\| = |c| \cdot \|\mathbf{A}\|$

   (b) $\kappa(c\mathbf{A}) = |c| \cdot \kappa(\mathbf{A})$

Solution: (a) This is true. By definition of the vector norm we know that for any vector $\mathbf{x}$ with $\|\mathbf{x}\| = 1$

$$\|c\mathbf{A}\mathbf{x}\| = |c| \cdot \|\mathbf{A}\mathbf{x}\|$$

Taking the max of all such $\mathbf{x}$ on both sides of the above equality and using the definition of the matrix norm gives

$$\|c\mathbf{A}\| = |c| \cdot \|\mathbf{A}\|$$

(b) This is not true unless $|c| = 1$, $c = 0$ or $\kappa(\mathbf{A}) = \infty$ since, by definition and part(a)

$$\kappa(c\mathbf{A}) = \|c\mathbf{A}\| \cdot \|c^{-1}\mathbf{A}^{-1}\| = |c| \cdot |c^{-1}| \cdot \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})$$

---

5. Let f be a function on $[a, b]$ with infinitely many continuous derivatives. In the homework, you showed that if $\bar{x}$ is a double root (i.e. $f(\bar{x}) = 0$, $f'(\bar{x}) = 0$, $f''(\bar{x}) \neq 0$) then the error for Newton's method at step i, $e_i = x_i - \bar{x}$, satisfies

$$e_{i+1} = \frac{1}{2}e_i + \mathcal{O}(e_i^2).$$

What happens when $\bar{x}$ is a triple root (i.e. $f(\bar{x}) = 0$, $f'(\bar{x}) = 0$, $f''(\bar{x}) = 0$, $f'''(\bar{x}) \neq 0$)? Give a formula relating e_{i+1} and e_i to leading order in e_i. What is the order of convergence in this case?

Solution: The error at step $i + 1$ is related to the error at step i by

$$e_{i+1} = e_i - \frac{f(\bar{x} + e_i)}{f'(\bar{x} + e_i)} = e_i - \frac{\frac{1}{6}f'''(\bar{x})e_i^3 + \mathcal{O}(e_i^4)}{\frac{1}{2}f'''(\bar{x})e_i^2 + \mathcal{O}(e_i^3)} = e_i - \frac{2}{6}\frac{f'''(\bar{x})}{f'''(\bar{x})}e_i + \mathcal{O}(e_i^2) = \frac{2}{3}e_i + \mathcal{O}(e_i^2).$$

Therefore the order of convergence is still order 1 just as with the double root.

---

6. Write down the secant method and state its order of convergence.

Solution: The secant method is given by

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}.$$

It's order of convergence is the golden ratio $\phi = \frac{1+\sqrt{5}}{2}$.

---

7. Four different methods were used to used to solve $f(x) = 0$ and the computed values $x_1, x_2, \ldots$ are shown below:

i	Method 1	Methods 2	Method 3	Method 4
1	1.10000000000000	1.02000000000000	1.05000000000000	1.03162277660168
2	1.01000000000000	1.00400000000000	1.02500000000000	1.00562341325190
3	1.00010000000000	1.00080000000000	1.01250000000000	1.00042169650343
4	1.00000001000000	1.00016000000000	1.00625000000000	1.00000865964323
5	1.00000000000000	1.00003200000000	1.00312500000000	1.00000002548297
6	1.00000000000000	1.00000640000000	1.00156250000000	1.00000000000407
7	1.00000000000000	1.00000128000000	1.00078125000000	1.00000000000000
8	1.00000000000000	1.00000025600000	1.00039062500000	1.00000000000000

(a) One of them is Newton's method. Which of the four is most likely Newton's method, and why?

(b) One of them is the bisection method. Which of the four is most likely the bisection method, and why?

Solution: (a) Method 1 is mostly likely Newton because of it's quadratic convergence. Specifically the error at each step is the square of the previous. All other methods are converging sub-quadratically.

(b) Method 3 is bisection because of the fact that it converges linearly, and at each step the error is divided by a factor of 2, which is a hall-mark of the bisection method.