# Midterm Solution

Math 3604: Introduction to Computational Mathematics

November 19, 2018

**Problem 1. [20 points (4+4+4+4+4)]** Consider a floating-point system that any real number $x$ is converted to the form of $0.\delta_1 0.ddddd \times 10\hat{}(\delta_2 ee)$ . Here, $\delta_1$ and $\delta_2$ indicates the sign (0 for positive and 1 for negative) of the mantissa and exponent, respectively; $d$ and $e$ are decimal digits (0 to 9) for the mantissa and the exponent, respectively. We assume that zero can be represented by either $+0.00000$ or $0.00000$ for the mantissa and $+0.00$ or $0.00$ for the exponent.

(a) What are the largest floating-point numbers in the floating-point system?

(b) What is the distance between 0 and the next positive floating-point number in the system?

(c) What is the smallest floating point that is larger than 8 in the floating-point system?

(d) How many distinct real numbers can be represented by this floating-point system.

(e) What relative error of numerical computations you can expect from this floating-point system?

**Solution**

(a) $0.99999 \times 10^{99}$

(b) $0.00001 \times 10^{-99}$

(c) $0.80001 \times 10^{1}$

(d) $2 \times (9 \times 10^4 \times 99 \times 2 + 10^5) - 1$

(e) $0.00001$

**Problem 2. [20 points (5+5+5+5)]** Let $f(x) = e^x - e^{-2x}$ and $\alpha$ be a small positive number.

(a) What numerical difficulty you may encounter while evaluating $f(\alpha)$?

(b) How can you overcome the difficulty to improve the accuracy of $f(\alpha)$ ?

(c) Calculate the condition number of the function $f(x)$.

(d) Identify all values of $x$ at which the condition number goes to infinity.

**Solution**

(a) Subtract cancellation. Since $e^x \approx e^{-2x}$ when $\alpha$ be a small positive number.

(b) Taylor expansion.

(c) $\kappa = \dfrac{x(e^x + 2e^{-2x})}{e^x - e^{-2x}}$

(d) Consider $x = 0$ since denominator is 0. However, $\lim_{x\to 0} \dfrac{x(e^x + 2e^{-2x})}{e^x - e^{-2x}} = 1$. Thus, no $x$ such that condition number goes to infinity.

**Problem 3. [20 points (12+8)]**

(a) Suppose $D$ is a real $n \times n$ diagonal matrix. Show that $||D||_2 = \max_{i=1:n} |D_{ii}|$. (Hint: Show that $||D||_2 \geq \max_{i=1:n} |D_{ii}|$ and $||D||_2 \leq \max_{i=1:n} |D_{ii}|$ .)

(b) Use Part (a) to show that $\kappa(D) = \frac{\max_{i=1:n} |D_{ii}|}{\min_{i=1:n} |D_{ii}|}$ in the 2-norm.

**Solution**

(a) By definition,

$$||D||_2 = \max_{x \neq 0} \frac{||Dx||_2}{||x||_2} = \max_{||x||_2=1} ||Dx||_2 \tag{1}$$

Let

$$D = \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{2}$$

Then Eq.(1) becomes

$$||D||_2 = \max_{||x||_2=1} \left( \sum_{i=1}^{n} |D_{ii} x_i|^2 \right)^{\frac{1}{2}} \tag{3}$$

[Prove $||D||_2 \geq \max_{i=1:n} |D_{ii}|$ (6 pts)]:
Note that for any $i = 1, \cdots, n$,

$$|D_{ii} x_i| \leq \left( \sum_{i=1}^{n} |D_{ii} x_i|^2 \right)^{\frac{1}{2}} \tag{4}$$

Assume $j$ is the index such that $|D_{jj}| = \max_{i=1:n} |D_{ii}|$. Choose

$$x = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{-th position}$$

This gives

$$\left( \sum_{i=1}^{n} |D_{ii} x_i|^2 \right)^{\frac{1}{2}} = |D_{jj}| \leq \max_{||x||_2=1} \left( \sum_{i=1}^{n} |D_{ii} x_i|^2 \right)^{\frac{1}{2}} = ||D||_2 \tag{5}$$

[Prove $||D||_2 \leq \max_{i=1:n} |D_{ii}|$ (6 pts)]:
For the other part of the inequality,

$$||D||_2 = \max_{||x||_2=1} \left( \sum_{i=1}^{n} |D_{ii} x_i|^2 \right)^{\frac{1}{2}} \leq \max_{||x||_2=1} |D_{jj}| \left( \sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} = |D_{jj}| \tag{6}$$

(b) [Consider $D$ is singular (1 pts)]:
If $D$ is singular, then some $D_{ii}$ must be zeros, so $\min_{i=1:n} |D_{ii}| = 0$ and $||D^{-1}|| = \infty$. Therefore,

$$\kappa(D) = ||D^{-1}||_2 ||D||_2 = \infty = \frac{\max_{i=1:n} |D_{ii}|}{\min_{i=1:n} |D_{ii}|} \tag{7}$$

[Consider $D$ is nonsingular (3+4 pts)]:
Now assume that $D$ is nonsingular, then

$$D^{-1} = \begin{bmatrix} \frac{1}{D_{11}} & & & \\ & \frac{1}{D_{22}} & & \\ & & \ddots & \\ & & & \frac{1}{D_{nn}} \end{bmatrix}. \tag{8}$$

[Find $||D^{-1}||_2$ (3 pts)]:
Therefore by (a),

$$||D^{-1}||_2 = \max_{i=1:n} \left| \frac{1}{D_{ii}} \right| = \frac{1}{\min_{i=1:n} |D_{ii}|} \tag{9}$$

[Compute $\kappa(D)$ (4 pts)]:
Finally, using the definition of matrix condition in 2-norm gives

$$\kappa(D) = ||D^{-1}||_2 ||D||_2 = \frac{\max_{i=1:n} |D_{ii}|}{\min_{i=1:n} |D_{ii}|} \tag{10}$$

**Problem 4. [20 points (10+10)]**

(a) Let $A = \begin{bmatrix} 2 & 4 & 2 \\ 4 & 6 & 2 \\ 2 & 2 & -26 \end{bmatrix}$. Factorize $A = LDL^T$, where $L$ is a lower matrix and $D$ is a diagonal matrix.

(b) Let $A$ be a $n \times n$ symmetric nonsingular matrix. Show that the matrix can be factorized as $A = LDL^T$.

**Solution**

(a) Perform LU factorization on $A$ gives

$$A = LU = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & 2 \\ & -2 & -2 \\ & & -26 \end{bmatrix} \tag{11}$$

The above equation can be rewritten to the form $A = LDU$ and $U = L^T$,

$$A = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & & \\ & -2 & \\ & & -26 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ & 1 & 1 \\ & & 1 \end{bmatrix} = LDL^T \tag{12}$$

[$L$ correct (5 pts) + $D$ correct (5 pts) = 10 pts]

(b) First, by LU factorization with partial pivoting,

$$PA = LDU, \tag{13}$$

where $P$ is the permutation matrix, $L$ is unit lower triangular, $U$ is unit upper triangular and $D$ is diagonal. Or equivalently, if some implicit permutations are allowed,

$$A = LDU. \tag{14}$$

Taking transpose on both sides of the above equation gives

$$A^T = U^T D L^T \tag{15}$$

The matrix $A$ is symmetric, hence $A = A^T$. Using the property $A = A^T$ gives

$$A = LDU = U^T D L^T \tag{16}$$

Then,

$$DU(L^T)^{-1} = L^{-1}U^T D. \tag{17}$$

Since $DU(L^T)^{-1}$ is upper triangular and $L^{-1}U^T D$ is lower triangular, these two matrices must be diagonal and

$$U(L^T)^{-1} = I \Rightarrow U = L^T \tag{18}$$

**Problem 5. [10 points (5+5)]**

(a) Define a general linear least square problem.

(b) Explain how you can use the QR factorization to solve the general linear least square problem.

**Solution**

(a) **The general linear least square problem:**
Given $\mathbf{A} \in \mathrm{R}^{m\times n}$ and $\mathbf{b} \in \mathrm{R}^m$, with $m > n$, find

$$\text{argmin}_{x\in \mathrm{R}^n} ||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2.$$

The notation "argmin" means to find an $\mathbf{b}$ that produces the minimum value.

(b) Te steps for solving the general linear least square problem $\mathbf{Ax} \approx \mathbf{b}$ are as follows:

(1.) Compute $\mathbf{N} = \mathbf{A^T A}$.

(2.) Compute $\mathbf{z} = \mathbf{A^T b}$.

(3.) Solve the $n \times n$ linear system $\mathbf{Nx} = \mathbf{z}$ for $\mathbf{x}$.

We substitute the matrix $\mathbf{A}$ to $\widehat{\mathbf{Q}}\widehat{\mathbf{R}}$ by using $\mathbf{QR}$ factorization, then

$$\begin{aligned} \mathbf{A^T A x} &= \mathbf{A^T b}, \\ \widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{Q}}^{\mathbf{T}}\widehat{\mathbf{Q}}\widehat{\mathbf{R}}\mathbf{x} &= \widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{Q}}^{\mathbf{T}}\mathbf{b}, \\ \widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{R}}\mathbf{x} &= \widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{Q}}^{\mathbf{T}}\mathbf{b} \end{aligned}$$

Hence

$$\mathbf{x} = \left(\widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{R}}\right)^{-1} \widehat{\mathbf{R}}^{\mathbf{T}}\widehat{\mathbf{Q}}^{\mathbf{T}}\mathbf{b}$$

**Problem 6. [20 points (5+5+10)]** Let $z$ be a $n \times 1$ vector and $v = ||z||e_1 - z$, and $P = I - 2\frac{vv^T}{v^T v}$.

(a) Show that $P$ is symmetric.

(b) Show that $P$ is orthogonal.

(c) Show that $Pz = ||z||e_1$.

**Solution**

(a) To show that $P$ is symmetric, that is $P^T = P$.

$$
\begin{aligned}
P^T &= \left(I - 2\frac{vv^T}{v^Tv}\right)^T \\
&= I^T - \frac{2}{||v||^2}(vv^T)^T \\
&= I - \frac{2}{||v||^2}(v^T)^T v^T \\
&= I - 2\frac{vv^T}{v^Tv} \\
&= P
\end{aligned}
$$

$P$ is symmetric since $P^T = P$. (Correct $P^T$ + Conclusion : 4 pts + 1 pt)

(b) To show that $P$ is orthogonal, that is $P^TP = I$.

$$
\begin{aligned}
P^TP &= \left(I - 2\frac{vv^T}{v^Tv}\right)^T \left(I - 2\frac{vv^T}{v^Tv}\right) \\
&= I^2 - \frac{2vv^T}{v^Tv} - \frac{2vv^T}{v^Tv} + \frac{4vv^Tvv^T}{v^Tvv^Tv} \\
&= I - \frac{4vv^T}{||v||^2} + \frac{4||v||^2vv^T}{||v||^4} \\
&= I
\end{aligned}
$$

$P$ is orthogonal since $P^TP = I$. (Correct $P^TP$ + Conclusion : 4 pts + 1 pt)

(c)

$$
\begin{aligned}
Pz &= \left(I - 2\frac{vv^T}{v^Tv}\right) z \\
&= z - 2\frac{vv^T}{v^Tv} z \\
&= z - 2\frac{v^Tz}{v^Tv} v
\end{aligned}
$$

(**Note:** $v^Tz$ is a constant.)

To show that $Pz = z - 2\frac{v^Tz}{v^Tv}v = ||z||e_1$, that is to show that $v^Tv + 2v^Tz = 0$.
Since

$$
\begin{aligned}
z - 2\frac{v^Tz}{v^Tv}v = ||z||e_1 &\iff -v = z - ||z||e_1 = 2\frac{v^Tz}{v^Tv}v \\
&\iff 2\frac{v^Tz}{v^Tv} + 1 = 0 \\
&\iff v^Tv + 2v^Tz = 0
\end{aligned}
$$

. ( Get the equivalence statement : 5 pts)

Then,

$$\begin{aligned} v^T v + 2v^T z &= v^T(v + 2z) \\ &= (||z||e_1^T - z^T)(||z||e_1 + z) \\ &= ||z||^2 e_1^T e_1 - ||z||z^T e_1 + ||z||e_1^T z - z^T z \\ &= ||z||^2 - ||z||e_1^T z + ||z||e_1^T z - ||z||^2 \\ &= 0 \end{aligned}$$

Hence, $Pz = ||z||e_1$. ( Prove $v^T v + 2v^T z = 0$ : 5 pts)