**458 Midterm 2 Solutions**[1]

## 1. Question 1

TRUE/FALSE

(a) The smallest positive number that exists in double precision floating point arithmetic is

$$2^{-1022}.$$

FALSE. This is the smallest positive normal number, which can be written as $1.0\cdots0 \times 2^{1023-1}$. There are subnormal numbers that are smaller, such as $2^{-1074}$.

(b) In double precision floating point arithmetic, the closest number to 4 that is larger than 4 is

$$4 + 2^{-52}.$$

(Put another way, if $x > 4$ is a double precision floating point number, then $|x - 4|$ is minimized when $x = 4 + 2^{-52}$.)

FALSE. The next largest double precision number after 4 can be written in a 52-digit binary after the decimal point as $1.0\ldots01 \times 2^2 = (1 + 2^{-52}) \times 4 = 4 + 4 \cdot 2^{-52}$.

(c) QR Decompositions are unique. That is, if $A$ is an $n \times n$ real matrix, then there is exactly one $n \times n$ orthogonal matrix $Q$ and there is exactly one $n \times n$ upper triangular matrix $R$ such that $A$ can be written as

$$A = QR.$$

FALSE. If $A = QR$ then $A = (-Q)(-R)$, with $-Q$ orthogonal and $-R$ upper triangular.

(d) Let $a_0, \ldots, a_{10}, b_0, \ldots, b_{10}$ be real numbers. Then there is a unique polynomial $p$ of degree at most 10 such that

$$p(a_i) = b_i \qquad \forall\, 0 \le i \le 10.$$

FALSE. If all of the $a_i$ are equal to 0, then $p$ is not unique.

(e) Let $m, n \ge 1$ be integers. Any real $m \times n$ matrix $A$ can be written as

$$A = UDV$$

where $U$ is an orthogonal $m \times m$ matrix, $V$ is an orthogonal $n \times n$ matrix, and $D$ is an $m \times n$ matrix whose non-diagonal entries are zero (i.e. $D_{ij} = 0$ whenever $1 \le i \le m, 1 \le j \le n$ and $i \ne j$.)

TRUE. This is a restated version of the existence of a singular value decomposition.

## 2. Question 2

Let $n \ge 1$ be an integer. Suppose I have a function $f\colon \mathbf{R} \to \mathbf{R}$ and I want to choose a polynomial $p_n$ that interpolates $f$ on the interval $[-1, 1]$. That is, we would like to choose nodes $a_0, \ldots, a_n \in [-1, 1]$ such that

$$f(a_i) = p_n(a_i), \qquad \forall\, 0 \le i \le n. \qquad (\ddagger)$$

- Suppose we want to choose the nodes $a_0, \ldots, a_n$ such that $\max_{y\in[-1,1]} |f(y) - p_n(y)|$ is as small as possible. Which nodes $a_0, \ldots, a_n$ would you choose? Justify your answer as best you can.

[1]November 5, 2022, 

- Suppose $n = 2$ and $a_0 = 0$, $a_1 = 1$ and $a_2 = 2$. Suppose also that

$$f(x) = x^3, \qquad \forall\, x \in \mathbf{R}.$$

Write an explicit formula for the degree 2 polynomial $p_2$ satisfying (‡).
Simplify your answer to the best of your ability.

*Solution.* We would choose the Chebyshev nodes $a_i := \cos((i + 1/2)\pi/(n + 1))$ for all $0 \leq i \leq n$, since these nodes lead to the best general error bound for $\max_{y\in[-1,1]} |f(x) - p_n(x)|$, i.e. this choice of nodes minimizes the quantity $\max_{|y|\leq 1} |\prod_{i=0}^{n}(y - a_i)|$. (The latter quantity is the only term in the error bound we wrote for $|f - p_n|$ that depends on the nodes $a_0, \ldots, a_n$.)

Using e.g. Theorem 5.1 in the notes, the polynomial $p_2$ can be written as

$$\begin{aligned} p_2(x) &= \sum_{j=0}^{2} f(a_j) \prod_{i\neq j} \frac{x - a_i}{a_j - a_i} = 0 + 1^3 \frac{x-0}{1-0}\frac{x-2}{1-2} + 2^3 \frac{x-0}{2-0}\frac{x-1}{2-1} \\ &= x(2 - x) + 8(x/2)(x - 1) = 2x - x^2 + 4x^2 - 4x \\ &= 3x^2 - 2x. \end{aligned}$$

## 3. Question 3

Suppose we have data points $(-1, 1), (0, 3), (1, 3) \in \mathbf{R}^2$ denoted as $\{(a_i, b_i)\}_{i=1}^{3}$. Find the line that best fits the data. That is, find the line $f\colon \mathbf{R} \to \mathbf{R}$ that minimizes the sum of squared differences $\sum_{j=1}^{3} |f(a_i) - b_i|^2$.

(You should find an exact formula for $f$. Do not write a Matlab program in this question.)

*Solution.* We would like to minimize $\|Ax - b\|^2$ where $f(t) = x_0 + tx_1$, and

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad x = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix}.$$

To do this, we could either try to solve this minimization problem directly, or we could solve $A^T Ax = A^T b$ (equivalence follows by Lemma 4.95 in the notes). In the second case, we have

$$A^T A = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

$$A^T b = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$$

So, solving $A^T Ax = A^T b$ amounts to solving

$$\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} x = \begin{pmatrix} 7 \\ 2 \end{pmatrix}.$$

That is, we find that $x_0 = 7/3$ and $x_1 = 1$. So, the best fit line is

$$f(t) = (7/3) + t, \qquad \forall\, t \in \mathbf{R}.$$

Alternatively (for those who know multivariable calculus), we could just minimize the function $g(x_0, x_1) := \|Ax - b\|^2$ directly, noting that

$$g = \left\| \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} x_0 - x_1 \\ x_0 \\ x_0 + x_1 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} x_0 - x_1 - 1 \\ x_0 - 3 \\ x_0 + x_1 - 3 \end{pmatrix} \right\|^2$$
$$= (x_0 - x_1 - 1)^2 + (x_0 - 3)^2 + (x_0 + x_1 - 3)^2.$$

We then have

$$\nabla g = \begin{pmatrix} 2(x_0 - x_1 - 1) + 2(x_0 - 3) + 2(x_0 + x_1 - 3) \\ -2(x_0 - x_1 - 1) + 2(x_0 + x_1 - 3) \end{pmatrix} = \begin{pmatrix} 6x_0 - 14 \\ 4x_1 - 4. \end{pmatrix}$$

Setting the gradient equal to zero, we get $x_1 = 1$ and $x_0 = 14/6 = 7/3$. Since $g$ is strictly convex and has a single critical point, this critical point is the global minimum of $g$.

## 4. Question 4

Let $A$ be a real $n \times n$ symmetric positive definite matrix all of whose eigenvalues are distinct.

Write a Matlab program that applies the QR algorithm to $A$. The output of the program should be two real $n \times n$ matrices $D$ and $Q$, such that $D$ is a diagonal matrix containing the eigenvalues of $D$, and $Q$ is an orthogonal matrix whose columns are eigenvectors of $A$, so that $A = QDQ^T$. (You are allowed to use the built-in Matlab program `qr`, whose syntax is `[Q,R]=qr(A)`, outputting a factorization $A = QR$.)

(It is okay if $QDQ^T$ is only approximately equal to $A$ and $D$ is only approximately diagonal, so that all non-diagonal entries of $D$ are small.)

In this problem, you will be graded on writing correct syntax in Matlab. Syntax mistakes will result in deductions of points.

*Solution.*

```
Qa=eye(length(A));
for i=1:100
    [Q R]=qr(A);
    A=R*Q;
    Qa=Qa*Q  % updated Qa, as in Theorem 4.90 in notes
end
Qa*A*Qa'   % should return the original matrix
% the output Qa is the matrix of eigenvectors
% the output A should be (nearly) diagonal with all eigenvalues
```

## 5. Question 5

Let $n = 500$. Let $A$ be a real $n \times n$ symmetric positive definite matrix all of whose eigenvalues are distinct. Suppose the largest eigenvalue of $A$ is 1 and all other eigenvalues of $A$ are at most $1/2$.

Suppose $x \in \mathbf{R}^n$ is a nonzero vector such that $Ax = x$.

Describe, to the best of your ability, the matrix

$$A^{1000}.$$

Justify your answer. Simplify your answer as best you can.

You should be able to say what each row of $A^{1000}$ is, within a reasonably small margin of error. For example, you should have a fairly precise description of the $356^{th}$ row of $A$.

*Solution.* From the spectral theorem, $A = QDQ^T$ where $D$ is a diagonal matrix containing the eigenvalues of $A$, and $Q$ is an orthogonal matrix whose columns contain the eigenvectors of $A$. We can choose $D$ such that $D_{11} = 1$ and by assumption $0 \leq D_{ii} \leq 1/2$ for all $i \geq 2$. Moreover, by this choice of $D$, the first column of $Q$ can be chosen as $x/\|x\|$. Then $A^{1000} = (QDQ^T)^{1000} = QD^{1000}Q^T$. By assumption, all entries of $D^{1000}$ will be smaller than $2^{-1000}$, except for the top left entry, which is one. Therefore, $A^{1000}$ is approximately equal to

$$Q \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} Q^T = \begin{pmatrix} \frac{x}{\|x\|} & \cdots \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \frac{x}{\|x\|} & \cdots \end{pmatrix}^T$$

$$= \begin{pmatrix} \frac{x}{\|x\|} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \frac{x}{\|x\|} & 0 & \cdots & 0 \end{pmatrix}^T = \frac{xx^T}{\|x\|^2}.$$

That is, $A^{1000}$ is approximately equal to $\frac{xx^T}{\|x\|^2}$.