Numerical Methods 2022 — Final Exam

Solutions

Problem 1 (40 pts). Consider a linear system $A\boldsymbol{x} = \boldsymbol{b}$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(a) (5 pts) Is the matrix A symmetric positive definite? If so, please prove it. Otherwise, give a counter example $\boldsymbol{x}$.

(b) (5 pts) Take $\boldsymbol{x}_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$. Do two CG iterations and show what $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are. Is $\boldsymbol{x}_2$ a solution or not? You must show the details such as α, β, ρ, etc. in the middle of the procedure.

(c) (10 pts) Let $\boldsymbol{r}_1$ and $\boldsymbol{p}_1$ be the vectors calculated in (b). Solve the following optimization problem of the variable $\boldsymbol{p}$ directly (By plugging the constraint into the objective, you get a unconstrained problem which can be solved easily.):

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \|\boldsymbol{p} - \boldsymbol{r}_1\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{p} \in \operatorname{span}\{A\boldsymbol{p}_1\}^{\perp} \end{aligned} \tag{1}$$

How does the solution of this problem connect to the $\boldsymbol{p}_2$ obtained in the CG procedure?

(d) (5 pts) In slides "sparse_CG4.pdf", we have a theorem saying that if

$$A = I + B$$

for some matrix B, then there is an upperbound on the number of CG steps in terms of rank(B). From what you have observed in (a)-(c), what is the lower bound of rank(B)? Check B to confirm the result.

(e) (5 pts) In our slides "sparse_CG3.pdf", we have a theorem in page 1, which said that

$$\begin{aligned} \operatorname{span}\left\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_j\right\} =& \operatorname{span}\left\{\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{j-1}\right\} \\ =& \operatorname{span}\left\{\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{j-1}\boldsymbol{b}\right\} \end{aligned} \tag{2}$$

after jth iteration of CG. Please check whether (2) holds for $j = 2$ in this case. Hint: You may consider some results in the process of doing sub-problem (b).

(f) (5 pts) Now consider

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \text{ and } \boldsymbol{b} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

Prove that A is only positive semi-definite but not positive definite.

(g) (5 pts) Following (f), run CG on the new problem and see if it fails at some time points. See if this corresponds to our explanation in the end of "sparse_CG4.pdf" about why CG requires A to be positive definite.

Solution.

(a) For any vector $\boldsymbol{x} = (x_1, x_2, x_3)$, we have

$$\begin{aligned} \boldsymbol{x}^T A \boldsymbol{x} &= \boldsymbol{x}^T \begin{bmatrix} x_1 \\ 2x_2 + x_3 \\ x_2 + x_3 \end{bmatrix} \\ &= x_1^2 + x_2(2x_2 + x_3) + x_3(x_2 + x_3) \\ &= x_1^2 + 2x_2^2 + 2x_2x_3 + x_3^2 \\ &= x_1^2 + x_2^2 + (x_2 + x_3)^2 \\ &\geq 0 \end{aligned}$$

If $\boldsymbol{x}^T A \boldsymbol{x} = 0$, then $x_1 = x_2 = (x_2 + x_3) = 0$. Therefore, $\boldsymbol{x}^T A \boldsymbol{x} = 0$ only when $\boldsymbol{x} = 0$. Therefore, A is positive definite.

(b) In the beginning, we have

$$\boldsymbol{x}_0 = 0, \boldsymbol{r}_0 = \boldsymbol{b}, \rho_0 = 3. \tag{3}$$

In the first iteration, we have

$$\boldsymbol{p}_1 = \boldsymbol{b},\ \boldsymbol{w}_1 = A\boldsymbol{p}_1 = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix},\ \alpha_1 = \frac{3}{\boldsymbol{p}_1^T \boldsymbol{w}_1} = \frac{1}{2}, \tag{4}$$

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \end{bmatrix},\ \boldsymbol{r}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \alpha_1 \boldsymbol{w}_1 = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix},\ \rho_1 = \frac{1}{2}. \tag{5}$$

In the second iteration, we have

$$\beta_2 = \frac{1/2}{3} = \frac{1}{6},\ \boldsymbol{p}_2 = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2/3 \\ -1/3 \\ 1/6 \end{bmatrix}, \tag{6}$$

$$\boldsymbol{w}_2 = A\boldsymbol{p}_2 = \begin{bmatrix} 2/3 \\ -1/2 \\ -1/6 \end{bmatrix},\ \alpha_2 = \frac{1/2}{\boldsymbol{p}_2^T \boldsymbol{w}_2} = \frac{6}{7},$$

$$\boldsymbol{x}_2 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} + \frac{6}{7} \begin{bmatrix} 2/3 \\ -1/2 \\ 1/6 \end{bmatrix} = \frac{1}{14} \begin{bmatrix} 15 \\ 3 \\ 9 \end{bmatrix}.$$

We can know that $\boldsymbol{x}_2$ is not the solution since $A\boldsymbol{x}_2 = \frac{1}{14} \begin{bmatrix} 15 \\ 15 \\ 12 \end{bmatrix} \neq \boldsymbol{b}$.

(c) The vector $\boldsymbol{p} = (p_1, p_2, p_3)$ has to satisfy the condition

$$\boldsymbol{p} \in \text{span}\{A\boldsymbol{p}_1\}^{\perp} \iff \boldsymbol{p}^T A \boldsymbol{p}_1 = 0 \iff p_1 + 3p_2 + 2p_3 = 0. \tag{7}$$

Therefore, the equivalent optimization problem is

$$\begin{aligned} \min_{\boldsymbol{p}} \ \ & F(\boldsymbol{p}) = \left\| \boldsymbol{p} - \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} \right\|_2^2 \\ \text{s.t.} \ \ & p_1 + 3p_2 + 2p_3 = 0 \end{aligned}$$

Plugging the constraint $p_1 = -3p_2 - 2p_3$ into $F(\boldsymbol{p})$, we get a new equivalent problem:

$$\min_{p_2, p_3} f(p_2, p_3) = (-3p_2 - 2p_3 - \frac{1}{2})^2 + (p_2 + \frac{1}{2})^2 + p_3^2$$

The is an unconstrained problem which can be solved by setting the derivative as zero:

$$\begin{aligned} \frac{\partial f}{\partial p_2} = 0 &\implies -6(-3p_2 - 2p_3 - \frac{1}{2}) + 2(p_2 + \frac{1}{2}) = 0 \\ \frac{\partial f}{\partial p_3} = 0 &\implies -4(-3p_2 - 2p_3 - \frac{1}{2}) + 2p_3 = 0 \end{aligned} \tag{8}$$

By solving (8) and then using (7), we get the solution

$$\boldsymbol{p} = \frac{1}{7} \begin{bmatrix} 4 \\ -2 \\ 1 \end{bmatrix} = \frac{6}{7} \boldsymbol{p}_2.$$

We can see that the solution $\boldsymbol{p}$ is parallel to the $\boldsymbol{p}_2$ obtained in (b).

(d) In (b), we already know that the algorithm have not converged after 2 iterations. Then, from the theorem in the slide we know that $\text{rank}(B) + 1 > 2$ and so $\text{rank}(B) \geq 2$. To confirm this, we can calculate that

$$B = A - I = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \implies \text{rank}(B) = 2.$$

(e) First, we show that $\text{span}\{\boldsymbol{p}_1, \boldsymbol{p}_2\} = \text{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1\}$. We know that

$$\boldsymbol{r}_0 = \boldsymbol{b} = \boldsymbol{p}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Also, from (6) we know that

$$\boldsymbol{r}_1 = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} = \boldsymbol{p}_2 - \frac{1}{6}\boldsymbol{p}_1,$$

so $\operatorname{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1\} \subseteq \operatorname{span}\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$. By rearranging the terms, we then have

$$\begin{aligned} \boldsymbol{p}_1 &= \boldsymbol{r}_0 \\ \boldsymbol{p}_2 &= \boldsymbol{r}_1 + \frac{1}{6}\boldsymbol{r}_0, \end{aligned}$$

so $\operatorname{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1\} = \operatorname{span}\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$.

Next, we show that $\operatorname{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1\} = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}\}$. From (4), we know that

$$A\boldsymbol{b} = \boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}.$$

Also, from (3) and (5) we respectively have

$$\begin{aligned} \boldsymbol{r}_0 &= \boldsymbol{b} \\ \boldsymbol{r}_1 &= \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} = \boldsymbol{b} - \frac{1}{2}A\boldsymbol{b}, \end{aligned}$$

which then implies that

$$\boldsymbol{b} = \boldsymbol{r}_0 \text{ and } A\boldsymbol{b} = -2\boldsymbol{r}_1 + 2\boldsymbol{r}_0.$$

Therefore, $\operatorname{span}\{\boldsymbol{r}_0, \boldsymbol{r}_1\} = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}\}$.

(f) For any vector $\boldsymbol{x} = (x_1, x_2, x_3)$, we have

$$\begin{aligned} \boldsymbol{x}^T A\boldsymbol{x} = \boldsymbol{x}^T &\begin{bmatrix} x_1 \\ x_2 + x_3 \\ x_2 + x_3 \end{bmatrix} \\ &= x_1^2 + x_2(x_2 + x_3) + x_3(x_2 + x_3) \\ &= x_1^2 + x_2^2 + 2x_2x_3 + x_3^2 \\ &= x_1^2 + (x_2 + x_3)^2 \\ &\geq 0 \end{aligned}$$

Since $\boldsymbol{x}^T A\boldsymbol{x} = 0$ when $x_1 = 0, x_2 = 1$ and $x_3 = -1$, A is positive semi-definite but not positive-definite.

(g) In the beginning, we have

$$\boldsymbol{x}_0 = 0, \boldsymbol{r}_0 = \boldsymbol{b}, \rho_0 = 3.$$

In the first iteration, we have

$$\boldsymbol{p}_1 = \boldsymbol{b},\ \boldsymbol{w}_1 = A\boldsymbol{p}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix},\ \alpha_1 = \frac{3}{\boldsymbol{p}_1^T\boldsymbol{w}_1} = 3,$$

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + 3\begin{bmatrix}1\\-1\\1\end{bmatrix} = \begin{bmatrix}3\\-3\\3\end{bmatrix},\ \boldsymbol{r}_1 = \begin{bmatrix}1\\-1\\1\end{bmatrix} - \alpha_1\boldsymbol{w}_1 = \begin{bmatrix}-2\\-1\\1\end{bmatrix},\ \rho_1 = 6.$$

In the second iteration, we have

$$\beta_2 = \frac{6}{3} = \frac{1}{6},\ \boldsymbol{p}_2 = \begin{bmatrix}-2\\-1\\1\end{bmatrix} + 2\begin{bmatrix}1\\-1\\1\end{bmatrix} = \begin{bmatrix}0\\-3\\3\end{bmatrix},\ \boldsymbol{w}_2 = A\boldsymbol{p}_2 = \begin{bmatrix}0\\0\\0\end{bmatrix},\ \alpha_2 = \frac{6}{\boldsymbol{p}_2^T\boldsymbol{w}_2} = \frac{6}{0},$$

so the CG procedure failed because of division by zero.

From the slides, we need $\boldsymbol{p}_2^T A\boldsymbol{p}_2 > 0$ to obtain α. However, we now have $\boldsymbol{p}_2^T A\boldsymbol{p}_2 = 0$ since A is only semi-definite, so CG fails.

Problem 2 (15 pts). On page 8 of "sparse_CG3.pdf", we wish to minimize

$$\min_{\boldsymbol{w},\mu} \|(1 + \frac{\mu}{\alpha_{k-1}})\boldsymbol{r}_{k-1}\|^2 + \| - \frac{\mu}{\alpha_{k-1}}\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{w}\|^2.$$

We claimed that if the optimal solution is $(\boldsymbol{w}^*, \mu^*)$, then

$$-\frac{\boldsymbol{w}^*}{\mu^*/\alpha_{k-1}} \text{ must be the solution to } \min_{\boldsymbol{z}} \|\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z}\|. \tag{9}$$

In this problem, we are going to verify this relation through direct calculation.

(a) (10 pts) Minimizing the function

$$\min_{\boldsymbol{z}} \|\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z}\|$$

is equivalent to minimizing

$$\min_{\boldsymbol{z}} \|\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z}\|^2,$$

which is a quadratic function of $\boldsymbol{z}$. Solve this optimization problem by setting the gradient with respect to $\boldsymbol{z}$ as zero. The solution $\boldsymbol{z}^*$ should be expressed in terms of A, P_{k-2} and $\boldsymbol{r}_{k-2}$.

(b) (5 pts) Similar to (a), by defining

$$F(\boldsymbol{w}, \mu) = \|(1 + \frac{\mu}{\alpha_{k-1}})\boldsymbol{r}_{k-1}\|^2 + \| - \frac{\mu}{\alpha_{k-1}}\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{w}\|^2,$$

we have a quadratic function of $(\boldsymbol{w}, \mu)$. By solving $\nabla_{\boldsymbol{w}} F(\boldsymbol{w}^*, \mu^*) = \boldsymbol{0}$, show that (9) is correct.

Solution.

(a) The function we are minimizing is

$$\begin{aligned} F(\boldsymbol{z}) &= \|\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z}\|^2 \\ &= (\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z})^T(\boldsymbol{r}_{k-2} - AP_{k-2}\boldsymbol{z}) \\ &= \boldsymbol{r}_{k-2}^T\boldsymbol{r}_{k-2} - 2\boldsymbol{r}_{k-2}^T AP_{k-2}\boldsymbol{z} + \boldsymbol{z}^T P_{k-2}^T A^T AP_{k-2}\boldsymbol{z}. \end{aligned}$$

The minimizer must satisfy

$$\nabla_{\boldsymbol{z}} F(\boldsymbol{z}^*) = -2P_{k-2}^T A^T \boldsymbol{r}_{k-2} + 2P_{k-2}^T A^T AP_{k-2}\boldsymbol{z}^* = 0.$$

Therefore, the solution is

$$\boldsymbol{z}^* = (P_{k-2}^T A^T AP_{k-2})^{-1} P_{k-2}^T A^T \boldsymbol{r}_{k-2}.$$

(b) By re-writing the norm as matrix products, the obtain the new objective we want to minimize:

$$F(\boldsymbol{w}, \mu) = (1 + \frac{\mu}{\alpha_{k-1}})^2 \boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1} + (\frac{\mu}{\alpha_{k-1}})^2 \boldsymbol{r}_{k-2}^T\boldsymbol{r}_{k-2} + \boldsymbol{w}^T P_{k-2}^T A^T AP_{k-2}\boldsymbol{w} + \frac{2\mu}{\alpha_{k-1}} \boldsymbol{r}_{k-2}^T AP_{k-2}\boldsymbol{w}$$

The minimum $(\boldsymbol{w}^*, \mu^*)$ must satisfy

$$\nabla_{\boldsymbol{w}} F(\boldsymbol{w}^*, \mu^*) = 2P_{k-2}^T A^T AP_{k-2}\boldsymbol{w}^* + \frac{2\mu^*}{\alpha_{k-1}} P_{k-2}^T A^T \boldsymbol{r}_{k-2} = 0$$

$$\implies \boldsymbol{w}^* = -\frac{\mu^*}{\alpha_{k-1}} (P_{k-2}^T A^T AP_{k-2})^{-1} P_{k-2}^T A^T \boldsymbol{r}_{k-2} \tag{10}$$

Indeed, we see that

$$-\frac{\boldsymbol{w}^*}{\mu^*/\alpha_{k-1}} = (P_{k-2}^T A^T AP_{k-2})^{-1} P_{k-2}^T A^T \boldsymbol{r}_{k-2} = \boldsymbol{z}^*.$$

Problem 3 (15 pts). Let the function f be

$$f(x) = 4\left(\frac{x}{\pi}\right)^2.$$

We wish to approximate $f(x)$ using a Fourier series with n term

$$S_n(x) = \frac{a_0 + a_n \cos nx}{2} + \sum_{k=1}^{n-1} \big(a_k \cos kx + b_k \sin kx\big) \tag{11}$$

and $2m$ points

$$(x_0, f(x_0)), \cdots, (x_{2m-1}, f(x_{2m-1})),$$

where

$$x_k = -\pi + \frac{k}{m}\pi,\ k = 0, \dots, 2m-1.$$

In the following subproblems, we consider the case where $m = n = 2$.

(a) (5 pts) Please give the matrix A_2, A_1 and P so that the Fourier Transform matrix F can be decomposed into

$$F = A_2 A_1 P.$$

(b) (5 pts) Following (a), calculate the coefficient vector $\boldsymbol{c}$ given by

$$\boldsymbol{c} = A_2 A_1 P \boldsymbol{y} \tag{12}$$

where $y_k = f(x_k)$.

(c) (5 pts) Following (b), calculate

$$a_0, a_1, a_2$$

and

$$b_1$$

from $\boldsymbol{c}$. Then, write down the transformed series in the form of (11).

Solution.

(a) The δ we will be using is

$$e^{-i\pi/m} = -i.$$

To calculate A_2, we have

$$L = 2^2 = 4, \quad r = \frac{2m}{L} = \frac{4}{4} = 1,$$

and so

$$\begin{aligned} A_2 &= I_r \otimes B_L \\ &= I_1 \otimes B_4 \\ &= \begin{bmatrix} 1 \end{bmatrix} \otimes \begin{bmatrix} I_2 & \Omega_2 \\ I_2 & -\Omega_2 \end{bmatrix}, \quad \text{where } \Omega_2 = \begin{bmatrix} 1 & 0 \\ 0 & \delta \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -i \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & i \end{bmatrix} \end{aligned} \tag{13}$$

To get A_1, we have

$$L = 2^1 = 2, \quad r = \frac{2m}{L} = 2.$$

and

$$\begin{aligned} A_1 &= I_r \otimes B_L \\ &= I_2 \otimes B_2 \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} I_1 & \Omega_1 \\ I_1 & -\Omega_1 \end{bmatrix}, \quad \text{where } \Omega_1 = \begin{bmatrix} 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \end{aligned} \tag{14}$$

The permutation is calculated by reversing the binary representation.

$00 \to 00$	column 0 swapped to column 0
$01 \to 10$	column 1 swapped to column 2
$10 \to 01$	column 2 swapped to column 1
$11 \to 11$	column 3 swapped to column 3

Therefore, we have

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

(b) First, the vector $\boldsymbol{y}$ is

$$\begin{bmatrix} f(-\pi) \\ f(-\frac{\pi}{2}) \\ f(0) \\ f(\frac{\pi}{2}) \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

Therefore, by (15) we have

$$P\boldsymbol{y} = \begin{bmatrix} 4 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Then by (14) and (13) we can then calculate

$$A_1(P\boldsymbol{y}) = \begin{bmatrix} 4 \\ 4 \\ 2 \\ 0 \end{bmatrix}$$

$$\text{and } \boldsymbol{c} = A_2(A_1 P\boldsymbol{y}) = \begin{bmatrix} 6 \\ 4 \\ 2 \\ 4 \end{bmatrix}.$$

(c) From $\boldsymbol{c}$ we can then calculate

$$\begin{aligned} a_0 &= \frac{\mathrm{Re}(c_0)(-1)^0}{2} = 3 \\ a_1 &= \frac{\mathrm{Re}(c_1)(-1)^1}{2} = -2 \\ a_2 &= \frac{\mathrm{Re}(c_2)(-1)^2}{2} = 1 \\ b_1 &= \frac{\mathrm{Im}(c_0)(-1)^1}{2} = 0 \end{aligned}$$

and so the transformed series is

$$S_m(x) = \frac{3 + \cos(2x)}{2} - 2\cos(x).$$

Problem 4 (30 pts). For least-square regression, we solve the following optimization problem

$$\min_{a,b} \sum_i (y_i - (ax_i + b))^2,$$

where $x \in R$. Now instead we consider

$$\sum_i \max(y_i - (ax_i + b) - \varepsilon, 0)^2 + \sum_i \max((ax_i + b) - y_i - \varepsilon, 0)^2.$$

In other words, originally we have

$$\text{error}_i = |y_i - (ax_i + b)|$$

but now

$$\text{error}_i = \begin{cases} 0 & \text{if } \varepsilon \le y_i - (ax_i + b) \le \varepsilon \\ |y_i - (ax_i + b)| & \text{otherwise} \end{cases}$$

(a) (5 pts) Prove that

$$\max(c, 0)^2$$

is a differentiable function on c.

(b) (5 pts) Use (a) and chain rule to calculate derivatives with respect to a and b. That is, with the definitions

$$\begin{aligned} G_i &= \max(y_i - (ax_i + b) - \varepsilon, 0)^2 \\ H_i &= \max((ax_i + b) - y_i - \varepsilon, 0)^2 \\ F_i &= G_i + H_i, \end{aligned}$$

please derive the formulation of

$$\frac{\partial G_i}{\partial a}, \frac{\partial G_i}{\partial b}, \frac{\partial H_i}{\partial a} \text{ and } \frac{\partial H_i}{\partial b}, \forall i.$$

at first, and then you can get the final results

$$\frac{\partial F_i}{\partial a} \text{ and } \frac{\partial F_i}{\partial b}, \forall i.$$

Hint: to simplify the representation, consider using

$$t_i = y_i - (ax_i + b).$$

Then you may need to separately consider situations such as

$$t_i - \varepsilon > 0, t_i - \varepsilon \le 0,$$

etc.

(c) (10 pts) Consider

$$\begin{aligned} (x_1, y_1) =& (0, 0) \\ (x_2, y_2) =& (1, 2) \\ (x_3, y_3) =& (2, 1) \end{aligned}$$

and

$$\varepsilon = \frac{1}{2}.$$

In this and the next sub-problem, we aim to solve equations derived in (b) to get a^* and b^*. To begin, consider the results

$$\frac{\partial F_i}{\partial a}, \forall i$$

from (b). Please identify three possible situations of

$$\frac{\partial F_1}{\partial a} + \frac{\partial F_2}{\partial a} + \frac{\partial F_3}{\partial a} = 0$$

according to the sign of

$$\frac{\partial F_i}{\partial a}, \forall i.$$

(d) (10 pts) Now move to check

$$\frac{\partial F_i}{\partial b}, \forall i$$

and the condition

$$\frac{\partial F_1}{\partial b} + \frac{\partial F_2}{\partial b} + \frac{\partial F_3}{\partial b} = 0.$$

See if for $i = 2, 3$, you can identify the relationship between

$$\frac{\partial F_i}{\partial a} \text{ and } \frac{\partial F_i}{\partial b}. \tag{16}$$

From (16) and the result of (c), you can then separately discuss three situations for finding the optimal solution a^* and b^*. Hint: One of the situation involves in analyzing inequalities. Specifically, you have three types of inequalities of a and b. You can draw a figure to see if there are points satisfying all inequalities. For the other two cases, you must solve two-variable linear systems.

Solution.

(a) Let

$$F(c) = \max(c, 0)^2 = \begin{cases} c^2 & c > 0 \\ 0 & c \leq 0 \end{cases}.$$

If $c \neq 0$, $F(c)$ is a polynomial function. Thus, we have

$$F'(c) = \begin{cases} 2c & c > 0 \\ 0 & c < 0 \end{cases}.$$

For the derivative of $F(0)$, we use the derivative's definition on two sides. The right side is

$$\lim_{t \to 0^+} \frac{F(0+t) - F(0)}{t} = \lim_{t \to 0^+} \frac{t^2}{t} = \lim_{t \to 0^+} t = 0,$$

and the left is

$$\lim_{t \to 0^-} \frac{F(0+t) - F(0)}{t} = \lim_{t \to 0^-} \frac{0}{t} = 0.$$

Therefore, we can conclude that

$$F'(c) = \begin{cases} 2c & , c > 0 \\ 0 & , c \leq 0 \end{cases},$$

which means F is differential function on c.

(b) Let us take

$$G_i = \max\left(t_i - \varepsilon, 0\right)^2$$

and

$$H_i = \max\left(-t_i - \varepsilon, 0\right)^2 .$$

The derivatives of G_i on a and b are

$$\frac{\partial G_i}{\partial a} = \begin{cases} \frac{\partial G_i}{\partial (t_i-\varepsilon)} \frac{\partial (t_i-\varepsilon)}{\partial a} \\ 0 \end{cases} = \begin{cases} 2\left(t_i-\varepsilon\right)\cdot\left(-x_i\right) & , \; t_i-\varepsilon > 0 \\ 0 & , \; t_i - \varepsilon \le 0 \end{cases}$$

and

$$\frac{\partial G_i}{\partial b} = \begin{cases} \frac{\partial G_i}{\partial (t_i-\varepsilon)} \frac{\partial (t_i-\varepsilon)}{\partial b} \\ 0 \end{cases} = \begin{cases} 2\left(t_i-\varepsilon\right)\cdot\left(-1\right) & , \; t_i-\varepsilon > 0 \\ 0 & , \; t_i - \varepsilon \le 0 \end{cases}$$

Similarly, the derivatives of H_i on a and b are

$$\frac{\partial H_i}{\partial a} = \begin{cases} 2\left(-t_i-\varepsilon\right)\cdot x_i & , \; -t_i-\varepsilon > 0 \\ 0 & , \; -t_i - \varepsilon \le 0 \end{cases}$$

and

$$\frac{\partial H_i}{\partial b} = \begin{cases} 2\left(-t_i-\varepsilon\right)\cdot 1 & , \; -t_i-\varepsilon > 0 \\ 0 & , \; -t_i - \varepsilon \le 0 \end{cases}$$

Moreover, the derivatives of $G_i + H_i$ can be derived as

$$\frac{\partial (G_i + H_i)}{\partial a} = \begin{cases} -2\left(t_i-\varepsilon\right)\cdot x_i & , \; t_i > \varepsilon \\ -2\left(t_i+\varepsilon\right)\cdot x_i & , \; t_i < -\varepsilon \\ 0 & , \; -\varepsilon \le t_i \le \varepsilon \end{cases}$$

and

$$\frac{\partial (G_i + H_i)}{\partial b} = \begin{cases} -2\left(t_i-\varepsilon\right) & , \; t_i > \varepsilon \\ -2\left(t_i+\varepsilon\right) & , \; t_i < -\varepsilon \\ 0 & , \; -\varepsilon \le t_i \le \varepsilon \end{cases}$$

(c) Let us define

$$F_i = G_i + H_i,$$

and we have

$$\begin{aligned} t_1 =& 0 - b \\ t_2 =& 2 - a - b \; . \\ t_3 =& 1 - 2a - b \end{aligned}$$

Then, we firstly derive

$$\frac{\partial F_1}{\partial a} = \begin{cases} -2\left(t_1 - \frac{1}{2}\right)\cdot 0 \\ -2\left(t_1 + \frac{1}{2}\right)\cdot 0 \\ 0 \end{cases} = \begin{cases} 0 & , \; t_1 > \frac{1}{2} \\ 0 & , \; t_1 < -\frac{1}{2} \\ 0 & , \; -\frac{1}{2} \le t_1 \le \frac{1}{2} \end{cases}$$

$$\frac{\partial F_2}{\partial a} = \begin{cases} -2\left(t_2 - \frac{1}{2}\right)\cdot 1 \\ -2\left(t_2 + \frac{1}{2}\right)\cdot 1 \\ 0 \end{cases} = \begin{cases} -2t_2 + 1 < 0 & , \; t_2 > \frac{1}{2} \\ -2t_2 - 1 > 0 & , \; t_2 < -\frac{1}{2} \\ 0 & , \; -\frac{1}{2} \le t_2 \le \frac{1}{2} \end{cases}$$

$$\frac{\partial F_3}{\partial a} = \begin{cases} -2\left(t_3 - \frac{1}{2}\right)\cdot 2 \\ -2\left(t_3 + \frac{1}{2}\right)\cdot 2 \\ 0 \end{cases} = \begin{cases} -4t_3 + 2 < 0 & , \; t_3 > \frac{1}{2} \\ -4t_3 - 2 > 0 & , \; t_3 < -\frac{1}{2} \\ 0 & , \; -\frac{1}{2} \le t_3 \le \frac{1}{2} \end{cases}$$

and focus on the equation

$$\frac{\partial F_1}{\partial a} + \frac{\partial F_2}{\partial a} + \frac{\partial F_3}{\partial a} = 0. \tag{17}$$

Because

$$\frac{\partial F_1}{\partial a} = 0, \ \forall t_1,$$

to satisfy (17), we can only have the following cases:

$\frac{\partial F_2}{\partial a}$	$\frac{\partial F_3}{\partial a}$	$\frac{\partial F_1}{\partial a} + \frac{\partial F_2}{\partial a} + \frac{\partial F_3}{\partial a}$
0	0	0
+	-	$-2t_2 - 1 - 4t_3 + 2$
-	+	$-2t_2 + 1 - 4t_3 - 2$

(i) In this situation

$$-\frac{1}{2} \leq t_2 \leq \frac{1}{2} \text{ and } -\frac{1}{2} \leq t_3 \leq \frac{1}{2}.$$

(ii) In this situation

$$t_2 < -\frac{1}{2}, \ t_3 > \frac{1}{2} \text{ and } -2t_2 - 4t_3 + 1 = 0.$$

(iii) In this situation

$$t_2 > \frac{1}{2}, \ t_3 < -\frac{1}{2} \text{ and } -2t_2 - 4t_3 - 1 = 0.$$

Next, we move on to the equation

$$\frac{\partial F_1}{\partial b} + \frac{\partial F_2}{\partial b} + \frac{\partial F_3}{\partial b} = 0 \tag{18}$$

with

$$\frac{\partial F_i}{\partial b} = \left\{ \begin{array}{c} -2\left(t_i - \frac{1}{2}\right) \\ -2\left(t_i + \frac{1}{2}\right) \\ 0 \end{array} \right. = \left\{ \begin{array}{cl} -2t_i + 1 < 0 & , \ t_i > \frac{1}{2} \\ -2t_i - 1 > 0 & , \ t_i < -\frac{1}{2} \\ 0 & , \ -\frac{1}{2} \leq t_i \leq \frac{1}{2} \end{array} \right. , \ i = 1, \ldots, 3.$$

Clearly,

$$\frac{\partial F_2}{\partial b} = \frac{\partial F_2}{\partial a}$$

and

$$\frac{\partial F_3}{\partial b} = \frac{\partial F_3}{\partial a} \cdot \frac{1}{2}.$$

Therefore, from (18) and the three cases to consider $\partial F_i / \partial a$, we can have only the following situations.

$\frac{\partial F_1}{\partial b}$	$\frac{\partial F_2}{\partial b}$	$\frac{\partial F_3}{\partial b}$	$\frac{\partial F_1}{\partial b} + \frac{\partial F_2}{\partial b} + \frac{\partial F_3}{\partial b}$
0	0	0	0
-	+	-	$-2t_1 + 1 - 2t_2 - 1 - 2t_3 + 1$
+	-	+	$-2t_1 - 1 - 2t_2 + 1 - 2t_3 - 1$

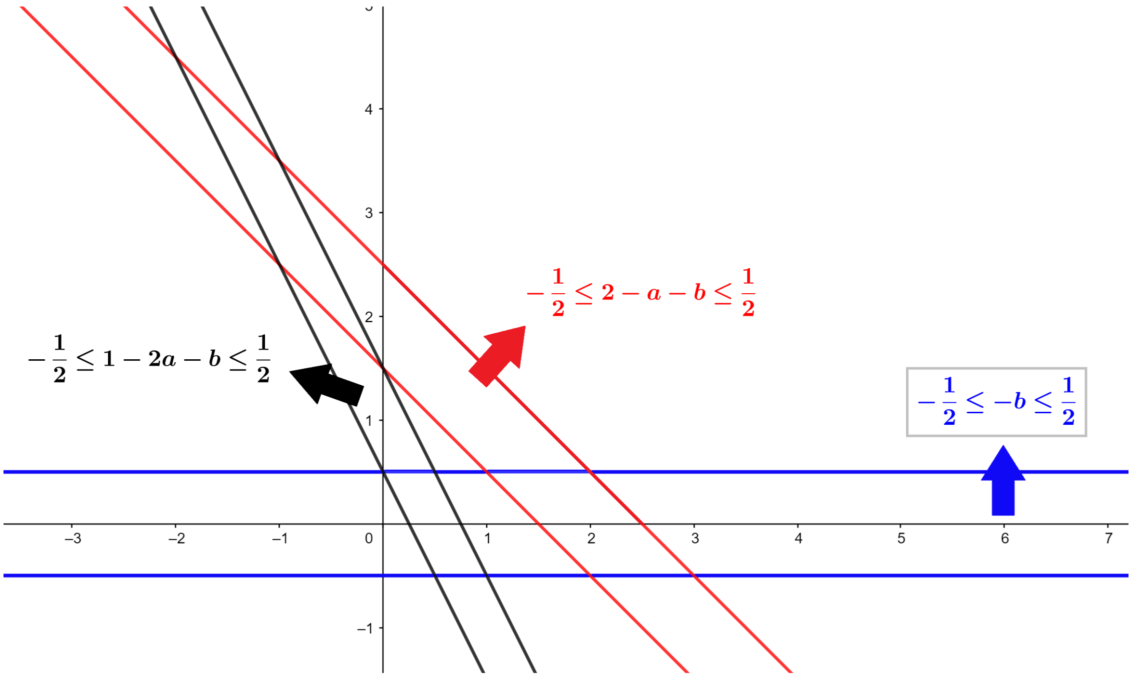


Figure 1: Explanation for situation (i)

(i) In this situation, let us suppose that there exists (a, b) such that

$$\begin{cases} -\frac{1}{2} \le t_1 \le \frac{1}{2} \\ -\frac{1}{2} \le t_2 \le \frac{1}{2} \\ -\frac{1}{2} \le t_3 \le \frac{1}{2} \end{cases} = \begin{cases} -\frac{1}{2} \le -b \le \frac{1}{2} \\ -\frac{1}{2} \le 2 - a - b \le \frac{1}{2} \\ -\frac{1}{2} \le 1 - 2a - b \le \frac{1}{2} \end{cases} \Rightarrow \begin{cases} -\frac{1}{2} \le b \le \frac{1}{2} \\ 1 \le a \le 3 \\ 0 \le a \le 1 \end{cases},$$

and it implies that $a = 1$. However, if we take $a = 1$,

$$\begin{cases} -\frac{1}{2} \le -b \le \frac{1}{2} \\ -\frac{1}{2} \le 2 - a - b \le \frac{1}{2} \\ -\frac{1}{2} \le 1 - 2a - b \le \frac{1}{2} \end{cases} = \begin{cases} -\frac{1}{2} \le -b \le \frac{1}{2} \\ -\frac{1}{2} \le 1 - b \le \frac{1}{2} \\ -\frac{1}{2} \le -1 - b \le \frac{1}{2} \end{cases},$$

which means there does not exist a b such that all of these conditions are satisfied. Hence, the optimal solution (a^*, b^*) is not in this case.
Figure 1 shows that the intersection of these three conditions

$$\begin{cases} -\frac{1}{2} \le -b \le \frac{1}{2} \\ -\frac{1}{2} \le 2 - a - b \le \frac{1}{2} \\ -\frac{1}{2} \le 1 - 2a - b \le \frac{1}{2} \end{cases},$$

is empty, which may help you realize what happens in this situation.

(ii) In this situation, we solve the linear equation of (17) and (18)

$$\begin{cases} -2t_2 - 4t_3 + 1 &= 0 \\ -2t_1 - 2t_2 - 2t_3 + 1 &= 0 \end{cases} = \begin{cases} 10a + 6b &= 7 \\ 6a + 6b &= 5 \end{cases} \Rightarrow \begin{cases} a &= \frac{1}{2} \\ b &= \frac{1}{3} \end{cases},$$

but

$$t_3 = 1 - 2 \cdot \frac{1}{2} - \frac{1}{3} = -\frac{1}{3} < \frac{1}{2}.$$

Hence, there is no solution that satisfies the conditions.

(iii) Similarly, we have the linear equation

$$\begin{cases} -2t_2 - 4t_3 - 1 & = 0 \\ -2t_1 - 2t_2 - 2t_3 - 1 & = 0 \end{cases} = \begin{cases} 10a + 6b & = 9 \\ 6a + 6b & = 7 \end{cases} \Rightarrow \begin{cases} a & = \frac{1}{2} \\ b & = \frac{2}{3} \end{cases},$$

and then confirm the solution on the conditions

$$\begin{cases} t_1 & = -\frac{2}{3} < -\frac{1}{2} \\ t_2 & = \frac{5}{6} > \frac{1}{2} \\ t_3 & = -\frac{2}{3} < -\frac{1}{2} \end{cases}$$

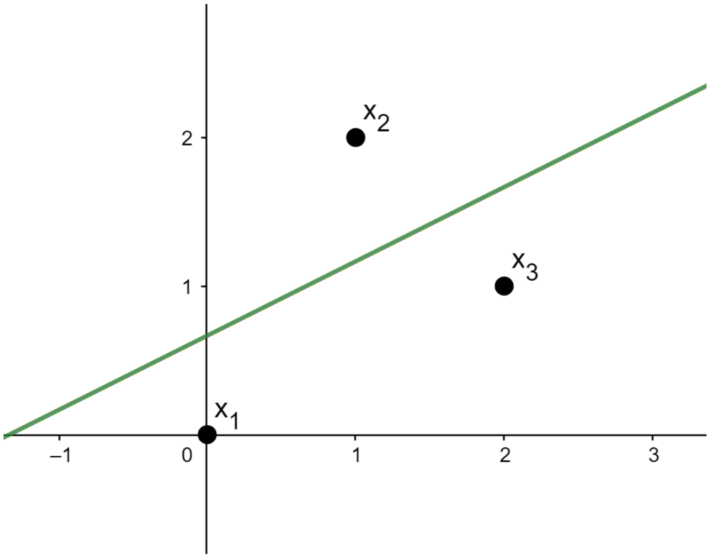
Therefore, the optimal solution (a^*, b^*) can be $(1/2, 2/3)$.

After the discussion, we have the only one optimal solution

$$(a^*, b^*) = (\frac{1}{2}, \frac{2}{3}).$$

The following figure shows the line of

$$y = ax + b$$


x1
x2
x3
−1
0
1
2
3
1
2