

# 5004 Homework 2

RONG Shuo

November 14, 2024

## Question 1:

1. For each of the following functions  $f(x_1, x_2)$ , find all critical points (i.e, all  $x_1, x_2$  such that  $\nabla f(x_1, x_2) = \mathbf{0}$ ).

- (a)  $f(x_1, x_2) = (4x_1^2 - x_2)^2$
- (b)  $f(x_1, x_2) = 2x_2^3 - 6x_2^2 + 3x_1^2x_2$
- (c)  $f(x_1, x_2) = (x_1 - 2x_2)^4 + 64x_1x_2$
- (d)  $f(x_1, x_2) = x_1^2 + 4x_1x_2 + x_2^2 + x_1 - x_2$

## Answer :

(a)

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= 2(4x_1^2 - x_2)(8x_1) = 16x_1(4x_1^2 - x_2) \\ \frac{\partial f}{\partial x_2} &= 2(4x_1^2 - x_2)(-1) = -2(4x_1^2 - x_2)\end{aligned}$$

set the gradient to 0:

$$16x_1(4x_1^2 - x_2) = 0 \tag{1}$$

$$-2(4x_1^2 - x_2) = 0 \tag{2}$$

if  $x_1 = 0$ , from (2) we can get that  $x_2 = 0$

if  $x_1 \neq 0$ , from equation (1), we know:

$$4x_1^2 = x_2$$

this satisfied  $(x_1, x_2) = (0, 0)$  Thus, we can conclude that the critical points are:

$$(x_1, x_2) = (x_1, 4x_1^2), \forall x_1 \in \mathbb{R}.$$

(b)

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= 6x_1x_2 \\ \frac{\partial f}{\partial x_2} &= 6x_2^2 - 12x_2 + 3x_1^2\end{aligned}$$

set the gradient to 0:

$$\begin{aligned}6x_1x_2 &= 0 \\6x_2^2 - 12x_2 + 3x_1^2 &= 0\end{aligned}$$

if  $x_1 = 0$ ,

$$\begin{aligned}6x_2^2 - 12x_2 &= 0 \\6x_2(x_2 - 2) &= 0\end{aligned}$$

we can conclude that  $(x_1, x_2) = (0, 0)$ , or  $(x_1, x_2) = (0, 2)$ .  
if  $x_2 = 0$ ,

$$\begin{aligned}3x_1^2 &= 0 \\x_1 &= 0\end{aligned}$$

This gives  $(x_1, x_2) = (0, 0)$

In conclusion, the critical points is  $(0, 0)$  and  $(0, 2)$

(c)

$$\begin{aligned}\frac{f}{\partial x_1} &= 4(x_1 - 2x_2)^3 + 64x_2 \\ \frac{f}{\partial x_2} &= -8(x_1 - 2x_2)^3 + 64x_1\end{aligned}$$

set the gradient to 0:

$$\begin{aligned}4(x_1 - 2x_2)^3 + 64x_2 &= 0 \\ \text{i.e. } (x_1 - 2x_2)^2 &= -16x_2 \\ -8(x_1 - 2x_2)^3 + 64x_1 &= 0 \\ \text{i.e. } (x_1 - 2x_2)^3 &= 8x_1\end{aligned}$$

$$\begin{aligned}(x_1 - 2x_2)^3 &= -16x_2 \\ (x_1 - 2x_2)^3 &= 8x_1 \\ -16x_2 &= 8x_1 \\ -2x_2 &= x_1\end{aligned}$$

Substituting  $-2x_2 = x_1$  to  $(x_1 - 2x_2)^2 = -16x_2$ , we can get:

$$\begin{aligned}64x_2^3 + 16x_2 &= 0 \\ x_2^2 &= \frac{1}{4}\end{aligned}$$

Thus, the result is

$$\begin{aligned}(x_1, x_2) &= (-1, \frac{1}{2}) \\ (x_1, x_2) &= (1, -\frac{1}{2})\end{aligned}$$

(d)

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= 2x_1 + 4x_2 + 1 \\ \frac{\partial f}{\partial x_2} &= 4x_1 + 2x_2 - 1\end{aligned}$$

set the gradient to 0:

$$\begin{aligned}2x_1 + 4x_2 + 1 &= 0 \\ 4x_1 + 2x_2 - 1 &= 0\end{aligned}$$

$$x_1 = -\frac{1}{2} - 2x_2$$

Substituting this to second equation.

$$\begin{aligned}4\left(-\frac{1}{2} - 2x_2\right) + 2x_2 - 1 &= 0 \\ -2 - 8x_2 + 2x_2 - 1 &= 0 \\ x_2 &= -\frac{1}{2} \\ x_1 &= -\frac{1}{2} + 1 = \frac{1}{2}\end{aligned}$$

Thus, the critical point is  $\left(\frac{1}{2}, -\frac{1}{2}\right)$

## Question 2:

2. Find the gradient of the following functions, where the space  $\mathbb{R}$  and  $\mathbb{R}^{n \times n}$  are equipped with the standard inner product.

(a)  $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ , where  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^m$ , and  $(\lambda > 0)$  are given.

(b)  $f(\mathbf{X}) = \mathbf{b}^T \mathbf{X} \mathbf{c}$ , where  $\mathbf{X} \in \mathbb{R}^{n \times n}$  and  $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$

(c)  $f(\mathbf{X}) = \mathbf{b} \mathbf{X}^T \mathbf{X} \mathbf{c}$ , where  $\mathbf{X} \in \mathbb{R}^{n \times n}$  and  $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$

**Answer :**

(a)

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{Ay} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{y}\|_2^2$$

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{Ay} - \mathbf{Ax} + \mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{y} - \mathbf{x} + \mathbf{x}\|_2^2$$

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{A}(\mathbf{y} - \mathbf{x}) + \mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda\|(\mathbf{y} - \mathbf{x}) + \mathbf{x}\|_2^2$$

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_2^2 + \langle \mathbf{A}(\mathbf{y} - \mathbf{x}), \mathbf{Ax} - \mathbf{b} \rangle + \frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{y} - \mathbf{x}\|_2^2 + 2\lambda\langle \mathbf{y} - \mathbf{x}, \mathbf{x} \rangle + \lambda\|\mathbf{x}\|_2^2$$

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_2^2 + \langle (\mathbf{y} - \mathbf{x}), \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) \rangle + \frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{y} - \mathbf{x}\|_2^2 + 2\lambda\langle \mathbf{y} - \mathbf{x}, \mathbf{x} \rangle + \lambda\|\mathbf{x}\|_2^2$$

$$f(\mathbf{y}) = \frac{1}{2}\|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_2^2 + \langle (\mathbf{y} - \mathbf{x}), \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) \rangle + f(\mathbf{x}) + \lambda\|\mathbf{y} - \mathbf{x}\|_2^2 + 2\lambda\langle \mathbf{y} - \mathbf{x}, \mathbf{x} \rangle$$

$$\begin{aligned}
& \lim_{\|\mathbf{y}-\mathbf{x}\|_2 \rightarrow 0} \frac{|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}), \mathbf{y} - \mathbf{x} \rangle + 2\lambda \langle \mathbf{y} - \mathbf{x}, \mathbf{x} \rangle)|}{\|\mathbf{y} - \mathbf{x}\|_2} \\
&= \lim_{\|\mathbf{y}-\mathbf{x}\|_2 \rightarrow 0} \frac{\frac{1}{2} \|\mathbf{A}(\mathbf{y} - \mathbf{x})\|_2^2 + 2\lambda \|\mathbf{y} - \mathbf{x}\|_2^2}{\|\mathbf{y} - \mathbf{x}\|_2} \\
&\leq \lim_{\|\mathbf{y}-\mathbf{x}\|_2 \rightarrow 0} \frac{\frac{1}{2} \|\mathbf{A}\|_2^2 \|\mathbf{y} - \mathbf{x}\|_2^2 + 2\lambda \|\mathbf{y} - \mathbf{x}\|_2^2}{\|\mathbf{y} - \mathbf{x}\|_2} \\
&= \lim_{\|\mathbf{y}-\mathbf{x}\|_2 \rightarrow 0} \frac{1}{2} \|\mathbf{A}\|_2^2 \|\mathbf{y} - \mathbf{x}\|_2 + 2\lambda \|\mathbf{y} - \mathbf{x}\|_2 \\
&= 0
\end{aligned}$$

According

$$f(\mathbf{y}) - (f(\mathbf{x}) + \langle \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}), \mathbf{y} - \mathbf{x} \rangle + 2\lambda \langle \mathbf{y} - \mathbf{x}, \mathbf{x} \rangle)$$

In conclusion, the gradient for (a) is  $\nabla f(\mathbf{x}) = \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) + 2\lambda \mathbf{x}$

(b) We know the definition of the  $(\nabla_{\mathbf{X}} f(\mathbf{X}))_{ij} = \frac{\partial f}{\partial X_{ij}}$

And we know that  $f(\mathbf{X})$

$$\begin{aligned}
f(\mathbf{X}) &= \sum_{i=1}^n \sum_{j=1}^n b_i X_{ij} c_j \\
\frac{\partial f}{\partial X_{ij}} &= b_i c_j \\
\nabla_{\mathbf{X}} f(\mathbf{X}) &= \begin{bmatrix} b_1 c_1 & b_1 c_2 & \cdots & b_1 c_n \\ b_2 c_1 & b_2 c_2 & \cdots & b_2 c_n \\ \vdots & \vdots & \vdots & \vdots \\ b_n c_1 & b_n c_2 & \cdots & b_n c_n \end{bmatrix} \\
\nabla_{\mathbf{X}} f(\mathbf{X}) &= \mathbf{bc}^T
\end{aligned}$$

In conclusion, the gradient of  $f(\mathbf{X}) = \mathbf{b}^T \mathbf{X} \mathbf{c}$  is  $\nabla_{\mathbf{X}} f(\mathbf{X}) = \mathbf{bc}^T$

(c) We consider

$$\begin{aligned}
f(\mathbf{X}) &= \sum_{i=1}^n \sum_{j=1}^n b_i X_{ji} X_{ij} c_j \\
\frac{\partial f}{\partial X_{ij}} &= b_i X_{ji} c_j + b_j X_{ji} c_i \\
\nabla_{\mathbf{X}} f(\mathbf{X}) &= \begin{bmatrix} b_1 \mathbf{X}_{11} c_1 & b_1 \mathbf{X}_{21} c_2 & \cdots & b_1 \mathbf{X}_{n1} c_n \\ b_2 \mathbf{X}_{12} c_1 & b_2 \mathbf{X}_{22} c_2 & \cdots & b_2 \mathbf{X}_{n2} c_n \\ \vdots & \vdots & \vdots & \vdots \\ b_n \mathbf{X}_{1n} c_1 & b_n \mathbf{X}_{2n} c_2 & \cdots & b_n \mathbf{X}_{nn} c_n \end{bmatrix} + \begin{bmatrix} b_1 \mathbf{X}_{11} c_1 & b_2 \mathbf{X}_{21} c_1 & \cdots & b_n \mathbf{X}_{n1} c_1 \\ b_1 \mathbf{X}_{12} c_2 & b_2 \mathbf{X}_{22} c_2 & \cdots & b_n \mathbf{X}_{n2} c_2 \\ \vdots & \vdots & \vdots & \vdots \\ b_1 \mathbf{X}_{1n} c_n & b_2 \mathbf{X}_{2n} c_n & \cdots & b_n \mathbf{X}_{nn} c_n \end{bmatrix} \\
\nabla_{\mathbf{X}} f(\mathbf{X}) &= \mathbf{X} \mathbf{b}^T \mathbf{c} + \mathbf{X} \mathbf{bc}^T
\end{aligned}$$

### Question 3:

3. Let  $\{\mathbf{x}_i, y_i\}_{i=1}^N$  be given with  $\mathbf{x}_i \in \mathbb{R}$  and  $y_i \in \mathbb{R}$ . Assume  $N < n$ . Consider the ridge regression

$$\text{minimize}_{\mathbf{a} \in \mathbb{R}^n} \sum_{i=1}^N (\langle \mathbf{a}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{a}\|_2^2,$$

where  $\lambda \in \mathbb{R}$  is a regularization parameter, and we set the bias  $b = 0$  for simplicity.

(a) Prove that the solution must be in the form of  $\mathbf{a} = \sum_{i=1}^N c_i \mathbf{x}_i$  for some  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^N$ .

(hint: similar to the proof of the representer theorem.)

(b) Re-express the minimization in terms of  $\mathbf{c} \in \mathbb{R}^N$ , which has fewer unknowns than the original formulation as  $N < n$ .

### Answer :

(a) We can denote  $\mathbf{a} = \mathbf{a}_s + \sum_{i=1}^N c_i \mathbf{x}_i$ , where  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^N$ , and  $\langle \mathbf{a}_s, \mathbf{x}_i \rangle = 0$ .

$$\begin{aligned} \sum_{i=1}^N (\langle \mathbf{a}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{a}\|_2^2 &= \sum_{i=1}^N (\langle \mathbf{a}_s + \sum_{j=1}^N c_j \mathbf{x}_j, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{a}_s + \sum_{j=1}^N c_j \mathbf{x}_j\|_2^2 \\ &= \sum_{i=1}^N (\sum_{j=1}^N c_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j_1=1}^N \sum_{j_2=1}^N c_{j_1} c_{j_2} \langle \mathbf{x}_{j_1}, \mathbf{x}_{j_2} \rangle + \lambda \|\mathbf{a}_s\|_2^2 \end{aligned}$$

Introduce  $\mathbf{K} = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ .

$$\begin{aligned} &\sum_{i=1}^N (\sum_{j=1}^N c_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j_1=1}^N \sum_{j_2=1}^N c_{j_1} c_{j_2} \langle \mathbf{x}_{j_1}, \mathbf{x}_{j_2} \rangle + \lambda \|\mathbf{a}_s\|_2^2 \\ &= \sum_{i=1}^N ((\mathbf{K}^T \mathbf{c})_i - y_i)^2 + \lambda \mathbf{c}^T \mathbf{K} \mathbf{c} + \lambda \|\mathbf{a}_s\|_2^2 \end{aligned}$$

We know:

$$\begin{aligned} &\text{minimize}_{\mathbf{a} \in \mathbb{R}^n} \sum_{i=1}^N ((\mathbf{K}^T \mathbf{c})_i - y_i)^2 + \lambda \mathbf{c}^T \mathbf{K} \mathbf{c} + \lambda \|\mathbf{a}_s\|_2^2 \\ &\iff \text{minimize}_{\mathbf{c} \in \mathbb{R}^N} \sum_{i=1}^N ((\mathbf{K}^T \mathbf{c})_i - y_i)^2 + \lambda \mathbf{c}^T \mathbf{K} \mathbf{c} \\ &\text{and minimize}_{\mathbf{a}_s \in \mathbb{R}^n, \langle \mathbf{a}_s, \mathbf{x}_i \rangle = 0} \lambda \|\mathbf{a}_s\|_2^2 \\ &\iff \mathbf{a}_s^* = \mathbf{0} \\ &\text{and } \mathbf{c}^* = \text{argmin}_{\mathbf{c} \in \mathbb{R}^N} \sum_{i=1}^N ((\mathbf{K}^T \mathbf{c})_i - y_i)^2 + \lambda \mathbf{c}^T \mathbf{K} \mathbf{c} \end{aligned}$$

In conclusion, the optimal  $\mathbf{a}^* = \sum_{i=1}^N c_i^* \mathbf{x}_i$

(b) The re-express formulation show below:

$$\text{minimize}_{\mathbf{c} \in \mathbb{R}^N} \sum_{i=1}^N ((\mathbf{K}^T \mathbf{c})_i - y_i)^2 + \lambda \mathbf{c}^T \mathbf{K} \mathbf{c}$$

## Question 4:

4. Let  $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{x} + c$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is a symmetric positive semidefinite matrix,  $\mathbf{b} \in \mathbb{R}^n$ , and  $c \in \mathbb{R}$ .

(a) Prove that  $\mathbf{x}$  is a global minimizer of  $f$  if and only if  $\mathbf{A}\mathbf{x} = -\mathbf{b}$ .

(b) Prove that  $f$  is bounded below over  $\mathbb{R}^n$  if and only if  $\mathbf{b} \in \{\mathbf{A}\mathbf{y} : \mathbf{y} \in \mathbb{R}^n\}$ .

## Answer :

(a) We know that:

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n x_i A_{ij} x_j \\ \frac{\partial f}{\partial x_k} &= \sum_{j=1}^n A_{kj} x_j + \sum_{i=1}^n x_i A_{ik} \\ \nabla f(\mathbf{x}) &= \mathbf{A} \mathbf{x} + \mathbf{A}^T \mathbf{x} \\ \nabla f(\mathbf{x}) &= 2\mathbf{A} \mathbf{x} \end{aligned}$$

So we can get the gradient of  $f(\mathbf{x})$ :

$$\nabla f(\mathbf{x}) = 2\mathbf{A} \mathbf{x} + 2\mathbf{b}$$

To prove that  $f(\mathbf{x})$  is convex, we have:

$$\begin{aligned} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &= \\ &= \mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{x} + c + 2\mathbf{x}^T \mathbf{A}(\mathbf{y} - \mathbf{x}) + 2\mathbf{b}^T(\mathbf{y} - \mathbf{x}) \\ &= \mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{x} + c + 2\mathbf{x}^T \mathbf{A} \mathbf{y} - 2\mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{y} - 2\mathbf{b}^T \mathbf{x} \\ &= \mathbf{x}^T \mathbf{A} \mathbf{x} + c + 2\mathbf{x}^T \mathbf{A} \mathbf{y} - 2\mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{y} \\ &= c + 2\mathbf{x}^T \mathbf{A} \mathbf{y} - \mathbf{x}^T \mathbf{A} \mathbf{x} + 2\mathbf{b}^T \mathbf{y} \end{aligned}$$

$$\begin{aligned} f(\mathbf{y}) - (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle) &= \\ &= \mathbf{y}^T \mathbf{A} \mathbf{y} + 2\mathbf{b}^T \mathbf{y} + c - c - 2\mathbf{x}^T \mathbf{A} \mathbf{y} + \mathbf{x}^T \mathbf{A} \mathbf{x} - 2\mathbf{b}^T \mathbf{y} \\ &= \mathbf{y}^T \mathbf{A} \mathbf{y} - 2\mathbf{x}^T \mathbf{A} \mathbf{y} + \mathbf{x}^T \mathbf{A} \mathbf{x} \end{aligned}$$

We know:

$$\begin{aligned} (\mathbf{x}^T \mathbf{A} \mathbf{y})^T &= \mathbf{y}^T (\mathbf{x}^T \mathbf{A})^T = \mathbf{y}^T (\mathbf{A}^T \mathbf{x}) = \mathbf{y}^T \mathbf{A} \mathbf{x} \\ \mathbf{x}^T \mathbf{A} \mathbf{y} &= \langle \mathbf{A} \mathbf{x}, \mathbf{y} \rangle \\ \mathbf{y}^T \mathbf{A} \mathbf{x} &= \langle \mathbf{A} \mathbf{y}, \mathbf{x} \rangle \end{aligned}$$

We know the transpose of a scalar is itself, so

$$\mathbf{x}^T \mathbf{A} \mathbf{y} = \mathbf{y}^T \mathbf{A} \mathbf{x}$$

Continue, we have

$$\begin{aligned}
&= \mathbf{y}^T \mathbf{A} \mathbf{y} - 2\mathbf{x}^T \mathbf{A} \mathbf{y} + \mathbf{x}^T \mathbf{A} \mathbf{x} \\
&= \mathbf{y}^T \mathbf{A} \mathbf{y} - \mathbf{x}^T \mathbf{A} \mathbf{y} - \mathbf{y}^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{A} \mathbf{x} \\
&= \langle \mathbf{A} \mathbf{y}, \mathbf{y} - \mathbf{x} \rangle - \langle \mathbf{A} \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \\
&= \langle \mathbf{A} \mathbf{y} - \mathbf{A} \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \geq 0
\end{aligned}$$

So we in conclusion, we have  $f(\mathbf{y}) \geq (f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)$ , which implies that  $f(\mathbf{x})$  is convex. So  $\nabla f(\mathbf{x}^*) = 0 \iff \mathbf{x}^*$  is the global minimizer. i.e.

$\mathbf{x}$  is a global minimizer  $\iff \mathbf{A} \mathbf{x} = -\mathbf{b}$

(b) Proof:  $f$  is bounded below  $\implies \mathbf{b} \in \{\mathbf{A} \mathbf{y}, \mathbf{y} \in \mathbb{R}^n\}$ .

(i) If  $\mathbf{A}$  is full rank, then  $\mathbf{b} \in \mathbf{A} \mathbf{y}, \mathbf{y} \in \mathbb{R}^n$  is easy to prove, since  $\mathbf{A} \mathbf{y} \equiv \mathbb{R}^n$

(ii) If  $\mathbf{A}$  is not full rank, then there exists a vector  $\mathbf{d} \neq \mathbf{0}$ , s.t.  $\mathbf{A} \mathbf{d} = \mathbf{0}$

we can choose  $\mathbf{b} = \mathbf{d}$

$$\begin{aligned}
f(-\mathbf{d}) &= \mathbf{d}^T \mathbf{A} \mathbf{d} - 2\mathbf{d}^T \mathbf{d} + c \\
f(-\mathbf{d}) &= -2\mathbf{d}^T \mathbf{d} + c \\
f(-\mathbf{d}) &= -2\|\mathbf{d}\|^2 + c
\end{aligned}$$

We know that  $\mathbf{d}^T \mathbf{d} > 0$ , since  $\mathbf{d} \neq \mathbf{0}$

$$\lim_{\|\mathbf{d}\|^2 \rightarrow \infty} -2\|\mathbf{d}\|^2 + c = -\infty$$

s.t. if  $f$  is bounded below, then  $\mathbf{A}$  must be full rank, which implies  $\mathbf{b} \in \mathbf{A} \mathbf{y}$

Proof:  $\mathbf{b} \in \{\mathbf{A} \mathbf{y}, \mathbf{y} \in \mathbb{R}^n\} \implies f$  is bounded below.

According to (a), we know that  $\nabla f(\mathbf{x}^*) = 0 \iff \mathbf{x}^*$  is the global minimizer.

$$\begin{aligned}
\nabla f(\mathbf{x}) &= 2\mathbf{A} \mathbf{x} + 2\mathbf{b} \\
\mathbf{b} &\in \mathbf{A} \mathbf{y}, \mathbf{y} \in \mathbb{R}^n \\
\exists \mathbf{x} \in \mathbb{R}^n, \text{ s.t. } \mathbf{A} \mathbf{x} &= -\mathbf{b} \\
\text{i.e. } \exists \mathbf{x}^* \in \mathbb{R}^n, \nabla f(\mathbf{x}^*) &= 0
\end{aligned}$$

$f(\mathbf{x}) \geq f(\mathbf{x}^*) \implies f$  is bounded below.

## Question 5:

5. We consider the following optimization problem:

$$\text{minimize}_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \log \left( \sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \right) \quad (3)$$

where  $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$  and  $b_1, \dots, b_m \in \mathbb{R}$  are given.

(a) Find the gradient of  $f(\mathbf{x})$ .

(b) If we use gradient descent to solve Problem (1), will it converge to the global minimizer? Please justify your answer.

## Answer

(a)

$$\begin{aligned}g(\mathbf{x}) &= \sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \\ \nabla f(\mathbf{x}) &= \frac{1}{g(\mathbf{x})} \nabla g(\mathbf{x}) \\ g(\mathbf{x}) &= \sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \\ \nabla g(\mathbf{x}) &= \sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{a}_i \\ \nabla f(\mathbf{x}) &= \frac{1}{\sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i)} \sum_{i=1}^m \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{a}_i\end{aligned}$$

(b) We can use Jensen's Inequality to prove this function is convex.

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

take  $t = 1/2$

$$f(t\mathbf{x} + (1-t)\mathbf{y}) = \log \left( \sum_{i=1}^n \exp\left(\frac{1}{2}(\mathbf{a}_i \mathbf{x} + b_i + \mathbf{a}_i \mathbf{y} + b_i)\right) \right)$$

for the right hand side, we know:

$$\begin{aligned}\text{RHS} &= \frac{1}{2} \log\left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{x} + b_i)\right) + \frac{1}{2} \log\left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{y} + b_i)\right) \\ &= \frac{1}{2} \log\left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{x} + b_i) \sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{y} + b_i)\right) \\ &= \log\left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{x} + b_i)^{1/2} \sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{y} + b_i)^{1/2}\right)\end{aligned}$$

According to CS-inequality:

$$\begin{aligned}\sum_{i=1}^n \exp\left(\frac{1}{2}(\mathbf{a}_i \mathbf{x} + b_i)\right) \exp\left(\frac{1}{2}(\mathbf{a}_i \mathbf{y} + b_i)\right) &\leq \left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{x} + b_i)\right)^{1/2} \left(\sum_{i=1}^n \exp(\mathbf{a}_i \mathbf{y} + b_i)\right)^{1/2} \\ f(1/2\mathbf{x} + 1/2\mathbf{y}) &\leq 1/2f(\mathbf{x}) + 1/2f(\mathbf{y})\end{aligned}$$

In conclusion,  $f(\mathbf{x})$  is midpoint convex, which is equivalent to the convexity if the function is continuous.

Since  $f(\mathbf{x})$  is convex, any local minimizer is also a global minimum. Gradient descent is guaranteed to converge to a global minimizer.