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

Answer :

(a)

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

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)

$$\frac{\partial f}{\partial x_1} = 6x_1 x_2$$
$$\frac{\partial f}{\partial x_2} = 6x_2^2 - 12x_2 + 3x_1^2$$

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)

$$\frac{\partial f}{\partial x_1} = 2x_1 + 4x_2 + 1$$
$$\frac{\partial f}{\partial x_2} = 4x_1 + 2x_2 - 1$$

set the gradient to 0:

$$2x_1 + 4x_2 + 1 = 0$$
$$4x_1 + 2x_2 - 1 = 0$$

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

Substituting this to second equation.

$$4(-\frac{1}{2} - 2x_2) + 2x_2 - 1 = 0$$
$$-2 - 8x_2 + 2x_2 - 1 = 0$$
$$x_2 = -\frac{1}{2}$$
$$x_1 = -\frac{1}{2} + 1 = \frac{1}{2}$$

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

Answer :

(a)

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

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

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

(b) We know the definition of the $(\nabla_{\boldsymbol{X}} f(\boldsymbol{X}))_{ij} = \frac{\partial f}{\partial \boldsymbol{X}_{ij}}$
And we know that $f(\boldsymbol{X})$

$$
\begin{aligned}
f(\boldsymbol{X}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i X_{ij} c_j \\
\frac{\partial f}{\partial X_{ij}} &= b_i c_j \\
\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) &= \begin{bmatrix} b_1c_1 & b_1c_2 & \cdots & b_1c_n \\ b_2c_1 & b_2c_2 & \cdots & b_2c_n \\ \vdots & \vdots & \vdots & \vdots \\ b_nc_1 & b_nc_2 & \cdots & b_nc_n \end{bmatrix} \\
\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) &= \boldsymbol{b}\boldsymbol{c}^T
\end{aligned}
$$

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

(c) We consider

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

Question 3:

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

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

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 $\boldsymbol{a} = \sum_{i=1}^{N} c_i \boldsymbol{x}_i$ for some $\boldsymbol{c} = [c_1, c_2, \cdots, c_N]^T \in \mathbb{R}^N$.
(*hint: similar to the proof of the representer theorem.*)
(b) Re-express the minimization in terms of $\boldsymbol{c} \in \mathbb{R}^N$, which has fewer unknowns than the original formulation as $N < n$.

Answer :

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

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

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

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

We know:

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

In conclusion, the optimal $\boldsymbol{a}^* = \sum_{i=1}^{N} c_i^* \boldsymbol{x}_i$
(b) The re-express formulation show below:

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

Question 4:

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

(a) Prove that $\boldsymbol{x}$ is a global minimizer of f if and only if $\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b}$.
(b) Prove that f is bounded below over $\mathbb{R}^n$ if and only if $\boldsymbol{b} \in \{\boldsymbol{A}\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^n\}$.

Answer :

(a) We know that:

$$\begin{aligned}
f(\boldsymbol{x}) &= \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j \\
\frac{\partial f}{\partial x_i} &= a_{ij}x_j + a_{ji}x_j = 2a_{ij}x_j \\
\nabla f(\boldsymbol{x}) &= 2\boldsymbol{A}\boldsymbol{x}
\end{aligned}$$

So we can get the gradient of f(x):

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

To prove that f(x) is convex, we have:

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

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

We know:

$$\begin{aligned}
(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y})^T = \boldsymbol{y}^T(\boldsymbol{x}^T\boldsymbol{A})^T = \boldsymbol{y}^T(\boldsymbol{A}^T\boldsymbol{x}) &= \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} \\
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y} &= \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle \\
\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} &= \langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{x} \rangle \\
\text{We know the transpose of a scalar is itself, so}& \\
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y} &= \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}
\end{aligned}$$

Continue, we have

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

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

(b) Proof: f is bounded below $\implies \boldsymbol{b} \in \{\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y} \in \mathbb{R}^n\}$.
Assume $\boldsymbol{b} \notin \boldsymbol{A}\boldsymbol{y}$, which $implies \nabla f(\boldsymbol{x}) \neq \mathbf{0}$

Proof: $\boldsymbol{b} \in \{\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y} \in \mathbb{R}^n\} \implies f$ is bounded below.
According to (a), we know that $\nabla f(\boldsymbol{x}*) = 0 \iff \boldsymbol{x}*$ is the global minimizer.

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

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

Question 5:

5. We consider the following optimization problem:

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

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

(a) Find the gradient of $f(\boldsymbol{x})$.
(b) If we use gradient descent to solve Problem (1), will it converge to the global minimizer? Please justify your answer.

Answer

(a)

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

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

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

take $t = 1/2$

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

for the right hand side, we know:

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

According to CS-inequality:

$$\sum_{i=1}^{n} \exp(\frac{1}{2}(\boldsymbol{a}_i\boldsymbol{x} + b_i))\exp(\frac{1}{2}(\boldsymbol{a}_i\boldsymbol{y} + b_i)) \leq \left(\sum_{i=1}^{n} \exp(\boldsymbol{a}_i\boldsymbol{x} + b_i)\right)^{1/2} \left(\sum_{i=1}^{n} \exp(\boldsymbol{a}_i\boldsymbol{y} + b_i)\right)^{1/2}$$

$$f(1/2\boldsymbol{x} + 1/2\boldsymbol{y}) \leq 1/2f(\boldsymbol{x}) + 1/2f(\boldsymbol{y})$$

In conclusion,$f(\boldsymbol{x})$ is midpoint convex, which is equivalent to the convexity if the function is continuous.
Since $f(\boldsymbol{x})$ is convex, any local minimizer is also a global minimum. Gradient descent is guaranteed to converge to a global minimizer.