

MSBD 5007 HW2

RONG Shuo

March 30, 2025

Question1

Determine the convexity of the following functions, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{S}_{++}^n$ (the set of symmetric positive definite matrices). Justify your answer.

(a) $f(\mathbf{x}) = \log(e^{x_1} + e^{x_2} + \dots + e^{x_n})$.

(b) $f(\mathbf{X}) = \log \det(\mathbf{X})$.

Answer

(a)

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$$

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \frac{1}{e^{x_1} + e^{x_2} + \dots + e^{x_n}} \times \frac{\partial}{\partial x_i} (e^{x_1} + e^{x_2} + \dots + e^{x_n}) \\ &= \frac{e^{x_i}}{e^{x_1} + e^{x_2} + \dots + e^{x_n}} \end{aligned}$$

Therefore,

$$\nabla f(\mathbf{x}) = \left(\frac{e^{x_1}}{\sum_{i=1}^n e^{x_i}}, \frac{e^{x_2}}{\sum_{i=1}^n e^{x_i}}, \dots, \frac{e^{x_n}}{\sum_{i=1}^n e^{x_i}} \right)$$

We know,

$$\begin{aligned} H_{ij} &= \frac{\partial^2 f}{\partial x_i \partial x_j} \\ H_{ii} &= \frac{\partial}{\partial x_i} \left(\frac{e^{x_i}}{S} \right) = \frac{e^{x_i}(S - e^{x_i})}{S^2} \text{ if } i = j \\ H_{ij} &= -\frac{e^{x_i} e^{x_j}}{S^2} \text{ if } i \neq j \\ \nabla^2 f(\mathbf{x}) &= \frac{1}{S} \text{diag}(\mathbf{e}) - \frac{1}{S^2} \mathbf{e} \mathbf{e}^T \end{aligned}$$

where $\mathbf{e} = (e^{x_1}, e^{x_2}, \dots, e^{x_n})^T$, and $S = \sum_{i=1}^n e^{x_i}$

$$\begin{aligned} \mathbf{z} \nabla^2 f(\mathbf{x}) \mathbf{z}^T &= \frac{1}{S} \sum_{i=1}^n e^{x_i} z_i^2 - \frac{1}{S^2} \left(\sum_{i=1}^n e^{x_i} z_i \right)^2 \\ \left(\sum_{i=1}^n e^{x_i} z_i \right)^2 &\leq \left(\sum_{i=1}^n e^{x_i} z_i^2 \right) \left(\sum_{i=1}^n e^{x_i} \right) \\ \frac{1}{S} \sum_{i=1}^n e^{x_i} z_i^2 &\geq \frac{1}{S^2} \left(\sum_{i=1}^n e^{x_i} z_i \right)^2 \\ \mathbf{z} \nabla^2 f(\mathbf{x}) \mathbf{z}^T &\geq 0 \end{aligned}$$

Thus, we prove that Hessian is positive semi-definite, so we can conclude $f(\mathbf{x})$ is convex.

(b)

Let $\mathbf{X} \in \mathbb{S}_{++}^n$ and $\mathbf{V} \in \mathbb{S}^n$, we define $g(t) = \log \det(\mathbf{X} + t\mathbf{V})$, where $\mathbf{X} + t\mathbf{V}$ is symmetric, positive and definite.

$$\begin{aligned} g(t) &= \log \det(\mathbf{X} + t\mathbf{V}) \\ &= \log \det(\mathbf{X}^{\frac{1}{2}}(\mathbf{I} + t\mathbf{X}^{-\frac{1}{2}}\mathbf{V}\mathbf{X}^{-\frac{1}{2}})\mathbf{X}^{\frac{1}{2}}) \\ &= \log(\det \mathbf{X}^{\frac{1}{2}} \cdot \det(\mathbf{I} + t\mathbf{X}^{-\frac{1}{2}}\mathbf{V}\mathbf{X}^{-\frac{1}{2}}) \cdot \det \mathbf{X}^{\frac{1}{2}}) \\ &= \log \det \mathbf{X} + \log \det(\mathbf{I} + t\mathbf{\Lambda}) \\ &= \log \det \mathbf{X} + \sum_{i=1}^n \log(1 + t\lambda_i) \end{aligned}$$

where λ_i is the eigenvalues of $\mathbf{\Lambda}$, and $\mathbf{X}^{-\frac{1}{2}}\mathbf{V}\mathbf{X}^{-\frac{1}{2}} = \mathbf{\Lambda}$

$$g''(t) = \sum_{i=1}^n \frac{-\lambda_i^2}{(1 + t\lambda_i)^2} \leq 0$$

Thus $g''(t) \leq 0$ for all t where $\mathbf{X} + t\mathbf{V}$ is symmetric, positive and definite, so $g(t)$ is concave. Since $g(t)$ is concave for any direction, $f(\mathbf{X}) = \log \det \mathbf{X}$ is concave on \mathbb{S}_{++}^n .

Let's denote $\mathbf{V} = (\mathbf{Y} - \mathbf{X})$. We have $g(t) = \log \det(\mathbf{X} + t(\mathbf{Y} - \mathbf{X}))$

$$g(0) = \log \det(\mathbf{X})$$

$$g(1) = \log \det(\mathbf{Y})$$

By concavity of $g(t)$, we know:

$$\begin{aligned} g(t) &\geq (1-t)g(0) + tg(1) \\ \log \det(\mathbf{X} + t(\mathbf{Y} - \mathbf{X})) &\geq (1-t)\log \det \mathbf{X} + t\log \det \mathbf{Y} \\ \log \det(t\mathbf{Y} + (1-t)\mathbf{X}) &\geq t\log \det \mathbf{Y} + (1-t)\log \det \mathbf{X} \end{aligned}$$

Question2

Consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, with $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$ and $\mathbf{b} \in \begin{bmatrix} 2 \\ -4 \end{bmatrix}$, and the initial guess $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

- (a) Present the first two update iterations using the **steepest descent algorithm**.
- (b) Present the first two updated iterations using the **conjugate gradient algorithm**.

Answer

(a)

We know if we want use steepest descent algorithm to solve the linear system, we need:

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_k^T \mathbf{A} \mathbf{r}_k}$$

where $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{A} \mathbf{r}_k$

$$\begin{aligned} \mathbf{r}_0 &= \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \begin{bmatrix} 2 \\ -4 \end{bmatrix} \\ \alpha_0 &= \frac{5}{14} \\ \mathbf{x}_1 &= \mathbf{x}_0 + \alpha_0 \mathbf{r}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{5}{14} \begin{bmatrix} 2 \\ -4 \end{bmatrix} = \begin{bmatrix} \frac{5}{7} \\ -\frac{10}{7} \end{bmatrix} \\ \mathbf{r}_1 &= \mathbf{b} - \mathbf{A}\mathbf{x}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \frac{5}{14} \begin{bmatrix} 2 \\ -4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \end{aligned}$$

next iteration:

$$\alpha_1 = \frac{5}{16}$$

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1 \mathbf{r}_1 = \begin{bmatrix} \frac{5}{7} \\ \frac{10}{7} \end{bmatrix} + \frac{5}{16} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{75}{56} \\ -\frac{125}{112} \end{bmatrix}$$

In conclusion, $\mathbf{x}_1 = \begin{bmatrix} \frac{5}{7} \\ \frac{10}{7} \end{bmatrix}$ $\mathbf{x}_2 = \begin{bmatrix} \frac{75}{56} \\ -\frac{125}{112} \end{bmatrix}$

(b)

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \begin{bmatrix} 2 \\ -4 \end{bmatrix}$$

$$\alpha_0 = \frac{\mathbf{r}_0^T \mathbf{r}_0}{\mathbf{r}_0^T \mathbf{A} \mathbf{r}_0} = \frac{5}{14}$$

$$\mathbf{x}_1 = \begin{bmatrix} \frac{5}{7} \\ \frac{10}{7} \end{bmatrix}$$

next iteration,

$$\mathbf{r}_1 = \mathbf{r}_0 - \alpha_0 \mathbf{A} \mathbf{r}_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\beta_0 = \frac{\mathbf{r}_1^T \mathbf{r}_1}{\mathbf{r}_0^T \mathbf{r}_0} = \frac{5}{20} = \frac{1}{4}$$

$$\mathbf{p}_1 = \mathbf{r}_1 + \beta_0 \mathbf{r}_0 = \begin{bmatrix} \frac{5}{2} \\ 0 \end{bmatrix}$$

$$\alpha_1 = \frac{\mathbf{r}_1^T \mathbf{r}_1}{\mathbf{p}_1^T \mathbf{A} \mathbf{p}_1} = \frac{2}{5}$$

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1 \mathbf{p}_1 = \begin{bmatrix} \frac{12}{7} \\ \frac{10}{7} \end{bmatrix}$$

In conclusion, $\mathbf{x}_1 = \begin{bmatrix} \frac{5}{7} \\ \frac{10}{7} \end{bmatrix}$ $\mathbf{x}_2 = \begin{bmatrix} \frac{12}{7} \\ \frac{10}{7} \end{bmatrix}$

Question3

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a twice continuously differentiable functions. Suppose that for every $\mathbf{x} \in \mathbb{R}^n$, the eigenvalues of the Hessian matrix $\nabla^2 f(\mathbf{X})$ lie uniformly in the interval $[m, M]$ with $0 < m \leq M < \infty$.

Prove that:

- (a) The function f has a unique global minimizer \mathbf{x}^* .
- (b) For all $\mathbf{x} \in \mathbb{R}^n$, the following inequality holds:

$$\frac{1}{2M} \|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2m} \|\nabla f(\mathbf{x})\|^2$$

Answer

(a)

We know the second order sufficient condition:

$$\nabla^2 f(\mathbf{x}) \succ 0 \implies f \text{ strictly convex}$$

Since all eigenvalues of $\nabla^2 f(\mathbf{x}) \geq m > 0$, we know

$$\mathbf{v} \nabla^2 f(\mathbf{x}) \mathbf{v} \geq m \|\mathbf{v}\|^2 > 0 \text{ for any non-zero vector } \mathbf{v}$$

$$\nabla^2 f(\mathbf{x}) \succ 0$$

Therefore, f is strictly convex. And we know the Theorem that for an optimization problem, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex on Ω and Ω is a convex set. Then the optimal solution must be unique.

To show f is coercive, we know

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \\ (\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) &\geq m \|\mathbf{y} - \mathbf{x}\|^2 \\ f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}m \|\mathbf{y} - \mathbf{x}\|^2 \\ f(\mathbf{y}) &\geq f(\mathbf{0}) + \nabla f(\mathbf{0})^T \mathbf{y} + \frac{m}{2} \|\mathbf{y}\|^2 \end{aligned}$$

Therefore, as $\|\mathbf{y}\| \rightarrow \infty$, $f(\mathbf{y}) \rightarrow \infty$. Hence, f is coercive. In conclusion, there exists a optimal solution(coercive), and it's unique(strictly convex).

(b)

We know

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}m \|\mathbf{y} - \mathbf{x}\|^2 \\ f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}^* - \mathbf{x}) + \frac{1}{2}m \|\mathbf{x}^* - \mathbf{x}\|^2 \\ f(\mathbf{x}) - f(\mathbf{x}^*) &\leq \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{x}^*) - \frac{1}{2}m \|\mathbf{x} - \mathbf{x}^*\|^2 \end{aligned}$$

According to Cauchy-Schwarz inequality, we know:

$$\begin{aligned} \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{x}^*) &\leq \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\| \\ f(\mathbf{x}) - f(\mathbf{x}^*) &\leq \|\nabla f(\mathbf{x})\| \|\mathbf{x} - \mathbf{x}^*\| - \frac{1}{2}m \|\mathbf{x} - \mathbf{x}^*\|^2 \end{aligned}$$

We know if $\|\mathbf{x} - \mathbf{x}^*\| = \frac{1}{m} \|\nabla f(\mathbf{x})\|$ (considering the $g(t) = \|\nabla f(\mathbf{x})\|t - \frac{1}{2}mt^2$), we get the maximum.

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \|\nabla f(\mathbf{x})\| \frac{1}{m} \|\nabla f(\mathbf{x})\| - \frac{1}{2}m \left(\frac{1}{m} \|\nabla f(\mathbf{x})\| \right)^2 = \frac{1}{2m} \|\nabla f(\mathbf{x})\|^2$$

Therefore, $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2m} \|\nabla f(\mathbf{x})\|^2$

And we know:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|^2$$

Since

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \\ (\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) &\leq M \|\mathbf{y} - \mathbf{x}\|^2 \\ f(\mathbf{y}) &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}M \|\mathbf{y} - \mathbf{x}\|^2 \end{aligned}$$

So

$$\begin{aligned}
f(\mathbf{y}) &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{M}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
\mathbf{y} &= \mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x}) \\
f(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})) &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(-\frac{1}{M}\nabla f(\mathbf{x})) + \frac{M}{2}\|\frac{1}{M}\nabla f(\mathbf{x})\|^2 \\
f(\mathbf{x}) + \nabla f(\mathbf{x})^T(-\frac{1}{M}\nabla f(\mathbf{x})) + \frac{M}{2}\|\frac{1}{M}\nabla f(\mathbf{x})\|^2 &= f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|^2 \\
f(\mathbf{x}^*) &\leq f(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|^2
\end{aligned}$$

Therefore $f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{1}{2M}\|\nabla f(\mathbf{x})\|^2$

Question4

Consider the optimization problem $\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuously differentiable function. To develop a weighted gradient descent method, let $\mathbf{W} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite (SPD) matrix. Denote by $\mathbf{W}^{\frac{1}{2}}$ the unique SPD square root of \mathbf{W} (i.e., $(\mathbf{W}^{\frac{1}{2}})^2 = \mathbf{W}$) and by $\mathbf{W}^{-\frac{1}{2}}$ its inverse. Given the current iterate $\mathbf{x}^{(k)}$, define the next iterate $\mathbf{x}^{(k+1)}$ as the solution of the following constrained optimization problem:

$$\begin{aligned}
&\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}^{(k)}) + \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \rangle \\
&\text{subject to } \|\mathbf{W}^{\frac{1}{2}}(\mathbf{x} - \mathbf{x}^{(k)})\|_2 \leq \alpha_k \|\mathbf{W}^{-\frac{1}{2}}\nabla f(\mathbf{x}^{(k)})\|_2
\end{aligned}$$

where $\alpha_k > 0$ is a step-size parameter.

Answer the following questions:

- (a) Derive an explicit formula for $\mathbf{x}^{(k+1)}$
- (b) Prove that $\mathbf{x}^{(k+1)}$ is equivalently the unique minimizer of the unconstrained quadratic problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \nabla f(\mathbf{x}^{(k)}) + \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \rangle + \frac{1}{2\alpha_k} \|\mathbf{W}^{\frac{1}{2}}(\mathbf{x} - \mathbf{x}^{(k)})\|_2^2 \right\}$$

Answer

(a)

We know the inequality, that:

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{W}^{-\frac{1}{2}}\mathbf{u}\|_2 \cdot \|\mathbf{W}^{\frac{1}{2}}\mathbf{v}\|_2$$

Let $\mathbf{d} = \mathbf{x} - \mathbf{x}^{(k)}$, the problem is equivalent to:

$$\min_{\mathbf{d}} \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d} \rangle$$

that subject to $\|\mathbf{W}^{\frac{1}{2}}\mathbf{d}\|_2 \leq \alpha_k \|\mathbf{W}^{-\frac{1}{2}}\nabla f(\mathbf{x}^{(k)})\|_2$

Using this inequality, we can get:

$$|\langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d} \rangle| \leq \|\mathbf{W}^{\frac{1}{2}}\mathbf{d}\|_2 \alpha_k \|\mathbf{W}^{-\frac{1}{2}}\nabla f(\mathbf{x}^{(k)})\|_2$$

We need \mathbf{u} and \mathbf{v} to be linear dependent to get the minimum. s.t.

$$\mathbf{W}^{\frac{1}{2}}\mathbf{d} = k\mathbf{W}^{-\frac{1}{2}}\nabla f(\mathbf{x}^{(k)})$$

We can derive $\mathbf{d} = k\mathbf{W}^{-1}\nabla f(\mathbf{x}^{(k)})$

And we need:

$$\|\mathbf{W}^{\frac{1}{2}}k\mathbf{W}^{-1}\nabla f(\mathbf{x}^{(k)})\|_2 = \alpha_k \|\mathbf{W}^{-\frac{1}{2}}\nabla f(\mathbf{x}^{(k)})\|_2$$

So we get $k = \alpha_k$. Therefore, we get the answer:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{W}^{-1}\nabla f(\mathbf{x}^{(k)})$$

(b)

We know the problem can be treat as (since $f(\mathbf{x}^{(k)})$ is constant):

$$\begin{aligned} \min_{\mathbf{d}} \{ \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{d} \rangle + \frac{1}{2\alpha_k} \|\mathbf{W}^{\frac{1}{2}} \mathbf{d}\|_2^2 \} \\ \min_{\mathbf{d}} \{ \nabla f(\mathbf{x}^{(k)})^T \mathbf{d} + \frac{1}{2\alpha_k} \mathbf{d}^T \mathbf{W} \mathbf{d} \} \end{aligned}$$

Let $\Phi(\mathbf{d}) = \nabla f(\mathbf{x}^{(k)})^T \mathbf{d} + \frac{1}{2\alpha_k} \mathbf{d}^T \mathbf{W} \mathbf{d}$

It's easy to find the Hessian matrix $\mathbf{H} = \frac{1}{\alpha_k} \mathbf{W}$, and since $\alpha_k > 0$ and \mathbf{W} is SPD, so we know $\Phi(\mathbf{d})$ is strictly convex.

$$\begin{aligned} \nabla \Phi(\mathbf{d}) &= \nabla f(\mathbf{x}^{(k)}) + \frac{1}{\alpha_k} \mathbf{W} \mathbf{d} = 0 \\ \mathbf{d} &= -\alpha_k \mathbf{W}^{-1} \nabla f(\mathbf{x}^{(k)}) \end{aligned}$$

Therefore, $\mathbf{x}^{(k+1)}$ is equivalently the unique minimizer of this unconstrained quadratic problem.