By Zhuo Chen
Group Member: Lemin Tian
Yiran Xie

Homework 1

1.

(a)

(i) It is true. $(A^{-1})^T (A^T) = (AA^{-1})^T = I$, Thus $(A^{-1})^T = (A^T)^{-1}$

(ii) It is False. Counterexample: $A = B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$,

Here, $(A+B)^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$, While $A^{-1} + B^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

(iii) It is true. PF: For a symmetric matrix A, i.e. $A^T = A$, Using (i),
we have $(A^{-1})^T = (A^T)^{-1} = A^{-1} \Rightarrow A^{-1}$ is itself symmetric

(b) $X^T = (U \Sigma V^T)^T = (V^T)^T \Sigma^T U^T = V \Sigma U^T$

$\therefore XX^T = (U \Sigma V^T) \cdot V \Sigma U^T = U \Sigma (V^T V) \Sigma U^T = U (\Sigma \cdot \Sigma) U^T$

Since $U^T U = I \Rightarrow U^T U \cdot U^{-1} = I U^{-1} = U^{-1} \Rightarrow U^T = U^{-1}$

Thus for XX^T, $\Lambda = \Sigma \cdot \Sigma$, $Q = U$

(c) (i) 757, 158, 130

(ii) 68125.6

2.

(a) (i) $\cancel{P(H=h \mid D=d) \leq P(H=h)}$ $\hookrightarrow$ On last Page.

~~Explain: Set $m = P(H=h \mid D=d)$, $n = P(D=d \mid H=h) P(H=h) = P(D=d, H=h)$~~

~~Then $m \cdot P(D=d) = P(H=h, D=d) = n$, $\Rightarrow m = \frac{n}{P(D=d)}$~~

~~Explain: Assume for all possible D, $D \in \Omega$, Then $P(H=h) = P(H=h \mid D \in \Omega)$~~

~~While $\{D=d\} \subseteq \{D \in \Omega\}$, Thus $P(H=h \mid D=d) \leq P(H=h)$~~

(ii) $P(H=h \mid D=d) \geq P(D=d \mid H=h) P(H=h)$

Explain: Set $m = P(H=h \mid D=d)$, $n = P(D=d \mid H=h) P(H=h) = P(D=d, H=h)$

Then $m \cdot P(D=d) = P(H=h, D=d) = n \Rightarrow m = \frac{n}{P(D=d)} \geq n$

(b) (i) PF: Assume the density function of the joint distribution is $f(x,y)$

Then $E_X(X \mid Y) = \int \frac{f_{X,Y}}{f_Y} x \, dx$, Thus

$E_Y[E_X(X \mid Y)] = \int_Y f_Y \int_X \frac{f_{X,Y}}{f_Y} x \, dx \, dy = \int_X x (\int_Y f_{X,Y} dy) dx = \int_X f_X \cdot x \, dx = EX$

(ii) PF: $E_Y[\mathrm{Var}_X(X \mid Y)] = E_Y[E_X(X^2 \mid Y) - (E_X(X \mid Y))^2]$ $\cdots$ ①

$\mathrm{Var}_Y[E_X(X \mid Y)] = E_Y(E_X(X \mid Y))^2 - [E_Y E_X(X \mid Y)]^2 = E_Y(E_X(X \mid Y))^2 - (EX)^2$

$① + ② \Rightarrow E_Y[Var_X(X|Y)] + Var_Y[E_X(X|Y)] = E_Y E_X(X^2|Y) - (E(X))^2$

From (i), we know $E_Y E_X(X^2|Y) = EX^2$, thus

$E_Y[Var_X(X|Y)] + Var_Y[E_X(X|Y)] = EX^2 - (EX)^2 = Var(X)$

3.

(a) i) ($\Rightarrow$) If A is PSD, Using spectral decomposition, $A = U\Lambda U^T$

Choosing $x = U$, we have $U^T U \Lambda U^T U \geq 0 \Rightarrow \Lambda \geq 0$, Thus $\lambda_i \geq 0$

ii) ($\Leftarrow$) if $\lambda_i \geq 0$, for any $x \neq 0$,

$x^T A x = x^T U \Lambda U^T x = (x^T U)\Lambda(x^T U)^T$, Set $x^T U = Q$, — $1 \times d$ vector

Then $x^T A x = \sum \lambda_i Q_i^2 \geq 0$, Thus A is PSD

(b) From the Proof in (a), it is easy to get the conclusion similar in (a). We only need to prove $\sum Q_i^2 \neq 0$ in addition. In fact, $\sum Q_i^2 = Q \cdot Q^T = (x^T U)(U^T x) = x^T x \neq 0$. Thus concluded.

4.

For $X \sim Pois(\lambda)$, $P(X = x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$

$\therefore$ The target function is $F = \sum_{i=1}^{n} \log \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = (\sum x_i)\log\lambda - n\lambda - \sum \log x_i!$

$F'(\lambda) = \frac{\sum x_i}{\lambda} - n$, it is decreasing. Thus the global maximizer is $\lambda = \frac{\sum x_i}{n}$, this is the maximum likehood estimator.

5.

(a) Using reduction to absurdity.

Assume f has two global minimizers, $x_1 > x_2$ (without loss of general). Then we have $f'(x_1) = f'(x_2) = 0$. However, if f is strictly convex, then $f''(x) > 0$, we have $f'(x_1) > f'(x_2)$. Contradictory.

(b) Assuming the neighborhood of x^* is $U(x^*)$, Then for $y \in U(x^*)$ we have $f(x^*) \leq f(y)$, $\nabla f(x^*) = 0$. Then From Taylor expansion,

$f(y) = f(x^*) + \langle \nabla f(x^*), y - x^* \rangle + \frac{1}{2}\langle (y - x^*), \nabla^2 f(x^*)(y - x^*) \rangle + O(\|x^* - y\|^2) \geq f(x^*)$

Since $y \in U(x^*)$, thus $O(\|x^* - y\|^2) \to 0$, then we have

$\langle (y - x^*), \nabla^2 f(x^*)(y - x^*) \rangle \geq 0$

Since $y \in U(x^*)$, for any $x \in \mathbb{R}^d$, it can be denoted by $\lambda(y - x^*)$, $\lambda \in \mathbb{R}$.
Thus $x^T \nabla^2 f(x^*) x \geq 0$, for any x, i.e. the Hessian of f is PSD at x^*.

(c) If f is twice continuous differenciable, then f is convex $\Leftrightarrow$ $f(x+y) \geq f(x) + \nabla f(x) \cdot y$, $\forall x, y$

i) ($\Rightarrow$) If f is convex, choose $x+y \in U(x)$, i.e. $y \to 0$

Then $f(x+y) = f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} \langle y, \nabla^2 f(x) y \rangle \geq f(x) + \nabla f(x) \cdot y$

$\Rightarrow \langle y, \nabla^2 f(x) y \rangle \geq 0$.

Similar to that in (b), $\nabla^2 f(x)$ is PSD.

ii) ($\Leftarrow$) From Taylor expansion in (5) of the question,

$f(x+y) = f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} \langle y, \nabla^2 f(x+ty) y \rangle$.

Since $\nabla^2 f(x)$ is PSD for all x, than

$f(x+y) \geq f(x) + \langle \nabla f(x), y \rangle$, $\Rightarrow$ f is convex

(d) $f(x) = \frac{1}{2} x^T A x + b^T x + c$

$\therefore$ $\nabla_x f(x) = Ax + b$, $\nabla_x^2 f(x) = A$,

Thus the Hessian of f is A.

~~If $A \geq 0$, f is convex. If $A > 0$, f is strictly convex.~~

If $A \succeq 0$, f is convex. If $A \succ 0$, f is strictly convex.

2.

(a) (i) The relationship is " depends "

Explain: Set $m = P(H=h \mid D=d)$, $n = P(H=h)$, then

① if H & D are independent, obviously $m = n$

② if the probability space is $\begin{cases} \frac{1}{4}: (D=d, H=h) \\ \frac{1}{4}: (D=d, H=h_1) \\ \frac{1}{2}: (D=d_1, H=h) \end{cases}$, then $m = \frac{1}{2} < n = \frac{3}{4}$

③ if the probability space is $\begin{cases} \frac{1}{2}: (D=d, H=h) \\ \frac{1}{2}: (D=d_1, H=h_1) \end{cases}$, then $m = 1 > n = \frac{1}{2}$