CS 189 Homework2

Xu Zhihao

July 8, 2019

I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted.

Signature: Zhihao Xu

1. Identities with Expectation

(1) By definition:

$$E(X^k) = \int x^k f(x)dx = \int_0^\infty x^k \lambda e^{-\lambda x} dx$$

When K = 1,

$$\begin{aligned} E(X) &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty x e^{-\lambda x} dx \\ &= \left[\lambda \times \frac{-\lambda x - 1}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda} \end{aligned}$$

So, when K = 1, the result is true, $E(X^k) = \frac{k!}{\lambda^k}$. Suppose, it is true for k = n-1,

$$E(X^{n-1}) = \frac{(n-1)!}{\lambda^{n-1}}$$

For k = n,

$$\begin{aligned} E(X^n) &= \int x^n f(x)dx = \int_0^\infty x^k \lambda e^{-\lambda x} dx \\ &= x^n \cdot \left[-e^{-\lambda x}\right]_0^\infty - \int_0^\infty n x^{n-1}(-e^{-\lambda x})dx \\ &= 0 + \frac{n}{\lambda} E(X^{n-1}) \\ &= n \times \frac{(n-1)!}{\lambda^{n-1}} = \frac{n!}{\lambda^n} \end{aligned}$$

which is true for x = n. So, by mathematical induction, $E(X^k) = \frac{k!}{\lambda^k}$ for $k \in \mathbb{Z}$

(2) By definition,

$$P(X \geq t) = 1 - F(t)$$

$$\begin{aligned}
\int_0^\infty P(X \geq t)dt &= \int_0^\infty [1 - F(t)]dt \\
&= t\,[1 - F(t)]_0^\infty + \int_0^\infty x f(x)dx \\
&= 0 + E(X)
\end{aligned}$$

which shows that $E(X) = \int_0^\infty P(X \geq t)dt$ if X is a non-negative real-valued Random Variable.

(3) We first set two indicator variables,

$$I_{\{X=0\}} = \begin{cases} 1 & \text{X = 0} \\ 0 & \text{X} > 0 \end{cases}$$

$$I_{\{X>0\}} = \begin{cases} 0 & \text{X = 0} \\ 1 & \text{X} > 0 \end{cases}$$

$$\begin{aligned} E(X) &= E(XI_{\{X=0\}}) + E(XI_{\{X>0\}}) \\ &= 0 + E(XI_{\{X>0\}}) = E(XI_{\{X>0\}}) \\ &\leq \sqrt{E(X^2)} \times \sqrt{E(I_{\{X>0\}})^2} \\ &= \sqrt{E(X^2) \cdot E(I_{\{X>0\}})} \\ &= \sqrt{E(X^2) \cdot P(X > 0)} \end{aligned}$$

So, we can get that

$$E(X)^2 \leq E(X^2) \cdot P(X > 0)$$

which is equivalent to

$$P(X > 0) \geq \frac{E(X)^2}{E[X^2]}$$

(4) First we define the similar indicator variables $I_{\{t-X>0\}}$. Since we already known

$$t - X \leq (t-X)I_{\{t-X>0\}}$$

We can get

$$E(t-X) \leq E\{(t-X)I_{\{t-X>0\}}\}$$

By Cauchy–Schwarz inequality:

$$E(t-X) \leq \sqrt{E[(t-X)^2]} \cdot \sqrt{E[I_{\{t-X>0\}}]^2}$$

Reformulate the inequality, we can get

$$\begin{aligned}
P(X<t) &= \frac{E^2(t-X)}{E[(t-X)^2]} \\
&= \frac{E^2(X) - 2tE(X) + t^2}{E(X^2) - 2tE(X) + t^2} \\
&= \frac{E^2(X) + t^2}{E(X^2) + t^2} \\
&\geq \frac{t^2}{E(X^2) + t^2} \\
\Longleftrightarrow 1 - P(X<t) &\leq 1 + \frac{-t^2}{E(X^2) + t^2} \\
&= \frac{E(X^2)}{E(X^2) + t^2} \\
\Longleftrightarrow P(X \geq t) \leq \frac{E(X^2)}{E(X^2) + t^2}
\end{aligned}$$

2. Properties of Gaussians

(1) Since $X \sim N(0, \sigma^2)$, the pdf of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\}$$

$$E[e^{\lambda x}] = \int e^{\lambda x} f(x) dx = \int_{-\infty}^{+\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\}$$

Use $t = \frac{x}{\sigma}$ to substitute x

$$\begin{aligned} E[e^{\lambda x}] &= \exp(\frac{\sigma^2\lambda^2}{2}) \int_{-\infty}^{+\infty} \exp(-\frac{\sigma^2\lambda^2}{2}) \exp(\sigma\lambda t) \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt \\ &= \exp(\frac{\sigma^2\lambda^2}{2}) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(t-\lambda)^2}{x}\} dx \end{aligned}$$

Since $\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(t-\lambda)^2}{x}\}$ is the pdf of $N(\lambda, 1)$

$$E[e^{\lambda x}] = \exp(\frac{\sigma^2\lambda^2}{2})$$

(2) Proof:

$$\begin{aligned}
P(X \geq t) &= P(e^{\lambda x} \geq e^{\lambda t}) \\
&\leq \frac{E[e^{\lambda x}]}{e^{\lambda t}} = \frac{e^{\frac{\sigma^2 \lambda^2}{2}}}{e^{\lambda t}} \\
&= \exp\{-\lambda t + \frac{1}{2}\sigma^2\lambda^2\} \\
&\leq \exp\{-\frac{t^2}{2\sigma^2}\}
\end{aligned}$$

Since $X \sim N(0, \sigma^2)$, and X is symmetric among $X = 0$.

$$P(|X| \geq t) = 2P(X \geq t) \leq 2\exp\{-\frac{t^2}{2\sigma^2}\}$$

(3) Since $X_1, X_2, ..., X_n \sim N(0, \sigma^2)$, by Central Limit Theorem:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(0, \frac{\sigma^2}{n})$$

Using the inequality proved in part (2)

$$\begin{aligned} P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t) &\leq \exp(-\frac{t^2}{2\frac{\sigma^2}{n}}) \\ &= \exp(-\frac{nt^2}{2\sigma^2}) \end{aligned}$$

when $n \to \infty$

$$\lim_{n\to\infty} \exp(-\frac{nt^2}{2\sigma^2}) = 0$$

So, the inequality goes to be

$$\lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t) = 0$$

(4) Let $X \sim N(0,1), Y = RX$, the pdf of R is $f_R(r) = \begin{cases} \frac{1}{2}, & r = 1 \\ -\frac{1}{2}, & r = -1 \\ 0, & Otherwise \end{cases}$

First we need to show Y is Gaussian.

$$\begin{aligned} P(Y \leq x) &= P(RX \leq x) \\ &= P(X \leq x | R = 1) + P(X \geq -x | R = -1) \\ &= P(X \leq x)P(R = 1) + P(X \geq -x)P(R = -1) \\ &= \frac{1}{2}\left[P(X \leq x) + P(X \geq -x)\right] \\ &= P(X \leq x) \\ &= \Phi(x) \end{aligned}$$

Here if we choose $a = \frac{1}{2}, b = \frac{1}{2}$, $aX + bY$ is not Gaussian.

$$P(aX + bY = 0) = 0$$

which does not satisfy Gaussian distribution

(5)

$$u_x = < u, X > = u_1 X_1 + u_2 X_2 + \cdots + u_n X_n$$
$$v_x = < v, X > = v_1 X_1 + v_2 X_2 + \cdots + v_n X_n$$

$$\begin{aligned}
Cov(u_x, v_x) &= E(u_x v_x) - E(u_x)E(v_x) \\
&= E\left[\sum_{i=1}^{n} u_i v_i X_i^2 + \sum_{i \neq j} (u_i v_j + u_j v_i) X_i X_j\right] - \left[\sum_{i=1}^{n} u_i E(X_i)\right]\left[\sum_{i=1}^{n} v_i E(X_i)\right] \\
&= \sum_{i=1}^{n} u_i v_i E(X_i^2) + \sum_{i \neq j} (u_i v_j + u_j v_i) E(X_i X_j) \\
&\quad - \left[\sum_{i=1}^{n} u_i v_i E^2(X_i) + \sum_{i \neq j} (u_i v_j + u_j v_i) E(X_i) E(X_j)\right] \\
&= \sum_{i=1}^{n} u_i v_i [E(X_i^2) - E^2(X_i)] + \sum_{i \neq j} (u_i v_j + u_j v_i)[E(X_i X_j) - E(X_i)E(X_j)] \\
&= \sum_{i=1}^{n} u_i v_i Var(X_i) + \sum_{i \neq j} (u_i v_j + u_j v_i) Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} u_i v_i = < u, v > = 0
\end{aligned}$$

Using the fact that jointly normal random variables are independent iff. they are uncorrelated u_x and u_y are independent.

(6) Take $Y = \max\limits_{1 \le i \le n} |X_i|$, using Jensen's inequality

$$\begin{aligned} e^{tE(Y)} &\le E(e^{tY}) \\ &= E\left(\max_{1\le i\le n} e^{t|X_i|}\right) \\ &\le \sum_{i=1}^{n} E\left(e^{t|X_i|}\right) \end{aligned}$$

where $|X_i|$ follows folded normal distribution with $\mu = 0$ and σ^2. Using the formula of folded normal distribution's mgf

$$\begin{aligned} E\left(e^{t|X_i|}\right) &= \varphi(-it) \\ &= 2e^{\frac{\sigma^2 t^2}{2}}\left[1 - \Phi(-\sigma t)\right] \\ &\le 2e^{\frac{\sigma^2 t^2}{2}} \end{aligned}$$

Then

$$\begin{aligned} e^{tE(Y)} &\le \sum_{i=1}^{n} E\left(e^{t|X_i|}\right) \\ &\le \sum_{i=1}^{n} 2e^{\frac{\sigma^2 t^2}{2}} \\ &= 2ne^{\frac{\sigma^2 t^2}{2}} \\ \Longleftrightarrow E(Y) &\le \frac{\ln 2n}{t} + \frac{t\sigma^2}{2} \end{aligned}$$

Take $f(t) = \frac{\ln 2n}{t} + \frac{t\sigma^2}{2}$, then let $\frac{df(t)}{dt} = 0$, we can get

$$t^* = \frac{\sqrt{2\ln(2n)}}{\sigma}$$

which leads to

$$E(Y) \le f(t^*) = \sigma\sqrt{2\ln(2n)}$$

So, we can get the result

$$E\left(\max_{1\le i\le n} e^{t|X_i|}\right) \le f(t^*) = C\sqrt{\ln(2n)}\sigma, \text{with C = } \sqrt{2}$$

3. Linear Algebra Review

(1) Since A is a real symmetric matrix, we can do the eigen-decomposition.

$$A = Q\Lambda Q^T = Q\Lambda Q^{-1}$$

where $\Lambda = diag\{\lambda_1, \lambda_2, ..., \lambda_n\}$, λ_i are all the eigenvalues of the matrix A.
Then for $\forall x \in \mathbb{R}^n$

$$x^T AX \geq 0 \Longleftrightarrow x^T Q\Lambda Q^T x \geq 0 \Longleftrightarrow (x^T Q)\Lambda(x^T Q)^T \geq 0$$
$$y\Lambda y^T \geq 0 \Longleftrightarrow \sum_{i=1}^{n} \lambda_i y_i^2 \geq 0$$

So, the definition (a) is equivalent to

$$\sum_{i=1}^{n} \lambda_i y_i^2 \geq 0 \text{ for } \forall y \in \mathbb{R}^n$$

This condition satisfies if and only if $\lambda_i \geq 0$, $i \in \{1, 2, ..., n\}$. Hence the definition (a) and (b) are equivalent.

Next we consider definition (c). We need to show the Sufficiency and Necessity between (c) and (a).

[Sufficiency] Define $U = Q\sqrt{\Lambda}Q^T$, we can easily verify that $U = U^T$. Then

$$\begin{aligned} UU^T = UU &= Q\sqrt{\Lambda}Q^T Q\sqrt{\Lambda}Q^T \\ &= Q\sqrt{\Lambda}Q^{-1}Q\sqrt{\Lambda}Q^T \\ &= Q\sqrt{\Lambda}\sqrt{\Lambda}Q^T \\ &= Q\Lambda Q^T = A \end{aligned}$$

[Necessity] If $\exists U \in \mathbb{R}^{n\times n}$, such that $A = UU^T$. Let y = $U^T x$,

$$x^T Ax = x^T UU^T x = (U^T x)^T (U^T x) = y^T y = \sum_{i=1}^{n} y_i^2 \geq 0$$

Hence the definition (a) and (c) are equivalent.
In conclusion, all the definition (a) (b) and (c) are equivalent.

(2) (a) For $\forall x \in \mathbb{R}^n$, $x^T Ax \geq 0$ and $x^T Bx \geq 0$, then

$$x^T(2A+3B)x = 2x^T Ax + 3x^T Bx \geq 0$$

Hence, 2A+3B is PSD

(b) If A is PSD, then for $\forall x \in \mathbb{R}^n$,$x^T Ax \geq 0$. Let$x = e_i$, the elementary vector. Only i-th element is 1, others are 0. Then

$$e_i^T Ae_i = a_{ii} \geq 0$$

(c) If A is PSD, then for $\forall x \in \mathbb{R}^n$, $x^T Ax \geq 0$. Let $x = \mathbf{1}$, all 1 vector. Then

$$\mathbf{1}^T A\mathbf{1} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \geq 0$$

(d) If A and B are PSD, then we get $A = UU^T$, $B = VV^T$, where U and V are $n \times n$ real valued matrix. Then

$$Tr(AB) = Tr(UU^T VV^T) = Tr[(V^T U)(V^T U)^T]$$

So we can clearly see that $(V^T U)(V^T U)^T$ is PSD. Using the result in (b) we can get that

$$Tr(AB) = \sum_{i=0}^{n} a_{ii} \geq 0$$

(e) [Sufficiency] Since any PSD matrix A can be decomposed into the product of two PSD matrix.

$$\begin{aligned} &Tr(AB) = 0 \\ \Rightarrow\ &Tr(A^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{2}}B^{\frac{1}{2}}) = 0 \\ \Rightarrow\ &Tr(B^{\frac{1}{2}}A^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{2}}) = 0 \\ \Rightarrow\ &Tr((A^{\frac{1}{2}}B^{\frac{1}{2}})^T A^{\frac{1}{2}}B^{\frac{1}{2}}) = 0 \end{aligned}$$

Using the property of square matrix

$$Tr(A^T A) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$$

If $Tr(A^T A) = 0$, then $sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 = 0 \Rightarrow x_{ij} = 0$, which is equivalent to $X = 0$. In this question,

$$\begin{aligned} &Tr((A^{\frac{1}{2}}B^{\frac{1}{2}})^T A^{\frac{1}{2}}B^{\frac{1}{2}}) = 0 \\ \Rightarrow\ &A^{\frac{1}{2}}B^{\frac{1}{2}} = 0 \\ \Rightarrow\ &AB = A^{\frac{1}{2}}(A^{\frac{1}{2}}B^{\frac{1}{2}})B^{\frac{1}{2}} = 0 \end{aligned}$$

[Necessity] It obvious that if AB=0, Tr(AB) = 0

(3) Since A is a real symmetric matrix, we can do the eigen-decomposition.

$$A = Q\Lambda Q^T = Q\Lambda Q^{-1}$$

where $\Lambda = diag\{\lambda_1, \lambda_2, ..., \lambda_n\}$, λ_i are all the eigenvalues of the matrix A. For $\forall x \in \mathbb{R}^n$, let $y = Q^T x$. Then we have

$$x^T Ax = x^T Q\Lambda Q^T x = yT^{\Lambda} y = \sum_{i=1}^{n} \lambda_i y_i^2$$

Obviously,

$$\lambda_{min}(A) \sum_{i=1}^{n} y_i^2 \le \sum_{i=1}^{n} \lambda_i y_i^2 \le \lambda_{max}(A) \sum_{i=1}^{n} y_i^2$$

Here we can add a constrain $\|x\|_2 = 1$,

$$\sum_{i=1}^{n} \lambda_i y_i^2 = y^T y = x^T QQ^T x = x^T Ix = x^T x = 1$$

So, substitute it back, we can get

$$\begin{aligned} &\lambda_{min}(A) \le x^T Ax \le \lambda_{max}(A) \\ &\Rightarrow \max \lambda(A) = \max_{\|x\|_2=1} x^T Ax \end{aligned}$$

4. Gradients and Norms

(1)

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$
$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Here we use the Minkowski Inequality continuously,

$$\begin{aligned}
\|x\|_2 &= \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} \\
&\leq \left(\sum_{i=1}^{n-1} |x_i|^2\right)^{\frac{1}{2}} + |x_n^2|^{\frac{1}{2}} \\
&\leq \left(\sum_{i=1}^{n-2} |x_i|^2\right)^{\frac{1}{2}} + |x_{n-1}^2|^{\frac{1}{2}} + |x_n^2|^{\frac{1}{2}} \\
&\cdots \\
&\leq \sum_{i=1}^{n} |x_i| = \|x\|_1
\end{aligned}$$

Then using Cauchy-Schwarz Inequality

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = \sum_{i=1}^{n} |x_i| \cdot 1 \leq \sum_{i=1}^{n} |x_i|^2 \cdot \sum_{i=1}^{n} 1^2 = \sqrt{n}\, \|x\|_2$$

(2) (a)

$$\frac{\partial \alpha}{\partial \beta_i} = \frac{y_i}{\beta_i}$$

(b)

$$\frac{\partial \beta_i}{\partial \gamma_j} = \begin{cases} 0, & i \neq j \\ \cosh(\gamma_i), & i = j \end{cases}$$

(c) Since $\gamma = A\rho + b$,

$$\gamma_i = \left(\sum_{i=1}^{m} a_{ij}\rho_j \right) + b_i$$

Then we can compute

$$\frac{\partial \gamma_i}{\partial \rho_j} = a_{ij}$$

(d) First we can compute $f(x)$,

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} y_i \ln \left[\sinh(Ax+b)_i \right] \\ &= \sum_{i=1}^{n} y_i \ln \left\{ \sinh \left[\left(\sum_{i=1}^{m} a_{ij}x_j \right) \right] + b_i \right\} \end{aligned}$$

Then we can compute

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n} \left\{ y_i \cdot \coth \left[\sum_{k=1}^{m} (a_{ik}x_k) + b_i \right] \cdot a_{ij} \right\}$$

(3)

$$
\begin{aligned}
A &= \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_n \end{bmatrix} \\
X &= \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \\
A^T X &= \begin{bmatrix} a_1^T x_1 & a_1^T x_2 & \cdots & a_1^T x_n \\ a_2^T x_1 & a_2^T x_2 & \cdots & a_2^T x_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n^T x_1 & a_n^T x_2 & \cdots & a_n^T x_n \end{bmatrix} \\
Tr(A^T X) &= \sum_{i=1}^{n} a_i^T x_i \\
&\Longrightarrow \frac{\partial Tr(A^T X)}{\partial x_{ij}} = a_{ij} \\
&\Longrightarrow \nabla_X Tr(A^T X) = A
\end{aligned}
$$

(4) (a) Let

$$f(x) = \frac{1}{2}x^T Ax - b^T x$$

Then

$$\nabla_x f(x) = \frac{1}{2}\left(Ax + A^T x\right) - b = Ax - b$$

We need to solve

$$\nabla_x f(x) = 0$$

which is equivalent to

$$Ax = b$$

So, $x^* A^{-1}b$ if $A^{-1} exists$

(b) Here we use gradient with step size equals 1, which is equivalent to Jacobian Method

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - 1 \cdot \nabla_{x^{(k)}} f(x^{(k)}) \\ \Longrightarrow x^{(k+1)} &= x^{(k)} - (Ax^{(k)} - b) \\ \Longrightarrow x^{(k+1)} &= (I - A)x^{(k)} + b \end{aligned}$$

(c) Since we already get $x^{(k+1)} = (I - A)x^{(k)} + b$, $b = Ax^*$

$$\begin{aligned} x^{(k)} - x^* &= (I - A)x^{(k-1)} + Ax^* - x^* \\ x^{(k)} - x^* &= (I - A)\left(x^{(k-1)} - x^*\right) \end{aligned}$$

(d) Use the fact that if λ is an eigenvalue of A, λ^2 is an eigenvalue of A^2

$$\|Ax\|_2^2 = x^T A^T Ax = x^T A^2 x$$

Using the result get in Problem 3 exercise 3

$$\|Ax\|_2^2 = x^T A^2 x \le \lambda_{max}(A^2)\|x\|_2^2 = \lambda_{max}^2(A)\|x\|_2^2$$

which is equivalent to

$$\|Ax\|_2 \le \lambda_{max}(A)\|x\|_2$$

(e) Take $u = x^{(k-1)} - x^*$

$$\begin{aligned} \left\|x^{(k)} - x\right\|_2^2 &= (x^{(k)} - x^*)^T (x^k - x^*) \\ &= \left[(I - A)\left(x^{(k-1)} - x^*\right)\right]^T \left[(I - A)\left(x^{(k-1)} - x^*\right)\right] \\ &= \left[(I - A)u\right]^T \left[(I - A)u\right] \\ &= u^T (I - A)^T (I - A)u \\ &= u^T (I - A)^2 u \end{aligned}$$

Using the fact that if λ is an eigenvalue of A, $1 - \lambda$ is an eigenvalue of $I - A$.
Since

$$0 < \lambda_{min}(A) < \lambda_{max}(A) < 1$$

we can get

$$0 < \lambda_{min}(I - A) < \lambda_{max}(I - A) < 1$$

Using the result calculated in (d)

$$u^T (I - A)^2 u \le \lambda_{max}^2 (I - A) \cdot \|u\|_2^2$$

Take $\rho = \lambda_{max}(I - A)$

$$\begin{aligned} & u^T(I-A)^2 u \le \rho^2 \|u\|_2^2 \\ \Longleftrightarrow\ & \|x^{(k)} - x^*\|_2^2 \le \rho^2 \|x^{(k-1)} - x^*\|_2^2 \\ \Longleftrightarrow\ & \|x^{(k)} - x^*\|_2 \le \rho \|x^{(k-1)} - x^*\|_2 \end{aligned}$$

(f) Substitute the inequality get in (e) continuously

$$\|x^{(k)} - x^*\|_2 \le \rho^k \|x^{(0)} - x^*\|_2$$

So, we just need to make sure $\rho^k \|x^{(0)} - x^*\|_2 \le \varepsilon$, which can guarantee $\|x^{(k)} - x^*\|_2 \le \varepsilon$

$$\begin{aligned} & \rho^k \|x^{(0)} - x^*\|_2 \le \varepsilon \\ \Longleftrightarrow\ & \ln\left(\rho^k \|x^{(0)} - x^*\|_2\right) \le \ln \varepsilon \\ \Longleftrightarrow\ & k \ln \rho + \ln \|x^{(0)} - x^*\|_2 \le \ln \varepsilon \\ \Longleftrightarrow\ & k \le \frac{\ln \varepsilon - \ln \|x^{(0)} - x^*\|_2}{\ln \rho} \end{aligned}$$

(5)

$$\begin{aligned} L(\theta) &= \|y - X\theta\|_2^2 = (y - X\theta)^T(y - X\theta) \\ \nabla_\theta L(\theta) &= -2X^T(y - X\theta) = 0 \\ \iff X^T X\theta &= X^T y \end{aligned}$$

If X is full rank and $X^T X$ is non-singular, we can get

$$\theta^* = (X^T X)^{-1} X^T y$$

5. (1) For $\forall x \in \mathbb{R}^n$,

$$\begin{aligned}
x^T \Sigma x &= x^T E\left[(Z-\mu)(Z-\mu)^T\right] x \\
&= E\left[x^T (Z-\mu)(Z-\mu)^T x\right] \\
&= E\left\{\left[x^T(Z-\mu)\right]^T \left[x^T(Z-\mu)\right]\right\} \\
&= E\left(\left\|x^T(Z-\mu)\right\|_2^2\right) \\
&\geq 0
\end{aligned}$$

which is the definition of PSD matrix.

(2) Do the eigen-decomposition of covariance matrix Σ. If Σ has one zero eigenvalue, $\lambda_k = 0$. Then

$$\Sigma v_k = \lambda_k v_k = 0 \text{ where } \lambda_k = 0$$

Take $Y = \sum_{i=1}^{n} v_{ki} X_i$,

$$\begin{aligned} Var(Y) &= Var(\sum_{i=1}^{n} v_{ki} X_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ki} v_{kj} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} v_{ki} \sum_{j=1}^{n} v_{kj} \sigma_{ij} \\ &= \sum_{i=1}^{n} v_{ki} \cdot 0 \\ &= 0 \end{aligned}$$

which means Y is a constant. And

$$\exists v_k \in \mathbb{R}^n, \text{ s.t. } < v_k, X >= 0$$

So, X lost 1 degree of freedom. If Σ had $m \leq n$ zero eigenvalues, so that

$$< v_i, X >= 0 \text{ for } \forall j \in [1, m]$$

Construct a new $\tilde{X} \in \mathbb{R}^{n-m}$ containing all the RV corresponding to non-zero eigenvalues, through Gaussian Elimination of

$$V = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_m^T \end{bmatrix}$$

Here $\tilde{X}$ contains all the infomation that needs to solve

$$VX = 0$$