Student Name: Saeedreza Zouashkiani

Student ID: 400206262

Deep Learning Homework 1

Nov-22

# 1)

## 1-a)

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{rank(A)}\,\|A\|_2$$

$$\|A\|_2 = \sigma_{max}, \|A\|_F = \sqrt{trace(A^H A)} = \sqrt{\sum_{i=1}^{rank(A)} \sigma_i^2} = \sqrt{\sigma_{max}^2 + \cdots + \sigma_{min}^2}$$

$$\|A\|_2 = \sigma_{max} = \sqrt{\sigma_{max}^2} \leq \sqrt{\sigma_{max}^2 + \cdots + \sigma_{min}^2} = \sqrt{\sum_{i=1}^{rank(A)} \sigma_i^2} = \|A\|_F \leq \sqrt{rank(A).\sigma_{max}^2}$$

$$= \sqrt{rank(A)}\,\|A\|_2 \blacksquare$$

## 1-b)

### 1-b-i)

$$P(X \geq a) \leq \frac{E(X)}{a}$$

$$E(X) = P(X < a).E(X|X < a) + P(X \geq a).E(X|X \geq a)$$

Since X is non-negative then $E(X|X < a)$ is positive, and $E(X|X \geq a)$ is larger than $a$. Thus:

$$E(X) \geq P(X \geq a).E(X|X \geq a) \geq P(X \geq a).a$$

Therefore

$$P(X \geq a) \leq \frac{E(X)}{a} \blacksquare$$

### 1-b-ii)

$$P(|Z - \mu| \geq \varepsilon) = P((Z - \mu)^2 \geq \varepsilon^2) \leq \frac{E((Z - \mu)^2)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \blacksquare$$

### 1-b-iii)
Let $Z_i$ denote the random variable (indicator) that determines whether the random point falls within the circle. $Z_i$ can take a value of 1 with probability of $\frac{\pi}{4}$ and 0 with a probability of $1 - \frac{\pi}{4}$. Therefore

$$E(Z_i) = 1.\frac{\pi}{4} + 0.\left(1 - \frac{\pi}{4}\right) = \frac{\pi}{4}$$

$$Var(Z_i) = \frac{\pi}{4}.\left(1 - \frac{\pi}{4}\right)$$

The estimator then is $\hat{\pi} = \frac{4}{n}\sum Z_i$. We check whether $\hat{\pi}$ is an unbiased estimator of $\pi$.

$$E(\hat{\pi}) = E\left(\frac{4}{n}\sum Z_i\right) = \frac{4}{n}\cdot\frac{n\pi}{4} = \pi$$

$$Var(\hat{\pi}) = \frac{16}{n^2}Var\left(\sum Z_i\right) = \frac{\pi(4-\pi)}{n}$$

### 1-b-iv)
Using Chebyshev's inequality, we have:

$$P(|\hat{\pi} - \pi| \geq 0.01) \leq \frac{Var(\hat{\pi})}{(0.01)^2} = \frac{\pi(4-\pi)}{n(0.01)^2} \leq 1 - 0.95 = 0.05$$

Then if we solve for $n$, we get $n \geq 539353.24$. Therefore $n = 539354$

## 2)

### 2-i)
$$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = \frac{\partial(a_1 x_1 + \cdots + a_n x_n)}{\partial x} = [a_1 \dots a_n] = a^T$$

$$\frac{\partial x^T A x}{\partial x} = \frac{\partial(x^T A)x}{\partial x} + \frac{\partial x^T (Ax)}{\partial x} = x^T A + (Ax)^T = x^T(A + A^T)$$

### 2-ii)
$$\frac{\partial AA^{-1}}{\partial \beta} = 0 = \frac{\partial A}{\partial \beta}A^{-1} + A\frac{\partial A^{-1}}{\partial \beta}$$

By rearranging the terms and multiplying by $A^{-1}$ from left we get

$$\frac{\partial A^{-1}}{\partial \beta} = -A^{-1}\frac{\partial A}{\partial \beta}A^{-1}$$

### 2-iii)
Let $M_{ij}, C_{ij}, adj(A)$ be the (i, j) minor of $A$, (i, j) element of cofactor matrix of $A$, and adjugate matrix of $A$ respectively.

$$adj(A) = C^T$$

$$A^{-1} = \frac{1}{\det(A)} adj(A)$$

$$(A^{-1})^T_{ij} = \frac{1}{\det(A)} C_{ij}$$

By cofactor expansion of **A**

$$\det(A) = \sum_{k=1}^{n} A_{ik} C_{ik}$$

$$\frac{\partial \det(A)}{\partial A_{ij}} = \sum_{k=1}^{n} \left( \frac{\partial A_{ik}}{\partial A_{ij}} C_{ik} + A_{ik} \frac{\partial C_{ik}}{\partial A_{ij}} \right) = C_{ij}$$

$$\frac{\partial \det(A)}{\partial A} = C = adj(A^T) = \det(A) A^{-T} = |A|A^{-T} = \nabla_A |A|$$

$$\nabla_A \log|A| = \frac{\partial \log(\det(A))}{\partial A} = \frac{1}{\det(A)} \frac{\partial \det(A)}{\partial A} = |A|^{-1}|A|A^{-T} = A^{-T}$$

## 3)

Let $A$ be an $n \times n$ matrix.

### 3-i) $tr(A) = \lambda_1 + \cdots + \lambda_n$

The characteristic polynomial of A is defined as

$$p_A(t) = \det(tI - A) = (-1)^n (t^n - (tr(A)t^{n-1} + \cdots + (-1)^n \det(A))$$

Also, the characteristic polynomial can be factorized as

$$p_A(t) = (-1)^n (t - \lambda_1) \ldots (t - \lambda_n)$$

So, by comparing terms we get

$$tr(A) = \lambda_1 + \cdots + \lambda_n$$

### 3-ii) $\det(A) = \lambda_1 \ldots \lambda_n$

By comparing terms from last part, it is easily derived.

## 4)

If A has full column rank, the inverse of $A^T A$ exists

To check whether $A^\dagger$ is a pseudoinverse, we should have:

$$A = A A^\dagger A$$

$$A. A^\dagger. A = A.\left(A^T A\right)^{-1} A^T. A = A$$

Or if we have $A^\dagger = V \Sigma^\dagger U^T$

$$\left(A^T A\right)^{-1} A^T = \left(V \Sigma^T U^T U \Sigma V^T\right)^{-1} V \Sigma^T U^T = \left(V \Sigma^T \Sigma V^T\right)^{-1} V \Sigma^T U^T = V\left(\Sigma^T \Sigma\right)^{-1} V^T V \Sigma^T U^T$$
$$= V \Sigma^\dagger \Sigma^{T\dagger} \Sigma^T U^T = V \Sigma^\dagger U^T = A^\dagger$$

If **A** has full row rank, the inverse of $A A^T$ exists

$$A. A^\dagger. A = A. A^T \left(A A^T\right)^{-1}. A = A$$

Or

$$A^T \left(A A^T\right)^{-1} = V \Sigma^T U^T \left(U \Sigma V^T V \Sigma^T U^T\right)^{-1} = V \Sigma^T U^T U \left(\Sigma \Sigma^T\right)^{-1} U^T = V \Sigma^T \left(\Sigma \Sigma^T\right)^{-1} U^T = V \Sigma^T \Sigma^{T\dagger} \Sigma^\dagger U^T$$
$$= V \Sigma^\dagger U^T = A^\dagger$$

We start by eliminating the block matrix under A

$$\begin{bmatrix} I_n & 0 \\ -CA^{-1} & I_k \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix}$$

Then by eliminating the element above D

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_n & -A^{-1}B \\ 0 & I_k \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & D - CA^{-1}B \end{bmatrix}$$

Therefore, by combining the two

$$\begin{bmatrix} I_n & 0 \\ -CA^{-1} & I_k \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_n & -A^{-1}B \\ 0 & I_k \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}$$

Thus

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ CA^{-1} & I_k \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix}$$

Which is the LDU decomposition of M.

Therefore

$$\begin{aligned}
\det(M) &= \det \left( \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} \right) \\
&= \det \left( \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \right) \det \left( \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \right) \det \left( \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} \right) \\
&= 1. \det(A) \det(D - CA^{-1}B). 1 = \det(A) \det(D - CA^{-1}B)
\end{aligned}$$

## 5-ii)
Like the last part, using block LDU decomposition when D is invertible, we get

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}$$

$$\det(M) = \det \left( \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \right) \det \left( \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \right) \det \left( \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix} \right) = \det(D) \det(A - BD^{-1}C)$$

## 5-iii)
Assume that **A** and **D** are both invertible, therefore by inverting the LDU decomposition of part i and ii we get. First invert using part i.

$$\begin{aligned}
M^{-1} &= \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}^{-1} \begin{bmatrix} I_n & 0 \\ CA^{-1} & I_k \end{bmatrix}^{-1} \\
&= \begin{bmatrix} I_n & -A^{-1}B \\ 0 & I_k \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ -CA^{-1} & I_k \end{bmatrix} \\
&= \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}
\end{aligned}$$

Now inverting part ii, results in:

$$\begin{aligned}
M^{-1} &= \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}^{-1} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix}^{-1} \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix}^{-1} \\
&= \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \begin{bmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}
\end{aligned}$$

By comparing the first block matrix we get

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$$

If we substitute **D** with $-D$

$$\left(A + BD^{-1}C\right)^{-1} = A^{-1} - A^{-1}B\left(D + CA^{-1}B\right)^{-1}CA^{-1}$$

## 5-iv)

Using part I, with $D = 1, \boldsymbol{u} = B, \boldsymbol{v}^T = C$:

$$\det\left(\begin{bmatrix} I_n & 0 \\ \boldsymbol{v}^T A^{-1} & 1 \end{bmatrix}\begin{bmatrix} A & 0 \\ 0 & 1 - \boldsymbol{v}^T A^{-1}\boldsymbol{u} \end{bmatrix}\begin{bmatrix} I & A^{-1}\boldsymbol{u} \\ 0 & 1 \end{bmatrix}\right) = \det(A)(1 - \boldsymbol{v}^T A^{-1}\boldsymbol{u})$$

## 6)

## 6-i)

Using question 5

$$\det\begin{pmatrix} t\boldsymbol{I} - \boldsymbol{B} & -\boldsymbol{x} \\ \boldsymbol{y}^* & t - a \end{pmatrix} = (t - a)\det\left((t\boldsymbol{I} - \boldsymbol{B}) - \frac{\boldsymbol{x}\boldsymbol{y}^*}{t - a}\right) = (t - a).\det(t\boldsymbol{I} - \boldsymbol{B})\left(1 - \frac{\boldsymbol{y}^*(t\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{x}}{t - a}\right)$$

$$= \det(t\boldsymbol{I} - \boldsymbol{B})\left(t - a - \boldsymbol{y}^*(t\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{x}\right) = (t - a).p_B(t) - \boldsymbol{y}^*adj(t\boldsymbol{I} - \boldsymbol{B})\boldsymbol{x}$$

## 6-ii)

Using Courant-Fischer's theorem

$$\lambda_i(\boldsymbol{M}) = \min_{\dim V = i} \max_{\substack{x \in V, \\ \|x\| = 1}} < \boldsymbol{M}\boldsymbol{x}, \boldsymbol{x} >$$

$$\lambda_i(\boldsymbol{A} + \boldsymbol{B}) = \min_{\dim V = i} \max_{\substack{x \in V, \\ \|x\| = 1}} < \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} > + < \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x} >$$

To give an upper bound on a minimum value of a function, we just need to give an upper bound on some value it takes.

Let $V_A, V_B$ be subspaces of $\mathbb{R}^n$ with dimensions of $i + j$ and $n - j$ respectively which achieve the minimum values of $\max_{\substack{x \in V_A, \\ \|x\| = 1}} < \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} >$, $\max_{\substack{x \in V_B, \\ \|x\| = 1}} < \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x} >$ and let $W = V_A \cap V_B$ be their intersection. $W$ has dimension of at least $i$.

$$\max_{\substack{x \in W, \\ \|x\| = 1}} (< (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x}, \boldsymbol{x} >) \leq \max_{\substack{x \in W, \\ \|x\| = 1}} (< \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} >) + \max_{\substack{x \in W, \\ \|x\| = 1}} (< \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x} >) \leq \lambda_{i+j}(\boldsymbol{A}) + \lambda_{n-j}(\boldsymbol{B})$$

Since $W$ has dimension of at least $i$, the above is an upper bound on the value of $\max_{\substack{x \in V, \\ \|x\| = 1}} (< (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x}, \boldsymbol{x} >)$

for any $i'th$ dimensional subspace $V \subseteq W$

## 7)

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta^2} x_i e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^{2n}} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} \prod_{i=1}^{n} x_i$$

$$Ln\big(L(x_1, \dots, x_n; \theta)\big) = -\frac{\sum_{i=1}^{n} x_i}{\theta} - 2nLn(\theta) + \sum_{i=1}^{n} Ln(x_i)$$

Differentiating with respect to theta and setting to zero, yields:

$$\frac{\partial Ln\big(L(x_1, \dots, x_n; \theta)\big)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta^2} - \frac{2n}{\theta} = 0$$

$$\widehat{\theta_{ML}} = \frac{\sum_{i=1}^{n} x_i}{2n}$$

## 8)

### 8-i) ML

$$L(x_1, \dots, x_n; \mu) = \prod_{i=1}^{n} f(x_i; \mu) = \frac{1}{\big(\sqrt{2\pi\sigma^2}\big)^n} e^{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}}$$

$$Ln\big(L(x_1, \dots, x_n; \mu)\big) = -\frac{n}{2} Ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

Taking derivative of log likelihood function with respect to $\mu$ and setting it to zero

$$\frac{\partial Ln\big(L(x_1, \dots, x_n; \mu)\big)}{\partial \mu} = 0 \rightarrow \hat{\mu}_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mathbb{E}[\hat{\mu}_{ML}] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \mu$$

Therefore $\hat{\mu}_{ML}$ is an unbiased estimator of $\mu$

$$Var(\hat{\mu}_{ML}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) = \frac{\sigma^2}{n}$$

Using Chebyshev's inequality:

$$P[|\mu - \hat{\mu}_{ML}| < \epsilon] \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n \to \infty} 0$$

## 8-ii) MAP

We want to maximize $f(\mu)f(x|\mu)$

$$\frac{1}{\left(\sqrt{2\pi\beta^2}\right)} e^{-\frac{(\mu-\gamma)^2}{2\beta^2}} \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}}$$

Taking the log

$$-\frac{1}{2} Ln(2\pi\beta^2) - \frac{(\mu-\gamma)^2}{2\beta^2} - \frac{n}{2} Ln(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

Taking derivative with respect to $\mu$ and setting it to zero

$$-\frac{(\mu-\gamma)}{\beta^2} + \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = -\frac{(\mu-\gamma)}{\beta^2} + \frac{\sum_{i=1}^n (x_i)}{\sigma^2} - \frac{n\mu}{\sigma^2} \to \hat{\mu}_{MAP} = \frac{\beta^2 \sum_{i=1}^n (x_i) + \sigma^2 \gamma}{n\beta^2 + \sigma^2}$$

$$\mathbb{E}[\hat{\mu}_{MAP}] = \mathbb{E}\left[\frac{\beta^2 \sum_{i=1}^n (x_i) + \sigma^2 \gamma}{n\beta^2 + \sigma^2}\right] = \lim_{n \to \infty} \frac{n\beta^2 \mu + \sigma^2 \gamma}{n\beta^2 + \sigma^2} = \mu$$

Therefore $\hat{\mu}_{MAP}$ is an unbiased estimator of $\mu$

$$Var(\hat{\mu}_{MAP}) = \left(\frac{\beta^2}{n\beta^2 + \sigma^2}\right)^2 Var\left(\sum_{i=1}^n (x_i)\right) = \left(\frac{\beta^2}{n\beta^2 + \sigma^2}\right)^2 n\sigma^2$$

Using Chebyshev's inequality:

$$P[|\mu - \hat{\mu}_{ML}| < \epsilon] \leq \frac{Var(\hat{\mu}_{MAP})}{\epsilon^2} \xrightarrow{n \to \infty} 0$$

## 9)

Let $x \in \mathbb{R}^n, x_a \in \mathbb{R}^{n_a}, x_b \in \mathbb{R}^{n_b}$ so that $n_a + n_b = n$

We will at first prove part ii

### ii)

Let $S_a$ be a subset matrix of dimension $n_a \times n$ such that $s_{ij} = 1$, if the $j'th$ element in $x_a$ corresponds to $i'th$ element in $x$, and zero otherwise.

$$S_a = \begin{pmatrix} I_{n_a} & 0 \\ 0 & 0_{n_b} \end{pmatrix}$$

$$x_a = S_a x$$

Therefore, by applying linear transformation

$$x_a \sim \mathcal{N}(S_a \mu, S_a \Sigma S_a^T)$$

Therefore $x_a$ is a normal distribution with

$$E(x_a) = \mu_a$$

$$Cov(x_a) = \Sigma_{aa}$$

i)

The joint distribution of $x_a$ and $x_b$ is $\mathcal{N}(\mu, \Sigma)$ and from part ii, the marginal distribution of $x_b$ is $\mathcal{N}(\mu_b, \Sigma_{bb})$. According to Bayes' law

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)} = \frac{\mathcal{N}(x; \mu, \Sigma)}{\mathcal{N}(x_b; \mu_b, \Sigma_{bb})} = \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\frac{1}{\sqrt{(2\pi)^{n_b} |\Sigma_{bb}|}} \exp\left(-\frac{1}{2}(x_b - \mu_b)^T \Sigma_{bb}^{-1}(x_b - \mu_b)\right)}$$

$$= \frac{1}{\sqrt{(2\pi)^{n-n_b}}} \sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \frac{1}{2}(x_b - \mu_b)^T \Sigma_{bb}^{-1}(x_b - \mu_b)\right)$$

Denote the inverse $\Sigma^{-1}$ as $\Sigma^{-1} = \begin{pmatrix} \Sigma'_{aa} & \Sigma'_{ab} \\ \Sigma'_{ba} & \Sigma'_{bb} \end{pmatrix}$

$$p(x_a|x_b) = \frac{1}{\sqrt{(2\pi)^{n-n_b}}} \sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}} \exp\left(-\frac{1}{2}\left(\begin{matrix} x_a - \mu_a \\ x_b - \mu_b \end{matrix}\right)^T \begin{pmatrix} \Sigma'_{aa} & \Sigma'_{ab} \\ \Sigma'_{ba} & \Sigma'_{bb} \end{pmatrix} \left(\begin{matrix} x_a - \mu_a \\ x_b - \mu_b \end{matrix}\right)\right.$$

$$+ \frac{1}{2}(x_b - \mu_b)^T \Sigma_{bb}^{-1}(x_b$$

$$\left. - \mu_b)\right) \overset{(\Sigma'_{ba}{}^T = \Sigma'_{ab})}{=} \frac{1}{\sqrt{(2\pi)^{n-n_b}}} \sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}} \exp\left(-\frac{1}{2}((x_a - \mu_a)^T \Sigma'_{aa}(x_a - \mu_a)\right.$$

$$+ 2(x_a - \mu_a)^T \Sigma'_{ab}(x_b - \mu_b) + (x_b - \mu_b)^T \Sigma'_{bb}(x_b - \mu_b))$$

$$\left. + \frac{1}{2}(x_b - \mu_b)^T \Sigma_{bb}^{-1}(x_b - \mu_b)\right) \qquad \text{(1)}$$

According to question 5

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} & -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \\ -\Sigma_{bb}^{-1}\Sigma_{ba}(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} & \Sigma_{bb}^{-1} + \Sigma_{bb}^{-1}\Sigma_{ba}(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \end{pmatrix}$$

Plugging the inverse into (1):

$$p(x_a|x_b) = \frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}}\exp\left(-\frac{1}{2}\left((x_a-\mu_a)^T\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}(x_a-\mu_a)\right.\right.$$

$$-2(x_a-\mu_a)^T\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)+(x_b-\mu_b)^T(\Sigma_{bb}^{-1}$$

$$\left.\left.+\Sigma_{bb}^{-1}\Sigma_{ba}\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1})(x_b-\mu_b))+\frac{1}{2}(x_b-\mu_b)^T\Sigma_{bb}^{-1}(x_b-\mu_b)\right)\right)$$

$$=\frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}}\exp\left(-\frac{1}{2}\left((x_a-\mu_a)^T\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}(x_a-\mu_a)\right.\right.$$

$$-2(x_a-\mu_a)^T\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)$$

$$\left.\left.+(x_b-\mu_b)^T\Sigma_{bb}^{-1}\Sigma_{ba}\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b))\right)\right)$$

$$=\frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\Sigma_{bb}|}{|\Sigma|}}\exp\left(-\frac{1}{2}\left[(x_a-\mu_a)-\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)\right]^T\left(\Sigma_{aa}\right.\right.$$

$$\left.\left.-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\left[(x_a-\mu_a)-\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)\right]\right)$$

Also using the determinant derived in question 5

$$\det\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}=|\Sigma_{bb}||\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}|$$

Thus

$$p(x_a|x_b)=\frac{1}{\sqrt{(2\pi)^{n_a}|\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}|}}\exp\left(-\frac{1}{2}\left[(x_a-\mu_a)-\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)\right]^T\left(\Sigma_{aa}\right.\right.$$

$$\left.\left.-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\left[(x_a-\mu_a)-\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)\right]\right)$$

Which is the probability distribution of multivariate normal distribution

$$\mu_{a|b}=\mu_a+\Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)$$

$$\Sigma_{a|b}=\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

## 10)

### 10-i)

$$\min_x\|Ax-b\|=\min_x(Ax-b)^T(Ax-b)=\min_x\left(x^TA^TAx-x^TA^Tb-b^TAx-b^Tb\right)$$

The last term does not affect the minimization over x. By differentiating with respect to x and setting to zero:

$$\frac{\partial x^TA^TAx-x^TA^Tb-b^TAx}{\partial x}=0=2x^TA^TA-2b^TA$$

Therefore

$$x = (A^T A)^{-1} A^T b = A^\dagger b$$

## 10-ii)

Assume that the algorithm has converged. We shall have $x^{(t+1)} = x^{(t)} = x$

$$x = x - v(A^T A x - A^T b) \rightarrow v(A^T A x - A^T b) = 0 \rightarrow x = (A^T A)^{-1} A^T b$$

## 10-iii)

Let $x^* = (A^T A)^{-1} A^T b$ be the optimum solution. By adding and subtracting $x^*$ from both sides (Using t'th iteration notation by subscript)

$$x_{t+1} - x^* = x_t - x^* - v(A^T A)\left(x_t - (A^T A)^{-1} A^T b\right) = x_t - x^* - v(A^T A)(x_t - x^*)$$

Define $y_t = x_t - x^*$

$$y_{t+1} = y_t - v(A^T A)y_t = (I - vA^T A)y_t$$

We know that $A^T A$ can be diagonalized by $A^T A = Q\Lambda Q^T$. Multiply both sides by $Q^T$ and define $z_t = Q^T y_t$.

$$Q^T y_{t+1} = z_{t+1} = Q^T(I - vA^T A)y_t = Q^T(QQ^T - vQ\Lambda Q^T)y_t = Q^T Q(I - v\Lambda)Q^T y_t = (I - v\Lambda)z_t$$

$(I - v\Lambda)$ is a diagonal matrix. Therefore, for the i'th element of $z_k$ after k iterations

$$z_k^i = (1 - v\lambda_i)^k z_0^i$$

Therefore, for the i'th mode to converge ($k \rightarrow \infty$) we shall have

$$|1 - v\lambda_i| < 1 \rightarrow 0 < v < \frac{2}{\lambda_i}$$

And to satisfy convergence of all modes we shall have

$$0 < v < \frac{2}{\lambda_{max}}$$